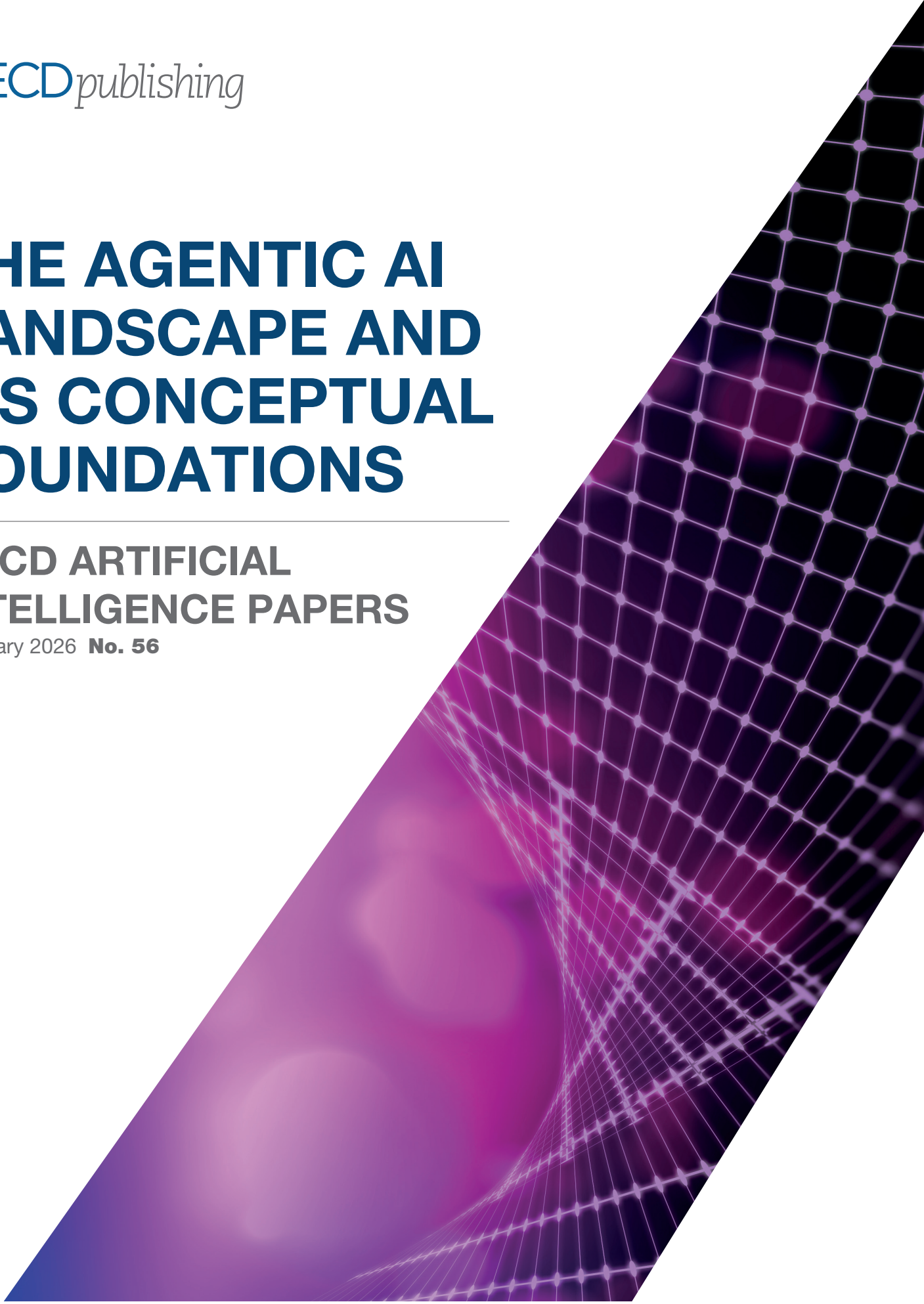
OECD publishing

THE AGENTIC AI LANDSCAPE AND ITS CONCEPTUAL FOUNDATIONS



OECD ARTIFICIAL INTELLIGENCE PAPERS

February 2026 **No. 56**

GPAI

OECD
BETTER POLICIES FOR BETTER LIVES

The agentic AI landscape and its conceptual foundations

Note to Delegations:

This document is also available on O.N.E Members & Partners under the reference code:

DSTI/DPC/GPAI(2025)/18/FINAL

Abstract

This report examines how the terms “AI agents” and “agentic AI” are defined and used across the literature by analysing their key features and points of overlap and distinction, and how these key features relate to the core elements of the OECD AI system definition. The analysis shows that while AI agents and agentic AI share foundational characteristics, agentic AI places stronger emphasis on co-ordination among multiple agents, task decomposition and delegation, sustained operation over time, and operation in more complex and less predictable environments with limited human oversight. The report also presents descriptive evidence on recent trends in the uptake of AI agents, indicating that although many developers are beginning to integrate them into their workflows, further progress is needed toward more trustworthy systems. Overall, the report contributes to a clearer conceptual understanding in this emerging area and provides a foundation for future analytical work.

Acknowledgements

This working paper describes preliminary results and research in progress and is published to stimulate discussion on the agentic AI landscape. It has been prepared by Luis Aranda and Kasumi Sugimoto from the OECD AI and Emerging Digital Technologies (AIEDT) Division, under the strategic direction of Audrey Plonk, Deputy Director of the OECD Directorate for Science, Technology and Innovation (STI) and Karine Perset, Deputy Head of AIEDT.

The report benefitted substantially from the contributions and guidance of Vincent Corruble (Sorbonne University) and Francesca Rossi (IBM), co-chairs of the OECD.AI Expert Group on Agentic AI. It was presented and discussed in November 2025 at the fourth Plenary meeting of the Global Partnership on Artificial Intelligence (GPAI) and at an ad hoc Agentic AI expert workshop attended by over 190 experts.

The report incorporates oral and written contributions from delegates and experts. In particular, the authors extend their sincere gratitude to the Delegations of Brazil, Canada, Chile, Colombia, Denmark, Germany, Greece, Israel, Japan, Mexico, Saudi Arabia, Singapore, Slovenia, Türkiye, the United Kingdom, the United States, Business at OECD (BIAC), and Civil Society Information Society Advisory Council (CSISAC).

The authors are also grateful to experts from the Agentic AI Expert Group, the Expert Group on AI Futures, and the OECD.AI network of experts, whose valuable feedback informed this report. In particular, contributions were received from Amit Ashkenazi (Tel Aviv University), Carolyn Ashurst (The Alan Turing Institute), Marta Bieńkiewicz (Cooperative AI Foundation), Virginia Dignum (Umeå University), Charles Fadel (Center for Curriculum Redesign), Yuko Harayama and Merve Hickok (Tokyo Centre of the GPAI Expert Community), Toshiya Jitsuzumi (Chuo University), Martin Marzidovšek (Jožef Stefan Institute), Nicolas Moës (The Future Society), Stuart Russell (University of California, Berkeley), Said Saillant (Societas Sapiens, Inc), Michael Schönstein (Federal Chancellery of Germany), Borys Stokalski (RETHINK), Osamu Sudoh (Chuo University), and Kyoko Yoshinaga (Keio University).

Finally, the authors thank all those who have contributed to the report throughout its development. This includes Jeff Mollins and Eunseo Choi (AIEDT); Gallia Daor (STI); Jamie Berryhill from the OECD Directorate for Public Governance (GOV); the Tokyo Centre of the GPAI Expert Community. The authors also thank Shellie Laffont, Christy Dentler, John Tarver (AIEDT), and Andreia Furtado from STI Communications for editorial support.

This paper is part of the OECD Horizontal Project on *Thriving with AI: Empowering Economies and Societies*.

Table of contents

FIGURES

TABLES

Executive summary

AI agents and agentic AI are receiving growing attention as artificial intelligence (AI) systems based on large language models (LLMs) become more autonomous and capable of interacting with their environments. While related concepts have long been studied in academic research, recent advances introduce new capabilities that challenge existing conceptual boundaries and highlight the need for a clearer, shared understanding of what constitutes an "agentic" AI system.

This report contributes to that clarity by examining how the terms "AI agents" and "agentic AI" are defined and used across the literature. By analysing definitions of both terms and highlighting their key features and points of overlap and distinction, and relating them to the core elements of the OECD AI system definition, the report supports more precise and consistent use of terminology and establishes a basis for further analytical work.

The analysis finds that both concepts share foundational characteristics, including a degree of autonomy, goal-directed behaviour, and the ability to perceive and act within their physical or virtual environment. However, agentic AI places greater emphasis on co-ordination among multiple agents, task decomposition and delegation, sustained operation over time, and functioning in more complex and less predictable environments with limited human oversight. Based on the analysis, the report provides the following common understanding:

- *AI agents* are systems that can perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts.
- *Agentic AI* generally refers to systems composed of multiple co-ordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods. Agentic AI systems are designed to operate in more open-ended, less predictable physical or virtual environments and to function with minimal human supervision.

Agentic AI systems are more than technical tools; they are increasingly regarded as systems embedded in social contexts and interactions, operating in a socio-technical paradigm. The value of agentic AI systems comes from their ability to act autonomously and interact with other agents – human, artificial, or institutional – through co-ordination and negotiation. Supporting these interactions requires advanced reasoning capabilities as well as robust infrastructure and communication protocols. This relational perspective is essential to designing agentic AI systems that can function responsibly and effectively in physical and virtual environments.

Many developers have integrated AI agents into their toolkits, and survey data indicate that nearly half of Stack Overflow respondents are using them or plan to do so. However, adoption does not indicate full technological maturity: respondents still highlight opportunities to further strengthen security, privacy, and accuracy, underscoring the need for continued progress toward more trustworthy AI agents.

Overall, the report provides a descriptive overview of the agentic AI landscape, clarifying key concepts and characteristics and establishing a shared analytical foundation. Going forward, improved understanding of use cases and technical architectures can help identify where safeguards and standards are most needed. Further analytical work could build on this foundation by developing policy-relevant typologies based on

features such as level of autonomy, adaptiveness, domain of operation, and system impact, as well as by improving empirical evidence on adoption and use across different contexts.

1 Background

The field of artificial intelligence (AI) is advancing rapidly, with many systems currently designed only to react on demand to human queries with information for humans to consider. However, there is growing interest in AI systems that function more like autonomous agents, capable of pursuing goals, making decisions, and taking actions with minimal human intervention across a wider range of tasks. These emerging AI agents and agentic AI systems hold the potential to drive innovation, attract investment, and improve productivity across multiple sectors by streamlining processes and enabling more efficient operations (Zeff and Wiggers, 2025[1]).

Despite growing attention, there is significant variation in how AI agents and agentic AI systems are understood. While the academic literature has long offered definitions of agents in the field of AI, recent developments – especially the integration of large language models (LLMs) – have introduced new complexities and competing interpretations. As the landscape continues to evolve, researchers, developers, and policymakers are working to track progress and explore governance approaches that support innovation while promoting trustworthy AI.

The OECD.AI Expert Group on AI Futures is a multidisciplinary, cross-sector group providing insights and advice to governments about current and possible future developments in AI. Its role includes equipping governments with the evidence and tools needed to help them devise future-ready AI policies. As part of this work, agentic AI has been identified as a priority workstream. This paper maps the current definitional landscape of “AI agents” and “agentic AI” and serves as the first project within the agentic AI stream. It is intended to be descriptive so as to lay the foundation for future analytical and policy work. This analysis identifies the most frequently cited features in existing definitions of agentic AI and AI agents, examines how these features are described across sources, and maps them to the key elements of the OECD definition of an AI system. By highlighting both shared traits and differences, the paper aims to support clearer conceptual understanding and inform future research and policymaking. It also provides descriptive data on recent trends in the uptake of AI agents and agentic AI.

2 What is an AI system?

2.1. OECD definition of an AI system

The OECD Council Recommendation on Artificial Intelligence defines an AI system as *"a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."* (OECD, n.d.[2])

2.2. Main elements of the OECD definition of an AI system

The Explanatory Memorandum on the Updated OECD Definition of an AI System and *the OECD Framework for the Classification of AI Systems* provide complementary information about the key elements of this definition (OECD, 2024[3]; OECD, 2022[4]). They include:

- **Objectives**: AI system objectives can be explicit or implicit; for example, they can belong to the following categories, which may overlap in some systems:
  a) Explicit and human-defined – where the developer encodes the objective directly into the system (e.g., through an objective function). Examples of systems with explicit objectives include simple classifiers, game-playing systems, combinatorial problem-solving systems, planning algorithms, and dynamic programming algorithms.
  b) Implicit in (typically human-specified) rules – rules dictate the action to be taken by the AI system according to the current circumstance. For example, a driving system might have a rule, "If the traffic light is red, stop." However, the objectives underlying these rules, such as compliance with the law or avoiding accidents, are not explicit, even though the rules are typically human-specified.
  c) Implicit in training data – where the ultimate objective is not explicitly programmed but incorporated through training data and a system architecture that learns to emulate that data (e.g., rewarding large language models for generating a plausible response).
  d) Not fully known in advance – some examples include recommender systems that use reinforcement learning to gradually narrow down a model of individual users' preferences.
- **Input and data**: Input is used both during development and after deployment. Input can take the form of data, knowledge, rules and code humans put into the system during development. Humans and machines can provide input. During development, input is leveraged to build AI systems, e.g., with machine learning that produces a model from training data and/or human input. Input is also used by a system in operation, for instance, to infer how to generate outputs. Input can include data relevant to the task to be performed or take the form of, for example, a user prompt or a search query.

- **Inference**: The concept of "inference" generally refers to the step in which a system generates intermediate or final outputs from its inputs, typically after deployment.
- **Outputs**: The output(s) of an AI system generally reflect the functions it performs, typically falling into broad categories such as recommendations, predictions, content generation, decisions, and actions. These categories are associated with varying levels of system autonomy and human involvement. Decisions and actions often indicate higher levels of autonomy, where the AI system directly affects its environment or directs another entity to do so. In contrast, predictions and recommendations usually involve more human involvement. However, the level of autonomy is not determined by output type alone – it also depends on how the system is designed, deployed, and built upon.
- **Influence on physical or virtual environments**: An AI system environment or context refers to the space – either physical or virtual – that the system can observe, fully or partially, through data or other inputs. An AI system's environment can be affected or influenced by the system's outputs.
- **Autonomy**: The degree to which a system can learn or act without human involvement. Human supervision can occur at any stage of the AI system lifecycle, such as during design, development or deployment. Some AI systems can generate outputs without these outputs being explicitly described in the AI system's objective and without specific instructions from a human. Some systems generate outputs only when queried and remain inactive otherwise; others continuously monitor the environment for cues; and some produce a stream of outputs once initiated. Action autonomy includes several levels:
  a) No-action autonomy (also referred to as "human support"): The system can make recommendations, but only the human decides whether to act on them.
  b) Low-action autonomy (also referred to as "human-in-the-loop"): The system suggests an action, but only proceeds if the human approves.
  c) Medium-action autonomy (also referred to as "human-on-the-loop"): The system acts on its own unless a human steps in to stop it.
  d) High-action autonomy (also referred to as "human-out-of-the-loop"): The system acts entirely on its own, without human involvement.
- **Adaptiveness**: Adaptiveness usually refers to AI systems that can continue to evolve after initial development. An adaptive system modifies its behaviour through direct interaction with input and data before or after deployment. Adaptiveness is frequently denoted as "learning" and is often used synonymously with "adaptability".

The following section leverages the key elements of the OECD definition of an AI system to examine various salient definitions of AI agents and agentic AI. It identifies shared characteristics and underscores important distinctions, providing a clear foundation to support informed policy discussions around these technologies.

The terms *autonomy* and *agency* are often used interchangeably, but they refer to distinct concepts with important implications for AI systems.

- Autonomy generally refers to a system's capacity to act without direct human involvement.
- Agency, by contrast, involves a system's capacity for independent goal formulation, long-term reasoning, and strategic adaptation (Tallam, 2025[5]). Agency typically presupposes interaction: without some form of interaction, there can be no agency. Interactivity and relational dynamics are foundational to all forms of agency (Floridi, 2024[6]; Dignum and Dignum, 2020[7]).

A system may be autonomous without being agentic. For example, a thermostat or some types of autonomous vehicles can operate independently yet lack the capacity to reason about goals or interact meaningfully with their environment. Agency, in its fuller sense, typically presupposes at least some degree

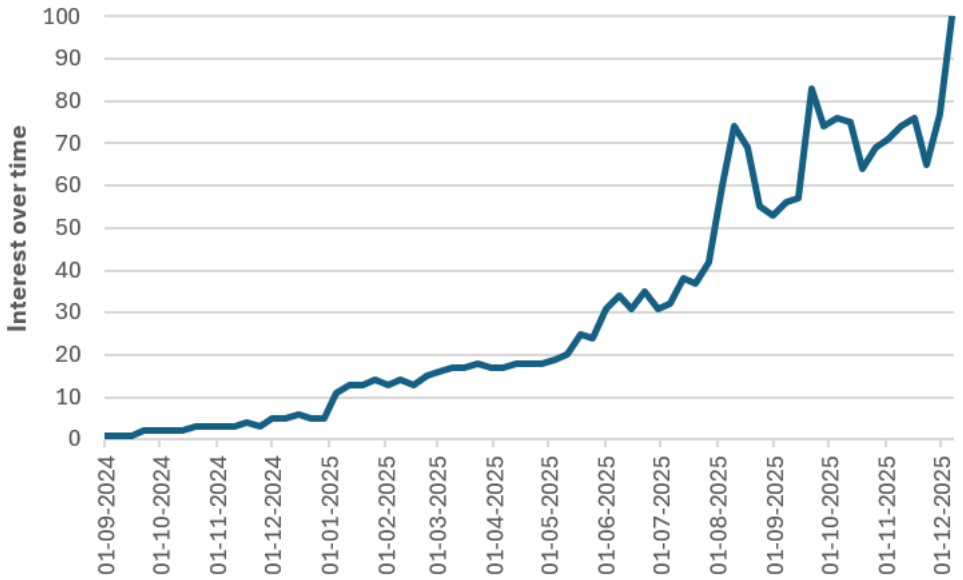
of autonomy, but also includes the capacity to engage with more complex, dynamic, and multi-agent environments, and often to translate higher-level goals into lower-level actions.

In the analysis that follows, the term autonomy aligns with the OECD definition of an AI system. Agency is not explicitly defined in the OECD definition, but can be inferred from elements such as adaptiveness, objectives, inference and the ability to influence the environment. A discussion of agency in socio-technical contexts is presented in Section 3.4.

3 What are AI agents and agentic AI?

The terms "*AI agents*" and "*agentic AI*" are now central to discussions of advanced AI systems. While both are widely used, agentic AI has risen in prominence more recently, largely driven by the emergence of more capable LLMs that can plan, act, and interact in agent-like ways. This development is reflected in Google Trends data: global searches for "agentic AI" increase sharply in 2025, indicating it is becoming a rapidly established concept in the AI landscape (Figure 3.1).

Figure 3.1. Global searches for "agentic AI" on Google surged in 2025



Note: Google Trends shows how popular a search is over time based on a sample of searches. The numbers do not show total search volume – instead, they are scaled from 0 to 100, with 100 being the peak interest for the time period of the chart. A downward trend means the search is becoming less popular compared to other searches, not necessarily that fewer people are searching for it overall.
Source: OECD analysis using data from Google Trends (Google, n.d.[8]).

This section explores the concepts of AI agents and agentic AI by reviewing and comparing salient definitions from multiple sources. Although some sources do not draw a clear distinction between "AI agents" and "agentic AI," the analysis seeks to highlight both the commonalities and the differences in how these terms are defined across the literature.

Based on the analysis, agentic AI refers to systems of multiple co-ordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods. These systems are designed to operate in more open-ended, less predictable physical or virtual environments and to function with minimal human supervision.

Box 3.1. Evolution of agentic research

The concept of an "agent" long predates its use in artificial intelligence. Derived from the Latin *agere*, meaning "to act", the term has deep roots in philosophy, where an agent is generally understood as a being with the capacity to act, and agency refers to the exercise of that capacity (Stanford Encyclopedia of Philosophy, 2015[9]). The notion of an "agent" is also well established in economics, where it refers to a decision maker whose behaviour is modelled as the solution to an optimisation or choice problem (Mas-Colell, Whinston and Green, 1995[10]).

Contemporary artificial intelligence research generally defines an intelligent agent as an entity that perceives its environment through sensors and acts upon it through actuators, often represented as a feedback loop between the agent and its operational context (Russell and Norvig, 1995[11]).

A common approach distinguishes between two broad types of agents: ***reactive*** and ***cognitive.***

- **Reactive agents** – also known as reflex agents – simply respond to stimuli without planning, co-ordinating, or setting goals (Teahan, 2014[12]; Russell and Norvig, 2022[13]). For example, a vacuum agent acts only on its current location and on whether that location contains dirt (Russell and Norvig, 2022[13]).
- **Cognitive agents** are designed to perceive their environment, reason about it, and make decisions based on objectives (Teahan, 2014[12]). A **rational agent** is a type of cognitive agent that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome (Russell and Norvig, 2022[13]).

The academic literature identifies several frameworks for categorising intelligent agents, primarily distinguished by either their functional sophistication or their underlying logical structure:

- **Capability-driven:** This categorisation organises agents along a spectrum of increasing autonomy and intelligence. It moves from simple reflex agents, which react instantly to the present moment, to learning agents, which can improve their own performance over time. Intermediate levels include model-based agents (maintaining an internal representation of the environment), goal-based agents (acting to achieve specific objectives), and utility-based agents (calculating the most "desirable" or efficient outcome) (Russell and Norvig, 2022[13]).
- **Architecture-driven:** This framework categorises agents by their internal logic. It distinguishes between logic-based agents (formal deduction), reactive agents (stimulus-response), and cognitive agents (goal-directed reasoning using internal world models). A key cognitive example is the Belief-Desire-Intention (BDI) architecture, which manages "mental states" representing an agent's beliefs, goals, and commitments (Rao and Georgeff, 1995[14]; Weiss, 2013[15]). Finally, layered architectures simply stack different levels of reasoning – from basic reflexes to high-level planning – into one co-ordinated system (Weiss, 2013[15]).

Building on these foundational concepts, research has progressed into multiple lines, including **multi-agent systems** (MAS) that focus on how agents co-ordinate, communicate, and negotiate, where the breakdown or distribution of roles is typically determined by human-designed system configurations (AAAI, 2025[16]; Wooldridge, 2002[17]). The emergence of LLMs from 2020 onwards has renewed interest in AI agents and broadened the landscape of AI-driven tools, leading to the introduction of new concepts such as "*copilots*" – that support users within specific workflows by offering real-time assistance – and more autonomous "*agentic AI systems*" which can carry out complex tasks with minimal human input (Kashyap, 2025[18]; AAAI, 2025[16]). A new frontier explores self-improving agents designed to recursively analyse their own performance during operation to become more efficient over time (Yin et al., 2024[19]).

3.1. AI Agents

3.1.1. Conceptual foundations of an AI agent

Discussions about AI agents have evolved over decades – from early rule-based systems to today's LLM-enabled agents (Box 3.1) (AAAI, 2025[16]; Dignum and Dignum, 2020[7]; IBM, 2025[20]). To make sense of these discussions, it is important to understand the foundational concepts and taxonomies that define AI agents. Table 3.1 presents characterisations of AI agents drawn from recent literature and standard academic definitions. A structured analysis of these definitions, mapped against the key components of the OECD AI system definition, reveals a hierarchy in the importance and frequency of definitional elements (Table 3.2). These elements have been grouped into three tiers: those most commonly mentioned ("prevalent"), those moderately referenced ("frequent"), and those less frequently included ("occasional"). This tiered view provides a practical framework to understand which attributes are most salient in AI agents.

- **Prevalent:** Three elements emerge as central to most AI agent definitions: **1) objectives, 2) outputs** (often in the form of actions)**,** and **3) autonomy**. These are the most consistently cited features across sources and form the foundation of AI agent definitions. For example, IBM (2025[20]) and NIST (2025[21]) both emphasise that AI agents are designed to pursue specific goals and produce outputs that include decisions and actions with some degree of autonomy. However, different definitions characterise these elements differently:
    - **1) Objectives:** Most sources recognise the role of objectives or goals in guiding the agent's behaviour and outputs. While most definitions include goals broadly, some restrict the application of AI agents to *complex* goals (NIST, 2025[21]; Oueslati and Staes-Polet, 2025[22]; Kasirzadeh and Gabriel, 2025[23]). Most sources do not mention whether the goals are implicit or explicit, except Anthropic (2024[24]) which seems to characterise objectives as human-specified via *prompts*, highlighting a unique framing tied to language model-based agents. Whereas certain studies, such as Sapkota, Roumeliotis and Karkee (2025[25]), characterise AI agents as task-specific systems designed to operate within narrowly scoped domains, Bengio et al. (2025[26]) define AI agents as "general-purpose AI." Other works, however, do not explicitly link task specificity to the definition of AI agents.
    - **2) Outputs:** Definitions of outputs in AI agents vary in scope and emphasis. Most sources describe outputs as *decisions* and *actions* – including task execution. Notably, Chan et al. (2025[27]) Bengio et al. (2025[26]), and Ferber (1999[28]) expand this to include *interactions* in open-ended environments, while Anthropic (2024[24]), Partnership on AI (2025[29]), Hugging Face (2025[30]), and LangChain Docs (2025[31]) highlight outputs shaped by *tool usage,* including through structured application programming interfaces (APIs) and by interacting directly with graphical user interfaces (GPUs) (Nguyen et al., 2024[32]). This reflects a spectrum from simple decision making and task execution to complex, context-sensitive responses.
    - **3) Autonomy:** Definitions of autonomy in AI agents vary in depth and scope. Some sources, like Russell and Norvig (1995[11]), IBM (2025[20]) and Fadel (2025[33]), imply basic agency and autonomy, noting "agency to act" or "take actions autonomously". Notably, NIST (2025[21]) seems to distinguish between autonomous *decision-making*, which occurs without human input, and *action-taking*, which may involve limited human supervision. This suggests a layered view of autonomy, where reasoning is fully independent but execution may be subject to human oversight and feedback loops (Anthropic, 2024[24]). Such configurations are sometimes referred to as hybrid systems, where the agent and the human interact in the process leading to an action (Samdani, Viswanathan and Jegadeesh, 2025[34]).

- **Frequent:** The second tier includes **1) influence on the environment**, **2) adaptiveness,** and **3) inference**. These elements are mentioned frequently but not as universally as the core trio.
    - **1) Influence on the physical or virtual environment:** Most sources describe agents as affecting their immediate physical or virtual environment through their outputs. However, what constitutes the "environment" varies across systems and applications. Notably, Chan et al. (2025[27]) and Oueslati and Staes-Polet (2025[22]) emphasis that agents can act in open-ended environments, while Sapkota, Roumeliotis and Karkee (2025[25]) limit influence to digital ecosystems. These different assumptions about where the agent can operate correspond to different definitions of what the "environment" is.
    - **2) Adaptiveness:** Some sources refer to the agent's ability to adjust its behaviour and outputs in response on changing inputs or contexts (Mitchell et al., 2025[35]). This often involves an iterative process to achieve objectives (Masterman et al., 2024[36]; Anthropic, 2024[24]; Capgemini Research Institute, 2025[37]; Russell and Norvig, 1995[11]; LangChain Docs, 2025[31]). In several recent sources, this iterative execution and learning pattern is called the agentic loop (Ashby, 2025[38]). While the ability to act in open-ended environments suggests a high degree of adaptiveness and autonomy (Chan et al., 2025[27]; Oueslati and Staes-Polet, 2025[22]), Sapkota, Roumeliotis and Karkee (2025[25]) characterise adaptiveness in AI agents as "limited".
    - **3) Inference:** While not explicitly referenced in the sources, inference is implied as the ability of an AI system to generate outputs from given inputs. Some sources also emphasise planning or reasoning – particularly in the case of cognitive agents – reflecting the complexity or sophistication often associated with inference (NIST, 2025[21]; Chan et al., 2025[27]; Masterman et al., 2024[36]; Anthropic, 2024[24]; Hugging Face, 2025[30]; Bengio et al., 2025[26]; LangChain Docs, 2025[31]; Oueslati and Staes-Polet, 2025[22]; Partnership on AI, 2025[29]).
- **Occasional:** The third tier includes **data and input**. This element appears less consistently in AI agent definitions. Although this component provides important technical context, it is usually treated as a supporting rather than a defining feature.
    - **Data and input:** This is among the least emphasised elements in AI agent definitions. Only a few sources (Sapkota, Roumeliotis and Karkee, 2025[25]; Anthropic, 2024[24]) explicitly mention how agents perceive or understand inputs. Notably, the reviewed definitions fail to clearly differentiate between pre-deployment training data – which include the data, experiments, or simulations used to shape an agent's capabilities – and post-deployment inputs, including interactions with the environment and other agents as different input forms. This limited explicit inclusion may reflect the assumption that data and inputs do not distinguish AI agents from other AI systems.

The analysed definitions commonly refer to autonomy, adaptiveness, output types, and interaction with the environment. However, there is no clear consensus on the required levels of autonomy or adaptiveness, the generality and nature of outputs, or the type of environment and degree of influence an agent should have within it.

This section aims to identify the features that distinguish AI agents from other AI systems. Like all AI models, AI agents depend on data for training; however, this reliance does not seem to be a distinguishing characteristic. Data and inputs nonetheless remain essential, and understanding how they are handled within AI agents is therefore important.

In summary, AI agents can be understood as systems that perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts.

Table 3.1. Illustrative definitions of AI agents

Source	Definition
Russell and Norvig (1995[11])	An agent is something that perceives and acts in an environment […] An ideal intelligent agent takes the best possible action in a situation […] computer agents are expected to […]: operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals.
Wooldridge (2002[17])	An agent is a computer system that is capable of independent action on behalf of its user or owner. In other words, an agent can figure out for itself what it needs to do in order to satisfy its design objectives, rather than having to be told explicitly what to do at any given moment.
Ferber (1999[28])	An agent is a physical or virtual entity (a) which is capable of acting in an environment, (b) which can communicate directly with other agents, (c) which is driven by a set of tendencies (in the form of individual objectives or of a satisfaction/survival function which it tries to optimise), (d) which possess resources of its own, (e) which is capable of perceiving its environment (but to a limited extent), (f) which has only a partial representation of this environment (and perhaps none at all), (g) which possesses skills and can offer services, (h) which may be able to reproduce itself, (i) whose behaviour tends towards satisfying its objectives, taking account of the resources and skills available to it and depending on its perception, its representations and the communications it receives.
IBM (2025[20])	An AI agent is a software entity that employs AI techniques and has agency to act in its environment based on set goals, which means it can decide which actions to perform and has the ability to execute them.
NIST (2025[21])	AI agent systems have the capability for autonomous decision-making and taking action to operate with limited human supervision to achieve complex goals. Characteristics of AI agent systems include the ability to understand context, reason, plan, adapt, and execute tasks.
Fadel (2025[33])	Agents: are software entities that take actions autonomously to achieve specific goals.
Chan et al. (2025[27])	AI agents: AI systems that can plan and execute interactions in open-ended environments, such as making phone calls or buying online. Agents differ from other computational systems in two significant ways. First, in comparison to foundation models used as chatbots, agents directly interact with the world (e.g., a flight booking website) rather than only with users. Second, in comparison to traditional software (e.g., an implementation of a sorting algorithm), agents can adapt to under-specified task instructions. Although the AI community has been developing agents for decades, these agents typically performed only a narrow set of tasks. In contrast, recent agents built upon language models can attempt—with varying degrees of reliability —a much wider array of tasks, such as software engineering or office support.
Kasirzadeh and Gabriel (2025[23])	We characterise AI agents as systems that have the ability to perform increasingly complex and impactful goal-directed action across multiple domains, with limited external control. In essence, our focus is on a large class of artificial systems that are able to independently pursue a wide range of goals and tasks, thereby exerting causal influence on the world.
Masterman et al. (2024[36])	AI agents are language model-powered entities able to plan and take actions to execute goals over multiple iterations. AI agent architectures are either comprised of a single agent or multiple agents working together to solve a problem.
Sapkota, Roumeliotis and Karkee (2025[25])	AI agents are an autonomous software entities engineered for goal-directed task execution within bounded digital environments. These agents are defined by their ability to perceive structured or unstructured inputs, reason over contextual information, and initiate actions toward achieving specific objectives, often acting as surrogates for human users or subsystems. Unlike conventional automation scripts, which follow deterministic workflows, AI agents demonstrate reactive intelligence and limited adaptability, allowing them to interpret dynamic inputs and reconfigure outputs accordingly…AI Agents are purpose-built for narrow, well-defined tasks. They are optimised to execute repeatable operations within a fixed domain, such as email filtering, database querying, or calendar co-ordination. This task specialisation allows for efficiency, interpretability, and high precision in automation tasks where general-purpose reasoning is unnecessary or inefficient.
Anthropic (2024[24])	Agents, are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks…Agents are emerging in production as LLMs mature in key capabilities—understanding complex inputs, engaging in reasoning and planning, using tools reliably, and recovering from errors. Agents begin their work with either a command from, or interactive discussion with, the human user. Once the task is clear, agents plan and operate independently, potentially returning to the human for further information or judgement. During execution, it's crucial for the agents to gain "ground truth" from the environment at each

	step (such as tool call results or code execution) to assess its progress. Agents can then pause for human feedback at checkpoints or when encountering blockers. The task often terminates upon completion, but it's also common to include stopping conditions (such as a maximum number of iterations) to maintain control. Agents can handle sophisticated tasks, but their implementation is often straightforward. They are typically just LLMs using tools based on environmental feedback in a loop.
Capgemini Research Institute (2025[37])	AI agents are programs/platforms/software that are connected to the business environment with a defined boundary, make decisions autonomously, and act to achieve specific goals with or without human intervention. With the latest advances in reasoning AI models, AI agents are able to break down tasks, "reason" through potential pathways to find solutions to the given problem, try those solutions, and present successful outcomes.
Bengio et al. (2025[26])	AI agent: A general-purpose AI which acts to achieve goals, possibly using plans, adaptively performing tasks involving multiple steps and uncertain outcomes along the way, and interacting with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight.
Oueslati and Staes-Polet (2025[22])	Agents are characterised by their capacity to (i) autonomously pursue complex, underspecified goals, engaging in long-term and adaptive planning and (ii) take actions in both virtual and real-world environments. Agents are a compound system, consisting of a GPAI model – currently an advanced LLM or multimodal model – that is connected to "scaffolding" software, affordances that aim to enable more effective planning and goal execution.
Mitchell et al. (2025[35])	"AI agents" are computer software systems capable of creating context-specific plans in non-deterministic environments.
Partnership on AI (2025[29])	Agents […] reason, plan, and perform sequences of actions to achieve user goals. Unlike generative AI, these systems directly execute actions by using digital tools to interact with complex environments.
Hugging Face (2025[39]; 2025[30])	An agent is a system that leverages an AI model to interact with its environment in order to achieve a user-defined objective. It combines reasoning, planning, and the execution of actions (often via external tools) to fulfill tasks. AI agents are programs where LLM outputs control the workflow. "Agency" evolves on a continuous spectrum, as you give more or less power to the LLM on your workflow.
LangChain Docs (2025[31])	Agents combine language models with tools to create systems that can reason about tasks, decide which tools to use, and iteratively work towards solutions.

Note: This table includes definitions drawn from a range of source types, including academic publications, websites, and developer documentation. These sources are considered together because all contribute to shaping contemporary understandings of AI agents. No weighting is applied, in order to capture the breadth of the ongoing conceptual debate. The table includes earlier definitions of "agents" that are not specific to AI agents, where such definitions have informed or contributed to the current understanding of AI agents.

Table 3.2. Most definitions characterise AI agents as systems that produce outputs to achieve objectives with a certain degree of autonomy

Mapping of key components of the OECD AI system definition explicitly highlighted in AI agent definitions.

Source	Objectives	Outputs	Autonomy	Influence on environment	Adaptiveness	Inference	Data and input
Russell and Norvig (1995[11])	X	X	X		X		X
Wooldridge (2002[17])	X	X	X				
Ferber (1999[28])	X	X	X	X	X		X
IBM (2025[20])	X	X	X	X			
NIST (2025[21])	X	X	X	X	X	X	
Fadel (2025[33])	X	X	X				
Chan et al. (2025[27])	X	X	X	X	X		
Kasirzadeh and Gabriel (2025[23])	X	X	X	X	X		
Masterman et al. (2024[36])	X	X	X		X	X	
Sapkota, Roumeliotis and Karkee (2025[25])	X	X	X	X	X	X	X
Anthropic (2024[24])	X	X	X	X	X	X	X
Capgemini Research Institute (2025[37])	X	X	X	X	X	X	
Bengio et al. (2025[26])	X	X	X	X	X	X	
Oueslati and Staes-Polet (2025[22])	X	X	X	X	X	X	
Mitchell et al. (2025[35])	X	X	X	X	X		
Partnership on AI (2025[29])	X	X	X	X		X	
Hugging Face (2025[39]; 2025[30])	X	X		X		X	
LangChain Docs (2025[31])	X	X	X			X	
Total	*18*	*18*	*17*	*13*	*12*	*10*	*4*

3.2. Agentic AI

3.2.1. Key features of agentic AI

The terms “AI agents” and “agentic AI” are sometimes used interchangeably, leading to confusion and a lack of conceptual clarity. However, a closer examination reveals important distinctions between these two concepts. This section explores the notion of agentic AI by comparing it systematically with definitions of AI agents.

By analysing and contrasting key definitions from academic and technical sources (Table 3.3), we aim to clarify how these terms differ at a semantic and functional level. Naturally, several core components of the OECD definition of an AI system – such as autonomy, adaptiveness, output types (e.g., decisions or actions), and interaction with the environment – are also frequently referenced in definitions of agentic AI. While there is significant overlap between definitions of AI agents and agentic AI, certain characteristics are either more strongly emphasised or unique to agentic AI:

- **Objectives**:
  - **Task decomposition and delegation:** Agentic AI systems can break down complex objectives into smaller, manageable tasks and delegate them to individual agents within the system (Miehling et al., 2025[40]; Sapkota, Roumeliotis and Karkee, 2025[25]; Challapally et al., 2025[41]).
  - **Extended timeframes:** Agentic AI systems can pursue goals over longer periods, introducing a temporal dimension absent in most AI agent definitions (CSET, 2024[42]; Chan et al., 2023[43]; Shavit et al., 2023[44]). This temporal aspect is closely connected to the notion of autonomy.
- **Inference**:
  - **System-level architecture:** Agentic AI is often described as a system of co-ordinated AI agents, typically organised within LLM-enabled architectures (Sapkota, Roumeliotis and Karkee, 2025[25]; IBM, 2025[20]; Miehling et al., 2025[40]; Dignum and Dignum, 2025[45]; Yousefi, Billi and Rotolo, 2025[46]; Infocomm Media Development Authority, 2026[47]).
  - **Distributed problem-solving:** Agents in an agentic system work together with other agents and components by inferring, planning, and co-ordinating tasks to solve problems collectively (Sapkota, Roumeliotis and Karkee, 2025[25]; Miehling et al., 2025[40]). In the context of agentic AI, inference is evolving from immediate generation to a sophisticated process of "deliberative reasoning" or "test-time compute". Rather than producing instantaneous outputs, agentic systems can engage in internal chains of thought as well as recursive multi-agent critique and self-reflection during inference (Kim et al., 2025[48]; Zhao et al., 2025[49]).
- **Outputs**:
  - **Task complexity**: Sapkota, Roumeliotis and Karkee (2025[25]) highlights the greater task complexity and the broader range of tasks and domains that agentic AI systems can handle compared to AI agents, owing to their ability to manage dynamic and large-scale workflows.
- **Autonomy**:
  - **Significant autonomy and flexibility:** Agentic AI systems operate in more open-ended and complex environments, with less reliance on step-by-step instructions and supervision than individual agents (Chan et al., 2023[43]; Shavit et al., 2023[44]; Sapkota, Roumeliotis and Karkee, 2025[25]).
  - **High degrees of agency and adaptiveness:** Sapkota, Roumeliotis and Karkee (2025[25]) characterise agentic AI as systems capable of adapting in real time to environmental shifts or partial task failures through the dynamic sequencing of subtasks. Several other definitions frame agency as a spectrum, suggesting varying degrees of autonomy, adaptiveness and complexity across agentic AI systems (Shavit et al., 2023[44]; CSET, 2024[42]). Pant and Viswanathan (2025[50]) contributes to this layered understanding by outlining five levels of agentic AI autonomy. These range from basic rule-based automation to highly autonomous agentic systems capable of dynamic adaptiveness and autonomous tool discovery.
- **Environment**:
  - **More complex physical or virtual environments:** Agentic AI systems can operate in more open-ended and unpredictable settings, where the immediate physical or virtual environment and the possible actions and outcomes are harder to model (CSET, 2024[42];

Shavit et al., 2023[44]; AAAI, 2025[16]). In such settings, the consequences of actions may be initially unknown or uncertain, requiring agents to engage in some degree of exploration. This requires technical infrastructure and shared protocols to guide how systems interact and impact their environment (Chan et al., 2025[27]). Section 3.4 details some of the elements that agentic AI systems are expected to possess in a multi-agent environment.

In summary, agentic AI refers to systems composed of multiple co-ordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods. These systems are designed to operate in more open-ended, less predictable physical or virtual environments and to function with minimal human supervision.

Table 3.3. Illustrative definitions of agentic AI

Source	Definition
IBM (2025[20])	Agentic AI systems are software systems that leverage AI agents (together with other components like tools, planners, memory, and datasets), pursue goals, and can operate autonomously.
Chan et al. (2023[43])	We identify 4 key characteristics associated with increasing agency in algorithmic systems, especially in combination: underspecification, directness of impact, goal-directedness, and long-term planning. (1) Underspecification: the degree to which the algorithmic system can accomplish a goal provided by operators or designers, without a concrete specification of how the goal is to be accomplished. (2) Directness of impact: the degree to which the algorithmic system's actions affect the world without mediation or intervention by a human, i.e. without a human in the loop. (3) Goal-directedness: the degree to which the system acts as if it is designed/trained to achieve a particular quantifiable objective. (4) Long-term planning: the degree to which the algorithmic system is designed/trained to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon.
Shavit et al. (2023[44])	Agentic AI systems are characterised by the ability to take actions which consistently contribute towards achieving goals over an extended period of time, without their behaviour having been specified in advance…. We define the degree of agenticness in a system as "the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision." [...] we will generally refer to systems exhibiting high degrees of agenticness as "agentic AI systems," to emphasise that agenticness [...] we will generally conceptualize agentic AI systems as operating in pursuit of goals defined by humans and in environments determined by humans (and often in co-operation with human "teammates"), rather than fully-autonomous systems that set their own goals.
Miehling et al. (2025[40])	An agentic AI system, or simply agentic system, is a collection of agents interacting with humans and the environment with the objective of fulfilling specified goals. Practically, an agent is an LLM or large multimodal model (LMM) with access to tools — specialised components/functionalities like APIs, external services, computational resources, or domain-specific software — that allow it to perform specific operations in the environment. In this sense, tools define both the capabilities (actions) of the agent and the information (via observations/signals) that can be obtained from the environment. The human is responsible for seeding the initial task specification, providing clarification, and authorizing any (agent) actions that need human approval (Shavit et al., 2023). Given the task specification, an agent is able to interact with other agents (agent-agent interaction) to facilitate task decomposition/planning and delegation. This interaction can be co-operative or competitive, e.g., in the case of limited compute. The environment consists of everything external to the agentic system. This includes infrastructure (computers), other humans, other agents, and even other agentic systems.
Sapkota, Roumeliotis and Karkee (2025[25])	Agentic AI systems represent an emergent class of intelligent architectures in which multiple specialised agents collaborate to achieve complex, high-level objectives. As defined in recent frameworks, these systems are composed of modular agents each tasked with a distinct subcomponent of a broader goal and co-ordinated through either a centralised orchestrator or a decentralised protocol. This structure signifies a conceptual departure from the atomic, reactive behaviours typically observed in single-agent architectures, toward a form of system-level intelligence characterised by dynamic inter-agent collaboration. A key enabler of this paradigm is goal decomposition, wherein a user-specified objective is automatically parsed and divided into smaller, manageable tasks by planning agents. These subtasks are then distributed across the agent network. Multi-step reasoning and planning mechanisms facilitate the dynamic sequencing of these subtasks, allowing the system to adapt in real time to environmental shifts or partial task failures. This ensures robust task execution even under uncertainty. Inter-agent communication is mediated through distributed communication channels, such as asynchronous messaging queues, shared memory buffers, or intermediate output exchanges, enabling co-ordination without necessitating continuous central oversight. Furthermore, reflective reasoning and memory systems allow agents to store context across multiple interactions, evaluate past decisions, and iteratively refine their strategies. Collectively, these capabilities enable Agentic AI systems to exhibit flexible, adaptive, and collaborative intelligence that exceeds the operational limits of individual agents.
CSET (2024[42])	Goal complexity: More agentic systems pursue complex, longer-term goals—or even a variety of different goals. Less agentic systems carry out individual, more explicitly defined tasks. Environment complexity: More agentic systems can operate effectively in more open-ended and complicated settings, where the number of possible states and actions available to the agent is larger and the dynamics governing what will happen next in the environment are more difficult to model. Less agentic systems can operate effectively only in simpler and more predictable settings.

	Independent planning and adaptation: More agentic systems can generate their own plan or pathway to meet the intended goal, adapting as needed to changing circumstances. Less agentic systems follow pre-specified step-by-step instructions. Direct action: More agentic systems take action directly in their environment (whether real or virtual). Less agentic systems provide information or recommendations for a human user to act on.
Capgemini Research Institute (2025[37])	Agentic AI is a broader term [than AI agents] and includes systems, platforms, practices, tools, and technologies that enable agents to function.
AAAI (2025[16])	The concept of Agentic AI refers to the integration of generative AI and LLMs into autonomous agent frameworks aiming to leverage the generative capabilities of such models to enhance interaction, creativity, and real-time decision-making in dynamic environments.
Yousefi, Billi and Rotolo (2025[46])	The agentic AI paradigm thus differs from traditional AI agents in both architecture and function. Whereas single AI agents are typically designed to complete well-defined, tool-assisted tasks in isolation, agentic AI systems consist of multiple, specialised agents that communicate and co-ordinate to achieve shared objectives within open, evolving environments.
Dignum and Dignum (2025[45])	Agentic AI denotes systems that couple large-scale foundation models with capabilities to reason, act (e.g., via tools or environments), and interact with users and other systems in a sustained, goal-directed manner.
Challapally et al. (2025[41])	Agentic AI, the class of systems that embeds persistent memory and iterative learning by design [...]. Unlike current systems that require full context each time, agentic systems maintain persistent memory, learn from interactions, and can autonomously orchestrate complex workflows.
Infocomm Media Development Authority (2026[51])	Agentic AI systems are systems that can plan across multiple steps to achieve specified objectives, using AI agents.

Note: This table includes definitions drawn from a range of source types, including academic publications, websites, and developer documentation. These sources are considered together because all contribute to shaping contemporary understandings of agentic AI. No weighting is applied, in order to capture the breadth of the ongoing conceptual debate. The definitions come at different levels of abstraction. Some focus on technical features, while others include socio-technical aspects, such as interactions with humans. Both are included to reflect the complementary perspectives that shape current understandings of agentic AI.

3.3. Key elements and distinct characteristics

The terms AI agents and agentic AI are increasingly used in discussions of advanced AI, yet their definitions often overlap and are not always clearly distinguished. This report reviews multiple sources to clarify these concepts, identifying both shared attributes and key differences, and uses the OECD AI system definition as an analytical framework for this analysis.

Table 3.4 summarises these findings and provides a structured overview of the nuanced differences between AI agents and agentic AI, helping policymakers navigate this evolving terminology. It articulates which elements of the OECD AI system definition are most relevant to AI agents and agentic AI, and how agentic AI builds on these elements by placing greater emphasis on features like co-ordination, task decomposition, and operating in more complex environments with less human oversight. While definitions vary, the table reflects the most common themes from a definitional analysis of AI agents and agentic AI.

Table 3.4. Key differences between AI agents and agentic AI systems

Comparative framework based on the key elements of the OECD AI System Definition

Key elements *from OECD AI system definition*	**AI agents**	**Agentic AI**
Objectives	• Simpler, narrower goals with shorter timeframes	• More complex goals with task decomposition, delegation and longer timeframes
Outputs	• Execution of more basic tasks and decision making within a more limited action space	• Execution of more complex tasks within a larger action space
Autonomy	• Higher reliance on step-by-step instructions and closer supervision	• Greater autonomy, flexibility and agency; may include capacity for autonomous tool discovery
Influence on environment	• Less complex environments typically within digital ecosystems	• More open-ended, complex, and unpredictable environments
Adaptiveness	• Lower capacity to adapt	• Higher degrees of adaptiveness and interactivity with the environment
Inference	• Single agent reasoning	• Co-ordinated AI agents with distributed problem-solving
Data and input	• More limited, less dynamic data sources	• More diverse, dynamic data including from highly interactive environments

Note: This table illustrates which elements of the OECD AI system definition are most relevant to AI agents and agentic AI (blue means prevalent, green means frequent, and grey means occasional) and how agentic AI builds on these elements while introducing distinct features compared to AI agents. All comparative expressions (e.g., 'higher', 'more', 'less', 'greater') refer to relative differences between AI agents and agentic AI within this table, unless explicitly stated otherwise. While definitions vary, the table summarises the most common themes from a definitional analysis of AI agents and agentic AI.

3.4. Agentic AI as a socio-technical paradigm

Agentic AI is often closely related to the field of multi-agent systems (Sapkota, Roumeliotis and Karkee, 2025[25]; IBM, 2025[20]). In multi-agent systems, agents interact with other agents – human, artificial, and institutional – rather than operate in isolation. This requires more than just individual "intelligence"; agents need relational capabilities such as negotiation, co-ordination, and adherence to certain norms (Dignum and Dignum, 2025[45]; Trivedi et al., 2024[52]).

A core insight from early multi-agent system research was the importance of reasoning in social contexts, especially when agents have conflicting goals (Dignum and Dignum, 2025[45]). This led to a shift in focus from isolated decision-making to "social intelligence", including co-operation, argumentation, and

negotiation. Viewing agentic AI through this relational lens moves the discussion beyond autonomy alone, highlighting the need for systems that can interact responsibly and effectively with others, each with their own objectives, incentives, and constraints (AAAI, 2025[16]).

The true benefit of agentic AI would lie in its ability to operate within and contribute to a social context (Dignum and Dignum, 2025[45]). Rather than simply following goals or maximising outcomes, such systems should be able to learn from experience, adapt their behaviour over time, make decisions and take actions based not only on efficiency, but also on values and the social context. According to this view, agentic AI systems should be capable of handling potentially conflicting goals, adjusting to real-world situations, and knowing when a "good enough" outcome is more appropriate than constant optimisation (Dignum and Dignum, 2020[7]; Dignum and Dignum, 2025[45]).

Together, these perspectives suggest that agentic AI systems should be understood not only as a technical construct, but as systems embedded in a social context. This socio-technical paradigm brings together reasoning, ethics, and context to support effective interaction and decision-making in complex, multi-agent environments. The development of agentic systems requires a clear conceptual foundation for what it means to act, decide, and co-ordinate within complex socio-technical environments (Dignum and Dignum, 2025[45]). Lessons from multi-agent system research can provide a valuable foundation for designing more reliable, context-aware LLM-enabled multi-agent interactions (AAAI, 2025[16]).

Effective interaction depends on robust technical infrastructures and shared protocols that enable agents to communicate, co-ordinate, and be orchestrated to influence their surroundings (Chan et al., 2025[27]). Despite decades of work on agent communication languages, co-operation protocols, and interoperability initiatives, more is needed to facilitate standardisation and wide adoption (Anthropic, 2024[53]; Gosmar et al., 2024[54]; Surapaneni et al., 2025[55]; Hammond et al., 2025[56]; AAAI, 2025[16]). Notwithstanding, recent initiatives such as the Model Context Protocol (MCP) and the Agent-to-Agent protocol (A2A) illustrate increasing uptake of open, shared standards for connecting AI agents and applications to external tools, data sources, and systems (MCP, 2025[57]). Beyond these protocols, experimental platforms such as Moltbook – a decentralised social network designed exclusively for AI agents, where humans are "welcome to observe" – offer early demonstrations of large-scale agentic interactions and emerging collective dynamics (Moltbook, 2026[58]).

4 Recent trends in the uptake of AI agents and agentic AI

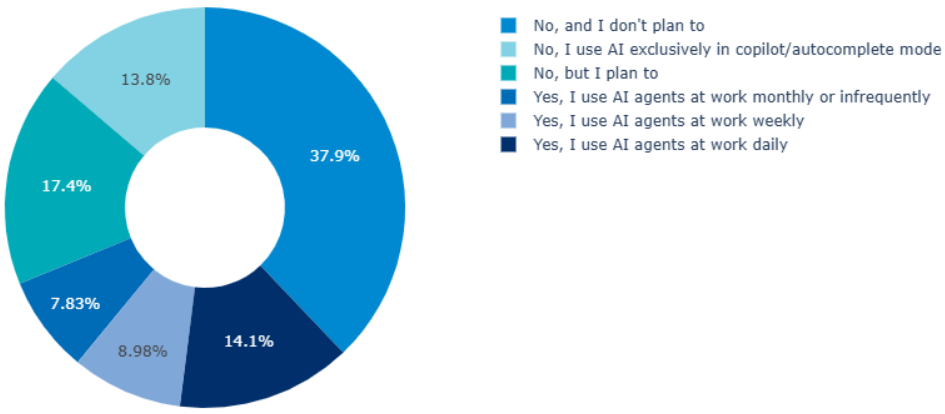
According to recent data, GitHub activity has seen a 920% increase in repositories using agentic AI frameworks such as AutoGPT, BabyAGI, OpenDevin, and CrewAI from early 2023 to mid-2025 (SuperAGI, 2025[59]). In parallel, emerging trends in GitHub repositories, including increasing adoption of the MCP and multi-agent orchestration strategies, highlight a significant shift toward agent-centric development paradigms (Ruiz, 2025[60]).

The latest Stack Overflow Developer Survey asked respondents a variety of questions about AI agents. The survey received more than 49000 responses from 177 countries and covered 62 questions in total. It defines AI agents as "autonomous software entities that can operate with minimal to no direct human intervention using artificial intelligence techniques" (Stack Overflow, 2025[61]). The answers provided useful insight into how developers use and think about AI agents.

Survey results show that about half of respondents are already using, or plan to use, AI agents in their work, while 38% have no plans to adopt them (Figure 4.1). At the same time, the vast majority of developers highlight opportunities to further strengthen the security, privacy, and accuracy of AI agents, underscoring the importance of continued progress in building more trustworthy AI agents (Figure 4.2).

Figure 4.1. Half of developers on Stack Overflow plan to use AI agents; 38% remain unswayed

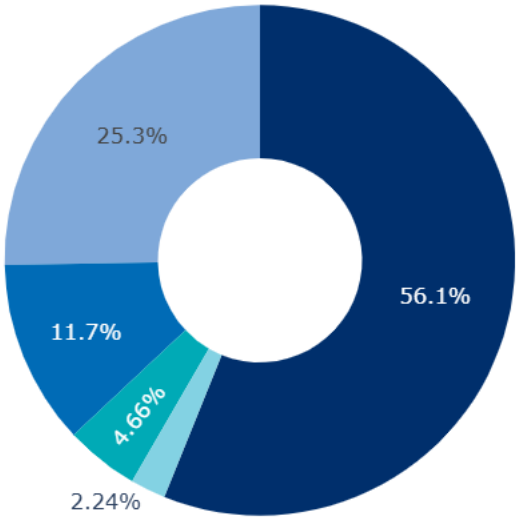
Answers to the question: "Are you using AI agents in your work (development or otherwise)?"



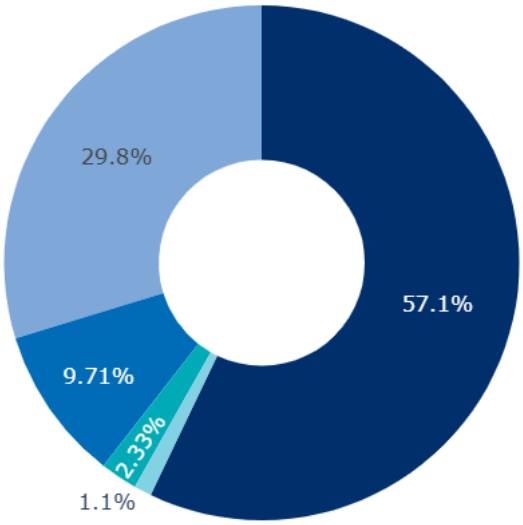
Note: Calculated as a percent of respondents that answered this question (n = 31 890).
Source: OECD elaboration based on data from Stack Overflow developer survey (2025[61]).

Figure 4.2. Vast majority of developers using AI agents concerned over privacy, security and accuracy

a) Answer "I have concerns about the security and privacy of data when using AI agents"



b) Answer "I am concerned about the accuracy of the information provided by AI agents"

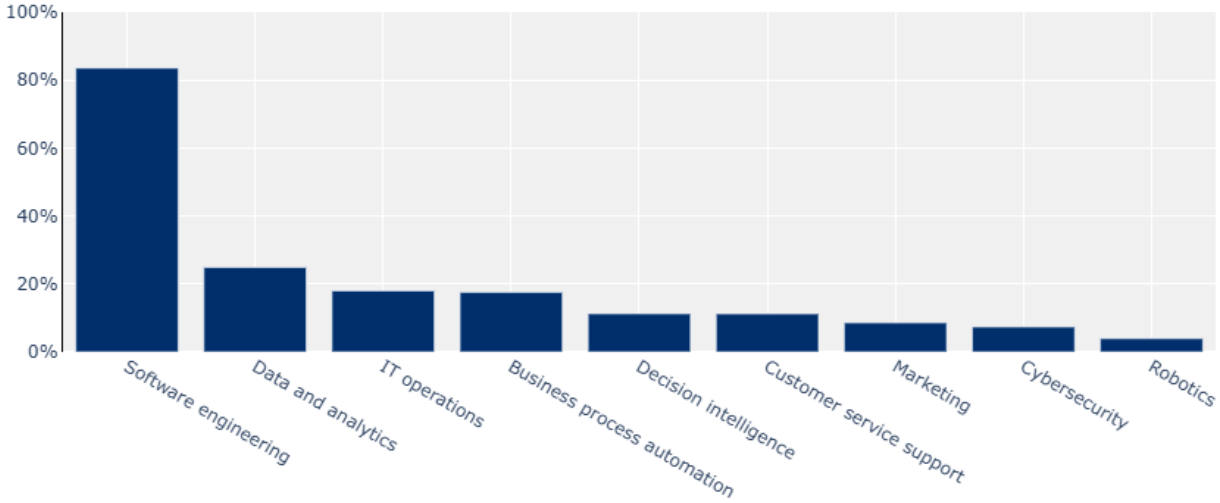


Note: Calculated as a percent of respondents that answered these question (panel a, *n* = 28 443; panel b, *n* = 28 826).
Source: OECD elaboration based on data from Stack Overflow developer survey (2025[61]).

Developers are primarily using agents for software engineering purposes (Figure 4.3). However, 64% of respondents identifying as either a data scientist, engineer, or analyst are using agents primarily for data and analytics.

Figure 4.3. Software engineering most common use for AI agents among developers

Answers to the question: "What industry purposes are you using AI agents in your development work?"



Note: Calculated as a percentage of respondents who reported using AI agent tools in their workflow and who also answered this question (*n* = 12 307).
Source: OECD elaboration based on data from Stack Overflow developer survey (2025[61]).

Developers are also using a variety of tools for AI agent orchestration, monitoring, security, and development. While most still rely heavily on general-purpose LLM tools like ChatGPT, Github Copilot, and Google Gemini, several dedicated AI agent tools are also gaining traction. Table 4.1 summarises the categories and uses of the tools mentioned by developers.

Table 4.1. Most common tools related to AI agents by use case

Use case	Most popular tools
AI agent memory or data management	Redis, Github MCP Server, Supabase, Chromadb
AI agent orchestration or agent frameworks	Ollama, Langchain, Langgraph, Vertex AI, Amazon Bedrock Agents
AI agent observability, monitoring or security	Grafana + Prometheus, Sentry, Snyk, New Relic, Langsmith
Out-of-the-box agents, copilots or assistants	ChatGPT, Github Copilot, Google Gemini, Claude Code, Microsoft Copilot

Source: OECD analysis based on data from Stack Overflow developer survey. AI agent orchestration tools include Ollama, LangChain, and LangGraph (2025[61]).

While these findings offer valuable insights, they should be considered indicative rather than exhaustive. Limited data on AI agent adoption and use constrain the evidence base, and the sources cited cover only a portion of the landscape, sometimes relying on self-reported information. As a result, they may not fully capture all economies, developer communities, or proprietary developments.

5 Discussion

This paper analysed definitions of AI agents and agentic AI from academic and technical sources to identify core concepts and taxonomies. Using the OECD AI system definition as an analytical framework, it highlighted both common features and points of divergence across definitions. This conceptual analysis is expected to be complemented by illustrative examples of how agentic AI systems are being developed and used in practice.

The analysis suggests that agentic AI most often refers to systems that integrate and co-ordinate multiple AI agents. While agentic AI systems are composed of AI agents, not all AI agents are part of an agentic AI system. Individual agents operating in isolation – without broader system-level orchestration – are generally not considered agentic AI under most definitions.

Agency exists on a spectrum: from reactive agents that simply respond to stimuli, through agent-assisted workflows such as "copilot" systems that support discrete tasks, to agentic AI systems that co-ordinate multiple agents and manage entire workflows with minimal human oversight.

As agentic AI systems become more capable and widely deployed, several areas merit further exploration to support effective policymaking. First, there is a need for greater clarity on the different architectures underpinning agentic AI and its technical stack (Dilmegani and Palazoglu, 2025[62]; Chaudhary, 2025[63]; Zhai et al., 2025[64]). Mapping these architectures can help identify where safeguards, standards, or oversight mechanisms may be most effective to support trustworthy innovation.

Second, the development of relevant typologies of agentic AI could support policy development and responsible deployments. Such typologies might distinguish systems by domain of application, level of autonomy, adaptiveness, tool access levels, or capacity to influence their physical or virtual environment (Stryker, 2025[65]; Chawla, 2025[66]; Srikumar, 2025[67]; Partnership on AI, 2025[29]; Pant and Viswanathan, 2025[50]).

Further work is also needed to assess the broader policy implications of agentic AI. Key issues include accountability, explainability, and transparency; system behaviour and performance within complex socio-technical environments; the protection of information integrity and human rights, including freedom of expression; and challenges related to global disparities, sovereignty, and resource efficiency, including water and energy use (Partnership on AI, 2025[29]; Partnership on AI, 2025[68]; Information Commissioner's Office, 2026[51]). In parallel, identifying systematic approaches that support the safe and trustworthy development and deployment of agentic AI – such as continuous monitoring, evaluation, and benchmarking – could help inform policymakers and complement emerging international efforts in this area (UK AI Security Institute, 2025[69]; Infocomm Media Development Authority, 2026[47]).

Future work could also prioritise the development of robust, cross-country indicators to track the development and adoption of AI agents. This could include analysing trends in job postings, reported incidents, investment activity, and research outputs that explicitly reference the development, deployment, or use of AI agents.

References

AAAI (2025), *AAAI 2025 presidential panel on the future of AI research*, https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf. [16]

Anthropic (2024), *Building effective agents*, https://www.anthropic.com/engineering/building-effective-agents (accessed on 4 February 2026). [24]

Anthropic (2024), *Introducing the Model Context Protocol*, https://www.anthropic.com/news/model-context-protocol (accessed on 4 February 2026). [53]

Ashby, D. (2025), *The agentic loop: Reimagining the future of QA*, https://www.functionize.com/blog/the-agentic-loop-reimagining-the-future-of-qa (accessed on 4 February 2026). [38]

Bengio, Y. et al. (2025), *International AI safety report: First key update, capabilities and risk implication*, https://internationalaisafetyreport.org/. [26]

Capgemini Research Institute (2025), *Rise of agentic AI: How trust is the key to human-AI collaboration*, https://www.capgemini.com/insights/research-library/ai-agents/. [37]

Challapally, A. et al. (2025), "The genAI divide: State of AI in business 2025", *MIT NANDA*, https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf. [41]

Chan, A. et al. (2023), "Harms from increasingly agentic algorithmic systems", *arXiv*, https://doi.org/10.48550/arXiv.2302.10329. [43]

Chan, A. et al. (2025), "Infrastructure for AI agents", *arXiv*, https://doi.org/10.48550/arXiv.2501.10114. [27]

Chaudhary, A. (2025), *7 layers again: From OSI to AI agents—An architecture reborn*, https://ai-with-aj.com/2025/07/10/7-layers-again-from-osi-to-ai-agents-an-architecture-reborn/ (accessed on 4 February 2026). [63]

Chawla, A. (2025), *5 levels of agentic AI systems*, https://www.dailydoseofds.com/p/5-levels-of-agentic-ai-systems/#:~:text=...,human%20guides%20the%20entire%20flow (accessed on 4 February 2026). [66]

CSET (2024), *Through the Chat Window and Into the Real World*, Center for Security and Emerging Technology, https://cset.georgetown.edu/wp-content/uploads/CSET-Through-the-Chat-Window-and-Into-the-Real-World.pdf. [42]

Dignum, V. and F. Dignum (2025), "Agentifying agentic AI", *arXiv*, https://doi.org/10.48550/arXiv.2511.17332. [45]

Dignum, V. and F. Dignum (2020), "Agents are dead. Long live agents!", *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1701–1705, https://dl.acm.org/doi/abs/10.5555/3398761.3398957. [7]

Dilmegani, C. and M. Palazoglu (2025), *The 7 Layers of agentic AI stack*, https://research.aimultiple.com/agentic-ai-stack/ (accessed on 5 January 2026). [62]

Fadel, C. (2025), *AI in 2025: A combinatorial explosion of possibilities, but NOT AGI*, https://curriculumredesign.org/wp-content/uploads/AI-in-2025-A-Combinatorial-Explosion-of-Possibilities-but-NOT-AGI-CCR.pdf. [33]

Ferber, J. (1999), *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence (1st ed.)*, Addison-Wesley. [28]

Floridi, L. (2024), *AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5135645. [6]

Google (n.d.), *Google Trends*, https://trends.google.com/trends/ (accessed on 1 December 2025). [8]

Gosmar, D. et al. (2024), "AI multi-agent interoperability extension for managing multiparty conversations", *arXiv*, https://doi.org/10.48550/arXiv.2411.05828. [54]

Hammond, L. et al. (2025), "Multi-agent risks from advanced AI", *arXiv*, https://doi.org/10.48550/arXiv.2502.14143. [56]

Hugging Face (2025), *What are agents?*, https://huggingface.co/docs/smolagents/conceptual_guides/intro_agents (accessed on 16 October 2025). [39]

Hugging Face (2025), *What is an agent?*, https://huggingface.co/learn/agents-course/unit1/what-are-agents (accessed on 16 October 2025). [30]

IBM (2025), *AI agents: Opportunities, risks, and mitigations*, https://www.ibm.com/granite/docs/resources/ai-agents-opportunities-risks-and-mitigations.pdf. [20]

Infocomm Media Development Authority (2026), *Model AI governance framework for agentic AI*, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf. [47]

Information Commissioner's Office (2026), *ICO tech futures: Agentic AI*, https://ico.org.uk/about-the-ico/research-reports-impact-and-evaluation/research-and-reports/technology-and-innovation/tech-horizons-and-ico-tech-futures/ico-tech-futures-agentic-ai/. [51]

Kashyap, P. (2025), "A practical guide: Understanding AI copilots and agents", *Conversive*, https://www.beconversive.com/blog/ai-copilot-vs-ai-agent (accessed on 4 February 2026). [18]

Kasirzadeh, A. and I. Gabriel (2025), "Characterizing AI agents for alignment and governance", *arXiv*, https://doi.org/10.48550/arXiv.2504.21848. [23]

Kim, J. et al. (2025), "The cost of dynamic reasoning: Demystifying AI agents and test-time scaling from an AI infrastructure perspective", *arXiv*, [48]

https://doi.org/10.48550/arXiv.2506.04301.

LangChain Docs (2025), *Agents*, https://docs.langchain.com/oss/python/langchain/agents (accessed on 16 October 2025). [31]

Mas-Colell, A., M. Whinston and J. Green (1995), *Microeconomic Theory*, Oxford University Press. [10]

Masterman, T. et al. (2024), "The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey", *arXiv*, https://doi.org/10.48550/arXiv.2404.11584. [36]

MCP (2025), *What is the Model Context Protocol (MCP)?*, https://modelcontextprotocol.io/docs/getting-started/intro (accessed on December 2025). [57]

Miehling, E. et al. (2025), "Agentic AI needs a systems theory", *arXiv*, https://doi.org/10.48550/arXiv.2503.00237. [40]

Mitchell, M. et al. (2025), "Fully autonomous AI agents should not be developed", *arXiv*, https://doi.org/10.48550/arXiv.2502.02649. [35]

Moltbook (2026), *Moltbook beta*, https://www.moltbook.com/ (accessed on 1 February 2026). [58]

Nguyen, D. et al. (2024), "GUI agents: A survey", *arXiv*, https://doi.org/10.48550/arXiv.2412.13501. [32]

NIST (2025), *SP 800-53 Control Overlays for Securing AI Systems Concept Paper*, https://csrc.nist.gov/csrc/media/Projects/cosais/documents/NIST-Overlays-SecuringAI-concept-paper.pdf. [21]

OECD (2024), "Explanatory memorandum on the updated OECD definition of an AI system", *OECD Artificial Intelligence Papers*, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/623da898-en. [3]

OECD (2022), "OECD framework for the classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [4]

OECD (n.d.), "Recommendation of the Council on Artificial Intelligence", *OECD/LEGAL/0449*, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 (accessed on 4 February 2026). [2]

Oueslati, A. and R. Staes-Polet (2025), *Ahead of the curve: Governing AI agents under the EU AI Act*, https://thefuturesociety.org/wp-content/uploads/2023/04/Report-Ahead-of-the-Curve-Governing-AI-Agents-Under-the-EU-AI-Act-4-June-2025.pdf. [22]

Pant, S. and M. Viswanathan (2025), "5 levels of agentic AI intelligence for enterprise use", *Outshift*, https://outshift.cisco.com/blog/agentic-ai-intelligence-for-enterprise-use (accessed on 4 February 2026). [50]

Partnership on AI (2025), *AI agents and global governance*, https://partnershiponai.org/resource/ai-agents-global-governance-analyzing-foundational-legal-policy-and-accountability-tools/. [68]

Partnership on AI (2025), *Preparing for AI agent governance*, https://partnershiponai.org/resource/preparing-for-ai-agent-governance/. [29]

Rao, A. and M. Georgeff (1995), "BDI Agents: From theory to practice", *Proceedings of the First International Conference on Multiagent Systems*, https://cdn.aaai.org/ICMAS/1995/ICMAS95-042.pdf. [14]

Ruiz, J. (2025), "From MCP to multi-agents: The top 10 new open source AI projects on GitHub right now and why they matter", *GitHub*, https://github.blog/open-source/maintainers/from-mcp-to-multi-agents-the-top-10-open-source-ai-projects-on-github-right-now-and-why-they-matter/ (accessed on 4 February 2026). [60]

Russell, S. and P. Norvig (2022), *Artificial Intelligence: A Modern Approach (4th ed.)*, Pearson. [13]

Russell, S. and P. Norvig (1995), *Artificial Intelligence: A Modern Approach (1st ed.)*, Prentice Hall. [11]

Samdani, G., G. Viswanathan and A. Jegadeesh (2025), *Human-AI collaboration: Balancing agentic AI and autonomy in hybrid systems*, https://aircconline.com/ijccsa/V15N1/15125ijccsa01.pdf. [34]

Sapkota, R., K. Roumeliotis and M. Karkee (2025), "AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges", *arXiv*, https://doi.org/10.48550/arXiv.2505.10468. [25]

Shavit, Y. et al. (2023), *Practices for governing agentic AI systems*, https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf. [44]

Srikumar, M. (2025), *Prioritizing real-time failure detection in AI agents*, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/. [67]

Stack Overflow (2025), *Stack Overflow annual developer survey*, https://survey.stackoverflow.co/. [61]

Stanford Encyclopedia of Philosophy (2015), *Agency*, https://plato.stanford.edu/entries/agency/ (accessed on 4 February 2026). [9]

Stryker, C. (2025), "Types of AI agents", *IBM*, https://www.ibm.com/think/topics/ai-agent-types (accessed on 4 February 2026). [65]

SuperAGI (2025), *Future-proofing your AI: Trends and innovations in open-source agentic frameworks for 2025 and beyond*, https://superagi.com/future-proofing-your-ai-trends-and-innovations-in-open-source-agentic-frameworks-for-2025-and-beyond/ (accessed on 4 February 2026). [59]

Surapaneni, R. et al. (2025), *Announcing the Agent2Agent protocol (A2A)*, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/?utm_source=www.dailyzaps.com&utm_medium=newsletter&utm_campaign=chatgpt-now-remembers-everything-you-tell-it&_bhlid=4dced5b40e6e3129347043d41cfce8fc1e352e22. [55]

Tallam, K. (2025), "Alignment, agency and autonomy in frontier AI: A systems engineering perspective", *arXiv*, https://doi.org/10.48550/arXiv.2503.05748. [5]

Teahan, W. (2014), *Artificial Intelligence - Agent behaviour I*, https://www.peterfisk.com/wp-content/uploads/2019/02/artificial-intelligence-agent-behaviour-i.pdf. [12]

Trivedi, R. et al. (2024), *Altared environments: The role of normative infrastructure in AI* [52]

alignment, https://openreview.net/forum?id=Gd6QrBLHBN.

UK AI Security Institute (2025), *International joint testing exercise: Agentic testing*, https://www.aisi.gov.uk/blog/international-joint-testing-exercise-agentic-testing (accessed on 4 February 2026). [69]

Weiss, G. (2013), *Multiagent Systems (2nd ed.)*, MIT Press. [15]

Wooldridge, M. (2002), *An Introduction to Multiagent Systems (1st ed.)*, John Wiley & Sons. [17]

Yin, X. et al. (2024), "Gödel agent: A self-referential agent framework for recursive self-improvement", *arXiv*, https://doi.org/10.48550/arXiv.2410.04444. [19]

Yousefi, Y., M. Billi and A. Rotolo (2025), *Agentic AI: An EU AI Act paradigm shift?*, http://dx.doi.org/10.2139/ssrn.5731424. [46]

Zeff, M. and K. Wiggers (2025), "No one knows what the hell an AI agent is", *TechCrunch*, https://techcrunch.com/2025/03/14/no-one-knows-what-the-hell-an-ai-agent-is/ (accessed on 4 February 2026). [1]

Zhai, L. et al. (2025), "The athenian academy: A seven-layer architecture model for multi-agent systems", *arXiv*, https://doi.org/10.48550/arXiv.2504.12735. [64]

Zhao, W. et al. (2025), "SiriuS: Self-improving multi-agent systems via bootstrapped reasoning", *arXiv*, https://doi.org/10.48550/arXiv.2502.04780. [49]