AI Agents

What They Are, What They Are Not, and the Boundaries of Current Understanding

(AIGuide 5)

A Structured Clarification of Meaning, Boundaries, and Current Capabilities in Global AI Discourse (Early 2026)

AI Sourced Facts (AISF) Pte. Ltd.

Edition: Version 1.0
Publication Date: February 2026

PREFACE

The term “AI Agent” has rapidly entered global discourse, appearing across technical, commercial, and public-facing contexts. It is used to describe a wide range of systems, often with differing characteristics, capabilities, and levels of complexity. Despite its growing prominence, the meaning of the term remains fluid.

This guide does not seek to define “AI Agent” in a fixed or prescriptive manner. Instead, it examines how the term is currently used, interpreted, and applied across multiple AI systems and contexts in early 2026.

The objective is to provide clarity without imposing certainty.

Through structured observation, this guide identifies recurring patterns, distinguishes between core and associated characteristics, and examines how the concept interacts with adjacent categories such as tools, chatbots, and automation workflows. It also situates the term within real-world usage, structural constraints, and the broader dynamics of language and interpretation.

The result is not a definitive definition, but a structured understanding of how meaning is formed, extended, and varied in current discourse.

This guide is intended for readers seeking clarity in a rapidly evolving space, where terms are widely used but not always consistently understood.

PURPOSE OF THE GUIDE

This guide provides a structured, time-bound clarification of the term “AI Agent” as it appears in global AI discourse in early 2026.

It is designed to:

• examine how the term is currently used across systems and contexts
• identify patterns of convergence and divergence in interpretation
• distinguish necessary characteristics from optional or associated features
• clarify boundaries between agents and adjacent concepts
• document observed usage, limitations, and structural conditions

This guide does not:

• define a single authoritative meaning
• provide implementation guidance
• recommend specific approaches or tools
• establish regulatory or technical standards

Its purpose is to:

structure understanding without imposing definition

SOURCE METHOD NOTE

This guide is based on structured analysis of outputs from multiple globally deployed AI systems.

These systems were queried using consistent prompts to elicit definitions, characteristics, and interpretations of “AI Agents.”

The resulting outputs were:

- compared across systems
- analysed for convergence and divergence
- synthesised into structured observations

No single system is treated as authoritative.

This guide reflects cross-system patterns, not individual model viewpoints.

GLOBAL PERSPECTIVE / NEUTRALITY STATEMENT

Artificial intelligence is a global phenomenon, shaped by diverse technological, economic, and cultural contexts.

This guide is written from a neutral, globally oriented perspective.

It does not:

- prioritise any region, country, or system
- promote any vendor, platform, or implementation
- reflect a single institutional or technical viewpoint

Instead, it presents:

- cross-system observations
- globally observable patterns
- context-independent analysis

The intention is to ensure that the content remains relevant across geographies, sectors, and levels of technical familiarity.

EDITORIAL STANDARD

This guide follows a consistent editorial principle:

describe what is observed, without extending into prescription or definitive interpretation.

To maintain this standard:

• no single definition is imposed
• variation is preserved where it exists
• ambiguity is recognised as structural
• language remains neutral and non-directive
• conclusions are avoided where evidence does not support them

The guide is structured to support clarity without reducing complexity.

HOW TO USE THIS GUIDE

This guide is structured as a sequence of interrelated chapters, each examining a different aspect of the concept of "AI Agents."

Readers may:

- read sequentially to build a layered understanding
- focus on specific chapters depending on interest
- use individual sections as reference points

The guide progresses through:

- emergence and meaning
- structural characteristics
- relationships to adjacent concepts
- observed usage patterns
- limitations and constraints
- autonomy and control
- ambiguity and variability
- near-term trajectory
- interpretive synthesis and open questions

It is intended to be:

- informative without being prescriptive
- structured without being restrictive

VERSION & GOVERNANCE

This guide is published as part of the AISF Education series.

Version 1.0
February 2026

This is a time-bound publication reflecting the state of global AI discourse at the time of writing.

Future editions may:

• refine observations as usage evolves
• incorporate additional patterns of interpretation
• adjust structural framing as convergence or divergence develops

All updates will be version-controlled.

AISF maintains editorial governance to ensure neutrality, consistency, and structural integrity.

AISF UNIVERSAL DISCLAIMER

This publication is provided for informational and educational purposes only.

It presents a structured and time-bound analysis based on information available at the time of writing. The content reflects observed patterns, interpretations, and perspectives and does not constitute definitive, authoritative, or universally accepted conclusions.

This publication does not provide:

- legal, regulatory, or compliance advice
- financial, investment, or business advice
- technical implementation guidance
- operational or strategic recommendations

No representation or warranty is made as to the accuracy, completeness, or suitability of the information for any specific purpose.

Artificial intelligence systems, technologies, and related terminology evolve rapidly. As such, the observations and interpretations contained in this publication may change over time and may not reflect subsequent developments.

AI Sourced Facts (AISF) does not endorse, promote, or recommend any specific technologies, systems, vendors, products, or approaches.

The use of any information, concepts, or interpretations from this publication is at the reader's own discretion and risk.

AISF publications are intended to support informed human judgment. Final responsibility for decisions, actions, and outcomes remains with the user.

Table of Contents

Chapter 1 — Introduction — Why "AI Agents" Requires Clarification

The term "AI Agents" has become increasingly visible across public discussion, technical commentary, enterprise messaging, and general conversation about artificial intelligence. In a relatively short period, it has moved from a more limited and specialised usage into broader circulation, where it is now applied to a wide range of systems, features, and software behaviours. In this wider usage, the term often appears to carry an intuitive but loosely defined meaning. It suggests action rather than response, initiative rather than passivity, and a degree of independent task completion rather than mere output generation. Yet once examined more closely, the term becomes less stable than its confident use might imply.

This instability is one of the main reasons clarification is needed. In current discourse, "AI Agent" does not function as a uniformly defined category. Rather, it operates as a contested and elastic label. Across systems, there is broad agreement that an agent is something more than a static software tool and something different from a purely conversational interface. At the same time, there is noticeable variation in how far that difference is thought to extend, which characteristics are treated as essential, and where the boundary lies between a genuine agent and a system that is simply being described in more ambitious language.

The need for clarification is therefore not driven by a lack of interest in the topic, but by the opposite condition: the term is now widely used in circumstances where its meaning is assumed rather than carefully established. In some contexts, "AI Agent" refers to a multi-step system that can interpret a goal, select tools, and act iteratively toward an outcome. In other contexts, the same term is used for a chatbot that can call a function once, or for an automation workflow that follows pre-written rules. The result is not merely terminological untidiness. It affects how capabilities are understood, how limitations are judged, and how claims about present-day AI systems are interpreted.

This guide begins from the observation that ambiguity does not disappear simply because a term becomes popular. On the contrary, popularity can intensify ambiguity by encouraging overextension. As the term becomes commercially useful and rhetorically persuasive, it is more readily applied to systems that share only partial similarities. Across multiple system outputs, there is explicit recognition that the term is undergoing semantic dilution, where features such as tool-calling or workflow integration are treated as sufficient to justify the label, even when the underlying system remains reactive or tightly structured.

Clarification is also needed because the term sits at the intersection of several neighbouring concepts that are easier to conflate than to separate. A standard AI tool may generate outputs in response to a prompt. A chatbot may sustain dialogue and retain limited conversational context. An automation workflow may execute predefined steps when triggered. An agent is often described as distinct from all three, yet the basis of that distinction varies. Sometimes it is framed as the ability to act. Sometimes it is framed as the ability to plan. Sometimes it depends on whether the system determines its own next step rather than waiting for instruction. These distinctions overlap, but they are not identical, and their interaction contributes to the present conceptual uncertainty.

The question is therefore not whether the term has meaning, but what kind of meaning it currently has. It would be misleading to approach the subject as though a single settled

definition is waiting to be extracted and defended. The cross-system material does not support that conclusion. What it supports instead is a pattern of partial convergence. Certain themes recur with notable consistency, including goal-directed behaviour, interaction with external tools or environments, iterative progression across multiple steps, and some capacity to adjust behaviour based on intermediate results. Yet even where these themes recur, differences remain over whether all are strictly necessary, whether some are optional refinements, and how particular boundary cases should be treated.

A further and more consequential source of ambiguity arises in the way the term “autonomy” is attached to agents. In current discourse, “AI Agents” are frequently described as autonomous. This description often carries more weight in general language than in technical reality. Across the surveyed material, there are repeated efforts to qualify this autonomy and to narrow its practical meaning. Current implementations are commonly described not as fully independent systems, but as bounded systems capable of making limited procedural decisions within human-defined constraints. They may sequence tasks, select from approved tools, or recover from certain errors, but they do not set their own objectives, define their own operating scope, or assume responsibility for outcomes. One formulation describes this condition as “dependent autonomy,” capturing a recurring distinction: systems may act autonomously within a task while remaining dependent on human goals, permissions, and oversight.

This distinction matters because inflated interpretations of autonomy can distort the understanding of current capabilities. If an agent is imagined as an independently operating entity, limitations may be overlooked. If it is understood instead as a system capable of bounded reasoning and action within configured constraints, its role becomes easier to describe without exaggeration. Clarification, in this sense, restores proportion between language and capability.

The guide also requires clarification because the subject is evolving, but not in a way that justifies imprecision. Early 2026 is a period in which the language around agents is expanding faster than stable consensus. New features, demonstrations, and deployment patterns are contributing to the term’s visibility, yet visibility alone does not create conceptual clarity. It is therefore appropriate to treat this moment as a snapshot of interpretation rather than a final settlement. The aim is to identify what can be stated with reasonable confidence about current usage, where variation persists, and which distinctions remain most useful for avoiding confusion.

This is why the guide is structured as a clarification rather than a proclamation. Its purpose is not to impose a single authoritative definition on a fluid field, nor to stabilise language by force. Terminology in this guide reflects observed usage rather than enforced standardisation. The task is to map how the term is currently used across multiple systems, identify recurring patterns, and separate stronger claims from weaker ones while maintaining clear conceptual boundaries.

That boundary-setting function is especially important because “AI Agents” currently attracts both analytical interest and rhetorical overreach. In practice, the same label can be applied to productivity assistants, software systems, customer-service processes, multi-step research tools, or orchestration layers interacting with external services. Some uses appear consistent with a broader interpretation of agency, while others rely more on suggestive language than

on a stable conceptual threshold. The task is not to eliminate ambiguity, but to distinguish meaningful variation from conceptual drift.

Accordingly, the chapters that follow begin with framing before moving into boundary clarification. They examine recurring definition patterns, distinguish necessary from optional characteristics, and clarify differences between agents and adjacent system types. They also address autonomy as a spectrum, the limits of current implementations, and the reasons human oversight remains structurally central. Throughout, the objective remains consistent: to structure the concept clearly without overstating certainty.

Chapter 2 — The Emergence of the Term in Global AI Discourse

The term “AI Agent” did not emerge in a single moment or from a single source. Its current prominence reflects a gradual convergence of ideas from multiple strands of artificial intelligence research, software development practices, and evolving user expectations. What is notable in early 2026 is not the novelty of the underlying concept, but the rapid expansion of the term’s usage beyond its earlier, more contained contexts.

Historically, the concept of an “agent” has long existed within the academic field of artificial intelligence. In earlier formulations, an agent was typically described in abstract terms as an entity that perceives its environment and takes actions to achieve a goal. This framing was deliberately broad, allowing it to encompass a wide range of systems, from simple rule-based mechanisms to more complex adaptive models. However, this earlier usage remained largely within technical and academic discourse, where its meaning was supported by shared theoretical context.

The current use of the term reflects a shift from that contained environment into broader technological and commercial language. This shift has not been accompanied by a corresponding consolidation of meaning. Instead, the term has expanded to accommodate a variety of interpretations, many of which are shaped by practical implementation patterns rather than theoretical definitions. As a result, the modern usage of “AI Agent” sits between its academic origins and its applied interpretations, without being fully anchored to either.

Across the surveyed systems, there is observable agreement that the term has gained traction alongside the increasing capability of systems to perform multi-step tasks. The transition from single-response systems to systems capable of chaining actions appears to have contributed significantly to the adoption of the term. Where earlier systems were primarily evaluated on their ability to generate outputs, more recent systems are described in terms of their ability to carry out processes. This shift in emphasis from output to process is one of the underlying drivers of the term’s wider usage.

At the same time, the emergence of tool integration has played a notable role in shaping how the term is understood. Systems that can interact with external functions, data sources, or services are frequently described as exhibiting agent-like behaviour. The ability to move beyond isolated computation and into interaction with an environment is often treated as a defining feature. However, as reflected in multiple system outputs, this association is not consistently applied. Some interpretations treat tool use as sufficient to justify the label, while others regard it as necessary but not sufficient, requiring additional elements such as planning, iteration, and goal persistence.

The expansion of the term has also been influenced by the increasing accessibility of AI systems. As interaction with such systems becomes more widespread, the language used to describe them adapts to more general expectations. Terms that were once reserved for technical description begin to take on broader meanings that align with everyday intuition. In this context, “agent” becomes a convenient way to describe systems that appear to “do things” rather than simply “say things.” This intuitive distinction between doing and responding appears repeatedly across system explanations, even where the underlying mechanisms differ significantly.

A further observable pattern is the role of narrative framing in the term's emergence. The language surrounding AI development often incorporates metaphors drawn from human roles, such as assistants, workers, or collaborators. The term "agent" fits naturally within this pattern, as it suggests delegated responsibility and purposeful action. While such framing can aid accessibility, it also contributes to the broadening of the term's meaning. Systems that share only some surface similarities with these metaphors may nonetheless be described using the same terminology, reinforcing the expansion of the label beyond its more precise boundaries.

The emergence of the term is therefore not simply a technical development, but a linguistic and conceptual one. It reflects the interaction between evolving system capabilities and the ways in which those capabilities are communicated and interpreted. In this interaction, precision is not always preserved. Instead, the term absorbs multiple layers of meaning, some grounded in observable behaviour and others shaped by expectation or analogy.

Across systems, there is also recognition that the term is being used at different levels of abstraction. In some cases, it refers to an architectural pattern involving iterative reasoning and action loops. In other cases, it is used more loosely to describe any system that performs a task with limited user intervention. These different levels of abstraction coexist without a clear boundary between them, contributing further to variation in usage. What one system classifies as an agent, another may classify as an enhanced tool or a structured workflow.

This variation is not necessarily contradictory, but it does indicate that the term operates across a spectrum rather than within a fixed category. The emergence of "AI Agents" as a widely used term can therefore be understood as a process of conceptual layering. Earlier definitions provide a foundational structure, while newer interpretations extend that structure in ways that are not always consistent with one another.

It is therefore appropriate to approach the term not as a stable classification that has recently been discovered, but as a developing construct whose boundaries are still being negotiated across different contexts. The observed patterns suggest that while certain core ideas are gaining traction, the surrounding interpretation remains fluid. This fluidity is not incidental; it is part of the term's current state.

In this context, the emergence of "AI Agents" in global AI discourse can be seen as both a reflection of genuine shifts in system capability and an example of how language adapts to those shifts. The two processes are intertwined. As systems become more capable of carrying out sequences of actions, the need for language that captures this behaviour increases. At the same time, the adoption of such language introduces new ambiguities that require clarification.

The purpose of this chapter is not to resolve those ambiguities, but to situate them. By tracing how the term has moved from a more defined academic concept into a broader and more varied usage, it becomes possible to understand why its meaning now appears both familiar and uncertain. This understanding provides the necessary context for the chapters that follow, which examine the structure, boundaries, and limitations of the term in greater detail.

Chapter 3 — Observational Boundaries of Meaning (Early 2026 Context)

The term “AI Agent,” as observed across multiple systems in early 2026, does not present itself as a fixed or uniformly bounded concept. Instead, it occupies a definitional space shaped by recurring patterns, partial agreements, and unresolved edges. These boundaries are not formalised through standards or universally accepted criteria. Rather, they emerge through usage—through how different systems describe capabilities, distinguish categories, and interpret the relationship between reasoning, action, and control.

This chapter does not seek to impose a precise boundary on the term. Instead, it identifies where boundaries appear to form, where they remain open, and where they become indistinct. The aim is to describe the observable limits of meaning as they currently exist, without resolving areas that remain structurally unsettled.

Across systems, there is a consistent tendency to associate the term “AI Agent” with systems that extend beyond single-step interaction. A baseline distinction is often made between systems that respond once to an input and systems that continue operating across multiple steps toward an outcome. This distinction does not always appear in identical language, but the underlying pattern is stable. An agent is generally understood to exhibit continuity of activity, where intermediate results influence subsequent actions.

This continuity introduces one of the first observable boundaries: the separation between **single-response systems** and **multi-step systems**. While this distinction is widely recognised, it does not by itself establish a definitive threshold. Some systems describe any multi-step behaviour as sufficient for agency, while others require additional characteristics such as internal decision-making or adaptive sequencing. The boundary, therefore, is not located at multi-step execution alone, but somewhere beyond it.

A second observable boundary relates to the source of decision-making. Across multiple interpretations, a distinction emerges between systems that follow externally defined sequences and systems that determine their own sequence of actions within a task. In the former case, the system executes steps that have been specified in advance, even if those steps involve multiple stages. In the latter, the system is described as selecting what to do next based on its interpretation of the task and the current state.

This distinction introduces a more structural boundary, often expressed in terms of **control over progression**. Systems that rely on pre-defined flows tend to be classified as automation, even when those flows are complex. Systems that exhibit internally generated progression—deciding how to move from one step to the next—are more consistently associated with agent-like behaviour. However, even here the boundary is not absolute. Some systems incorporate both elements, combining fixed structures with flexible decision points, making classification dependent on emphasis rather than strict categorisation.

A further boundary is observed in the relationship between reasoning and action. The term “AI Agent” is frequently applied where reasoning processes are linked to the ability to affect change beyond the generation of text or data. This linkage is often described as the ability to interact with an environment, whether through external tools, system interfaces, or other

mechanisms of execution. The presence of such interaction is widely treated as significant, yet its role in defining the boundary varies.

In some interpretations, the ability to act on external systems is treated as essential. In others, the capacity to produce actionable outputs—such as complete instructions or executable artefacts—is considered sufficient, even if direct execution is not permitted. This distinction reflects a broader ambiguity between **capability and permission**. A system may be capable of acting but configured not to act without approval. Whether such a system is considered an agent depends less on what it is allowed to do and more on what it is designed to be able to do.

The boundary between **capability and configuration** is therefore another defining feature of the current landscape. Systems are often evaluated not only on their observable behaviour, but on their underlying design intent. This introduces a layer of interpretation that can vary across contexts, contributing to differences in classification.

Memory provides another area where boundaries are both present and unsettled. Across systems, there is general recognition that some form of state awareness is necessary for multi-step activity. A system must, at minimum, maintain awareness of its current position within a task to avoid repetition or loss of progression. This form of short-term continuity is widely treated as integral to agent-like behaviour.

At the same time, longer-term memory—persistence of information across tasks or sessions—is not consistently treated as required. Some interpretations include it as an advanced feature, while others treat it as optional or unrelated to the core definition. The boundary, in this case, appears to separate **task-level continuity**, which is broadly accepted, from **cross-task persistence**, which remains optional. This distinction reinforces the pattern that not all frequently associated features are necessary for classification.

Goal orientation introduces a further boundary that appears more stable than others. Across systems, there is repeated emphasis on the presence of an objective that persists across multiple steps. The system is described not merely as responding to inputs, but as working toward an outcome that remains constant throughout the process. This persistence differentiates agent-like systems from reactive systems that treat each interaction independently.

However, even within this area of relative convergence, variation exists. Some systems emphasise explicit goals defined at the outset, while others allow for goals that evolve through interaction. Some treat failure recovery as essential to goal persistence, while others describe it as an enhancement. The boundary therefore holds conceptually, but its operational details remain variable.

A particularly significant boundary appears in the distinction between **orchestration and execution**. Across interpretations, there is recurring emphasis on the idea that an agent does not merely execute actions, but determines how and when those actions should occur. Execution, in this sense, is not sufficient to establish agency. What appears to matter more consistently is the presence of a coordinating function that integrates reasoning, sequencing, and action into a coherent process.

This distinction provides one of the more stable anchors in the current landscape. Systems that execute predefined steps, even if complex, are typically not described as agents. Systems that orchestrate their own sequence of actions are more consistently included within the term. However, as with other boundaries, hybrid cases exist where orchestration is partial or constrained, leading to variation in classification.

The boundary surrounding autonomy remains one of the most visible and yet most variable. While the term is frequently associated with autonomy, its meaning shifts depending on context. In practice, autonomy is often limited to **task-level procedural independence**, where the system can operate without step-by-step instruction once a goal is defined. This form of autonomy is bounded by constraints, permissions, and oversight mechanisms that remain externally defined.

The concept of **dependent autonomy** captures this condition. Systems may act independently within a defined scope, but they do not determine that scope themselves. They do not set their own objectives, expand their own authority, or operate beyond configured limits. This introduces a boundary between **local autonomy**, which is widely observed, and **global autonomy**, which is not supported by current implementations. The distinction is important because it separates the observable behaviour of systems from broader interpretations that may be inferred from the language used to describe them.

Beyond these more structured boundaries, there are also areas where meaning becomes diffuse. The term “AI Agent” is sometimes applied to systems that exhibit only one or two of the characteristics described above, particularly where those characteristics are perceived as significant. For example, the presence of tool use or the ability to perform a task without continuous input may be sufficient for the label to be applied, even if other elements such as planning or iterative reasoning are limited.

This selective emphasis contributes to what has been identified as semantic dilution. The term expands to include systems that partially resemble an agent without fully aligning with more structured interpretations. This does not necessarily invalidate those uses, but it does broaden the range of systems covered by the term, making its boundaries less distinct.

The result is a conceptual landscape characterised not by a single boundary, but by multiple overlapping thresholds. Some thresholds are more widely recognised, such as the distinction between reactive and goal-directed systems. Others are more context-dependent, such as the role of memory or the necessity of direct execution. Still others remain open, particularly where hybrid systems combine elements of different categories.

It is therefore possible to describe the meaning of “AI Agent” in early 2026 as bounded but not closed. Boundaries exist, but they are permeable. There are areas of convergence that provide structure, but there are also areas of divergence that prevent strict definition. The term functions less as a precise classification and more as a cluster of related characteristics, with different systems emphasising different parts of that cluster.

Understanding these observational boundaries is necessary for the chapters that follow. Without such an understanding, distinctions between agents and adjacent concepts risk being interpreted as arbitrary rather than grounded in observed patterns. By identifying where boundaries appear to hold and where they remain open, it becomes possible to examine the concept with greater clarity while maintaining the discipline of not overstating certainty.

In this sense, the boundaries described here are not final. They reflect the current state of usage, shaped by both capability and interpretation. As such, they provide a framework for understanding the term as it is presently used, while leaving open the possibility that those boundaries may continue to evolve.

Chapter 4 — Necessary Characteristics of AI Agents

The question of what constitutes a necessary characteristic of an AI Agent arises from the absence of a single agreed definition. Across systems, there is observable convergence around certain features, but this convergence does not always extend to strict agreement on thresholds. As a result, identifying necessary characteristics requires distinguishing between features that appear consistently as foundational and those that, while common, are treated as enhancements rather than requirements.

This chapter does not seek to formalise a definitive checklist. Instead, it identifies characteristics that, across multiple interpretations, are repeatedly treated as forming the minimum conditions under which the term "AI Agent" is meaningfully applied. These characteristics define what may be described as the **functional core** of agency in the current context.

A first and consistently observed characteristic is the presence of a **reasoning component capable of interpreting a goal**. Across systems, an agent is not described as acting randomly or merely reacting to fixed triggers. It is described as operating in relation to an objective, whether explicitly provided or inferred within a bounded interaction. This objective provides direction to the system's behaviour and distinguishes it from processes that operate without contextual interpretation.

The importance of this characteristic lies not in the complexity of reasoning, but in its role. The system must be able to translate an instruction into an internal representation that can guide subsequent steps. Without this translation, behaviour remains either deterministic or purely reactive. Systems lacking this capability are generally not described as agents, even when they perform useful functions.

Closely related to this is a second characteristic: **goal-directed progression across multiple steps**. An agent is not typically understood as completing a task in a single response. Instead, it is associated with processes that unfold over time, where intermediate outputs contribute to the continuation of the task. This progression is not merely sequential but directed. Each step is understood to relate to the overall objective, rather than existing as an isolated action.

This introduces a distinction between multi-step activity and goal-directed activity. While the two often coincide, they are not identical. A system may execute multiple steps without maintaining a coherent objective, particularly if those steps are predefined. In contrast, agent-like systems are described as maintaining alignment with a goal throughout the process, even when intermediate steps change or fail.

A third characteristic, frequently identified across systems, is the presence of an **iterative loop linking reasoning and action**. This loop reflects the system's ability to take a step, observe the result, and determine what to do next. The loop continues until a condition is met, whether that condition is successful completion, failure, or escalation for external input.

This iterative structure is central to how agency is described. It reflects a shift from static execution to dynamic adjustment. The system is not limited to a single cycle of input and output; it operates within a process that evolves based on its own intermediate outputs.

Without such a loop, behaviour tends to revert to either one-off response generation or fixed workflows, both of which are generally treated as distinct from agency.

A fourth characteristic concerns the system's ability to **determine its own sequence of actions within a task**. This characteristic introduces the distinction between externally defined workflows and internally generated progression. In agent-like systems, the next step is not always specified in advance by a human or a predefined script. Instead, the system evaluates its current state and selects an appropriate action.

This does not imply unrestricted freedom. The system operates within constraints defined by its design and configuration. However, within those constraints, it exhibits a degree of initiative in selecting how to proceed. This internal determination of sequence is one of the more consistently cited features distinguishing agents from automation systems, where the sequence of steps is fixed in advance.

A fifth characteristic, closely associated with the previous one, is the **ability to interact with an environment in a way that affects state**. Across systems, agents are not described as purely internal processors of information. They are described as capable of influencing external systems, whether through direct execution or through the generation of outputs that lead to execution.

The nature of this interaction varies. In some cases, it involves direct integration with tools, services, or data sources. In others, it involves producing complete outputs that can be acted upon externally. What appears to be necessary is not the form of interaction, but its presence. The system must be able to extend its operation beyond internal reasoning into some form of externally relevant action.

It is important to note that this characteristic relates to **capability rather than authority**. A system may be designed to act on external systems but configured to require confirmation before doing so. The presence of such constraints does not appear to remove its classification as an agent. The distinction lies in whether the system can propose or prepare actions as part of its process, rather than whether it is permitted to execute them autonomously.

A sixth characteristic, which appears with strong consistency, is **task-level state awareness**. This refers to the system's ability to maintain continuity within a given task. It must be able to recognise what has already been done, what remains to be done, and how intermediate results relate to the overall objective. Without this awareness, the system risks repetition, loss of context, or failure to progress.

This form of state awareness does not necessarily require persistent memory across sessions. Rather, it requires sufficient continuity within the scope of a task. The distinction between **short-term state** and **long-term memory** is significant here. The former appears to be necessary for agent-like behaviour, while the latter is more often treated as an optional enhancement.

A seventh characteristic, often closely linked to the others, is **goal persistence across intermediate variation**. An agent is described as maintaining its objective even when individual steps do not proceed as expected. If an attempt fails, the system does not simply stop; it adjusts its approach while retaining the original aim. This persistence differentiates agent-like systems from reactive systems that treat each step as independent.

The presence of goal persistence introduces the idea of conditional continuation. The system continues its operation until a termination condition is reached. This condition may be success, failure, or a decision to defer to external input. What appears necessary is the presence of a guiding objective that remains active throughout the process.

Taken together, these characteristics form a pattern rather than a rigid definition. They describe a system that:

• interprets an objective
• progresses toward that objective across multiple steps
• operates through an iterative reasoning–action loop
• determines its own sequence of actions within constraints
• interacts with an environment in a way that affects state
• maintains continuity within the task
• persists in pursuit of the objective despite intermediate variation

Across systems, when these characteristics are present in combination, the term "AI Agent" is more consistently applied. When one or more are absent, classification becomes less stable, and alternative labels such as tool, chatbot, or automation are more likely to be used.

It is also observable that not all characteristics carry equal weight. Some, such as goal-directed progression and iterative operation, appear more central than others. However, the relative importance of each characteristic is not always explicitly stated, and different systems emphasise different aspects. This variation reinforces the need to treat these characteristics as a cluster rather than as individually sufficient conditions.

The identification of necessary characteristics also highlights the distinction between **minimum viability** and **practical usefulness**. A system may meet the minimum conditions for being described as an agent while remaining limited in capability. Conversely, systems that incorporate additional features such as long-term memory or learning may extend beyond the minimum without altering the core classification. This distinction is relevant for understanding how the term is applied across systems with differing levels of sophistication.

It is therefore appropriate to interpret these characteristics as forming a **baseline threshold of agency** in early 2026. They do not define the outer limits of what an agent could become, nor do they eliminate ambiguity in borderline cases. Instead, they provide a structured way to recognise the conditions under which the term is most consistently used.

This baseline does not resolve all questions of classification. Hybrid systems, partial implementations, and evolving capabilities continue to challenge clear boundaries. However, by identifying the characteristics that most consistently appear across interpretations, it becomes possible to establish a working understanding of what is generally required for the term "AI Agent" to be applied in a meaningful way.

The next chapter builds on this foundation by examining characteristics that, while frequently associated with agents, are not consistently treated as necessary. This distinction between necessary and optional features further clarifies the boundaries of the concept without imposing a fixed definition.

Chapter 5 — Optional and Associated Characteristics of AI Agents

While the preceding chapter identified characteristics that appear consistently as forming a minimum threshold for agency, the broader discourse surrounding “AI Agents” includes a wider set of features that are frequently associated with the term but not uniformly treated as necessary. These features contribute to how agents are perceived, evaluated, and differentiated in practice, yet their presence is neither required nor consistently defined across systems.

This chapter examines those characteristics that are commonly observed but variably interpreted. The aim is not to diminish their importance, but to distinguish their role as **extensions of capability** rather than **conditions of classification**. In doing so, it becomes possible to separate what is essential for agency from what enhances or modifies its expression.

One of the most frequently referenced associated characteristics is **long-term memory**. Across many system descriptions, agents are described as remembering user preferences, past interactions, or prior outcomes. This persistence is often presented as contributing to continuity and personalisation. However, there is consistent variation in whether such memory is considered necessary.

In multiple interpretations, the distinction is made between **task-level state awareness**, which supports progression within a single objective, and **cross-task memory**, which extends beyond the immediate task. The former is widely treated as integral to agent-like behaviour, while the latter is treated as optional. Systems that lack long-term memory but maintain continuity within a task are still described as agents, whereas systems that incorporate memory across sessions are described as more advanced or specialised forms of agents.

A related characteristic is **learning across interactions**. Some descriptions include the idea that agents improve over time based on experience, adjusting their behaviour to become more effective. This form of learning is often associated with adaptation and optimisation. However, across systems, there is observable restraint in treating learning as a defining feature.

In practice, many systems described as agents do not exhibit continuous learning during operation. Instead, they operate within a fixed capability set, with improvements introduced externally through updates rather than through autonomous adaptation. As a result, learning is more accurately described as an associated capability that may be present in some implementations but is not required for the classification of a system as an agent.

Another commonly associated characteristic is **direct execution authority over external systems**. Some interpretations emphasise the ability of an agent to carry out actions without human confirmation, such as sending communications, initiating transactions, or modifying data. This capability is often linked to perceptions of autonomy.

However, across systems, there is consistent recognition that execution authority is subject to configuration rather than definition. Systems may be designed to perform actions but operate under constraints that require human approval. In such cases, the absence of direct execution

does not remove the classification of the system as an agent. The distinction, therefore, lies between **capability to act** and **permission to act**, with only the former appearing relevant to the core definition.

The characteristic of **breadth of tool integration** is also frequently associated with agents. Systems that can access a wide range of external tools, services, or data sources are often described as more capable or more agent-like. This breadth can include interaction with communication platforms, databases, computational tools, or external information sources.

Despite this association, the number or diversity of tools does not appear to determine whether a system is considered an agent. Systems with limited tool access may still meet the necessary characteristics if they demonstrate goal-directed progression and iterative behaviour. Conversely, systems with extensive tool integration may not be described as agents if their operation remains reactive or pre-scripted. Tool integration, therefore, functions as an amplifier of capability rather than a defining condition.

A further associated characteristic is **complexity of task decomposition**. Agents are often described as capable of breaking down high-level objectives into smaller, manageable steps. This ability contributes to their perceived usefulness in handling tasks that are not explicitly structured. However, the degree of decomposition varies across systems.

Some systems perform minimal decomposition, handling only simple multi-step processes, while others exhibit more elaborate planning behaviour. The presence of decomposition appears to support agency, but the level of sophistication is not uniformly required. Even limited forms of step structuring may be sufficient when combined with other core characteristics.

The concept of **self-correction or error recovery** is also commonly associated with agents. Systems are often described as detecting when an action has failed or produced an unexpected result and adjusting their behaviour accordingly. This capability contributes to continuity and resilience within a task.

Across interpretations, however, self-correction is not always treated as essential. Some systems may exhibit limited or inconsistent recovery behaviour while still being described as agents. The presence of self-correction enhances the robustness of the system but does not appear to define its classification.

Another characteristic that frequently appears in descriptions is **multi-agent interaction**. In some contexts, agents are described as operating alongside other agents, either cooperating or interacting within a shared environment. This arrangement introduces additional layers of coordination and complexity.

While multi-agent systems represent an extension of the concept, they are not necessary for defining a single agent. The existence of such systems demonstrates how the concept can scale, but it does not alter the baseline characteristics identified earlier. As such, multi-agent interaction is best understood as a structural variation rather than a defining feature.

The **natural language interface** is also often associated with agents, particularly in contexts where interaction occurs through conversational input. This association reflects the

accessibility of such interfaces and their role in making agent-like behaviour observable to users.

However, across systems, it is clear that natural language is not required. Agents may operate through structured inputs, programmatic interfaces, or embedded processes without direct human interaction. The presence of conversational capability may influence perception, but it does not determine classification.

Another associated characteristic is the **degree of autonomy exhibited within a task**. Systems that operate with minimal intervention are often perceived as more agent-like than those requiring frequent input. This perception is linked to the broader discussion of autonomy, where degrees of independence influence interpretation.

Despite this association, the degree of autonomy does not appear to define agency in a binary way. Systems that require confirmation at key stages may still be described as agents if they meet the underlying criteria of goal-directed progression and iterative operation. Autonomy, in this sense, functions as a spectrum that modifies the expression of agency rather than determining its existence.

A further characteristic sometimes associated with agents is **generality across domains**. Systems capable of operating in multiple contexts or handling diverse tasks are often described as more advanced forms of agents. This generality contrasts with systems designed for narrow, domain-specific tasks.

However, domain breadth is not consistently treated as necessary. Many systems described as agents operate within specific contexts, such as customer support, data analysis, or software development. The ability to generalise across domains may extend the scope of an agent, but it is not required for its classification.

These associated characteristics collectively illustrate how the concept of an AI Agent extends beyond its minimum threshold. They contribute to the richness and variability of the term, shaping how it is applied across different contexts. At the same time, their variability reinforces the importance of distinguishing between **core definition** and **capability enhancement**.

It is also observable that these characteristics often influence perception more than classification. Systems that exhibit multiple associated features—such as memory, tool integration, and self-correction—are more readily described as agents, even when the underlying classification criteria are not explicitly examined. Conversely, systems that meet the minimum conditions but lack additional features may be perceived as less agent-like, despite satisfying the core characteristics.

This distinction between perception and structure contributes to the ongoing variability in how the term is used. The presence of associated characteristics can create a gradient of "agent-likeness," where systems are informally ranked based on their capabilities. However, this gradient does not correspond to a formal boundary, and different systems may place emphasis on different aspects of this spectrum.

It is therefore appropriate to treat these characteristics as forming a **layer of optional capability** surrounding the core concept of agency. They expand what agents can do, how

they operate, and how they are experienced, but they do not redefine the minimum conditions under which the term applies.

Understanding this distinction is necessary for maintaining clarity in the use of the term. Without it, there is a tendency to conflate enhancement with definition, leading to either overextension or unnecessary restriction. By recognising that these characteristics are associated rather than essential, it becomes possible to describe variation without collapsing the concept into inconsistency.

The following chapters build on this distinction by examining how these characteristics interact with adjacent concepts and how they contribute to the broader landscape in which the term “AI Agent” is applied.

Chapter 6 — AI Agents vs Tools, Chatbots, and Automation

The term “AI Agent” gains much of its meaning not in isolation, but through contrast. Across systems, descriptions of agents frequently rely on distinguishing them from adjacent categories, particularly tools, chatbots, and automation workflows. These distinctions are not always expressed in identical terms, yet they form a consistent part of how the concept is understood. At the same time, the boundaries between these categories are not fixed. They overlap in capability, share underlying components, and increasingly appear in hybrid forms. This chapter examines how these distinctions are drawn in practice, and where they remain fluid.

A useful starting point is the category of **tools**. In general usage, an AI tool is described as a system that performs a specific function in response to an input. The relationship between user and system is typically direct: an instruction is given, and a result is returned. The tool may be highly capable within its domain, but its operation is bounded by the immediate interaction. It does not extend its activity beyond the scope of the request unless prompted to do so again.

The distinction between tools and agents is often framed in terms of **continuity and initiative**. Tools are described as responding to inputs, while agents are described as continuing activity toward an objective. This distinction is not based on intelligence alone. A tool may exhibit sophisticated reasoning within a single step, yet still be classified as a tool if it does not maintain progression across steps. Conversely, an agent may operate with relatively simple reasoning, but still be described as such if it sustains goal-directed activity through an iterative process.

This contrast introduces a structural difference between **single-step execution** and **multi-step progression**. However, as observed in earlier chapters, multi-step behaviour alone does not fully define agency. Tools can be combined into sequences, and some systems extend tool functionality into limited chains of operations. The distinction therefore extends beyond the number of steps to include who determines those steps and how they are connected.

The category of **chatbots** introduces a different type of comparison. Chatbots are typically described as systems designed for conversational interaction. Their primary function is to engage in dialogue, interpret user input, and generate responses that maintain coherence within that interaction. In many cases, chatbots incorporate elements of reasoning, context retention, and even limited tool use.

The distinction between chatbots and agents is often expressed through the difference between **conversation and execution**. A chatbot is described as facilitating dialogue, while an agent is described as acting on the outcome of that dialogue. This distinction is sometimes summarised as the difference between systems that “say” and systems that “do.” While such phrasing is simplified, it reflects a recurring pattern in how the categories are separated.

However, this distinction is not always clear in practice. Some chatbots are equipped with the ability to call external functions or perform limited actions. In such cases, the boundary becomes dependent on whether those actions are reactive or part of a broader, self-directed process. A chatbot that calls a tool in direct response to a user request may still be described

as a chatbot with enhanced capabilities. An agent, by contrast, is more consistently described as initiating subsequent actions as part of a plan rather than as isolated responses.

This introduces a distinction based on **control over progression**. In chatbot interactions, the user typically defines the sequence of steps through successive prompts. In agent-like systems, the system itself contributes to defining that sequence, determining what to do next based on the evolving state of the task. The presence of conversational capability does not determine classification; rather, it is the role that conversation plays within the broader process.

The category of **automation workflows** presents a more complex comparison. Automation systems are typically described as executing predefined sequences of actions in response to specific triggers. These sequences may involve multiple steps, interactions with external systems, and conditional logic. In some cases, automation workflows can appear similar to agent-like systems, particularly when they handle tasks that involve multiple stages.

The distinction between automation and agents is often framed in terms of **determinism versus adaptability**. Automation workflows are generally described as following fixed rules. The sequence of actions is defined in advance, and the system executes those actions consistently under the same conditions. While conditional branching may exist, the range of possible behaviours is predetermined.

Agents, by contrast, are described as operating with a degree of flexibility in how they achieve an objective. They are not limited to a single predefined path. Instead, they evaluate the current state and determine an appropriate next step, which may vary depending on circumstances. This introduces a distinction between **predefined sequencing** and **dynamic sequencing**.

Despite this distinction, the boundary between automation and agents is not absolute. Hybrid systems exist in which predefined workflows incorporate elements of reasoning or decision-making. In such cases, parts of the system may function as automation, while other parts exhibit agent-like behaviour. The classification of the overall system may depend on which aspect is considered dominant, rather than on a strict separation.

A recurring theme across these comparisons is the role of **orchestration**. Tools, chatbots, and automation systems can all perform actions, process information, and interact with external systems. What appears to distinguish agents more consistently is the presence of a coordinating function that integrates these capabilities into a coherent process. The agent is described not only as performing actions, but as determining how those actions are sequenced, adjusted, and connected to an objective.

This distinction between **orchestration and execution** provides a unifying perspective across categories. Tools execute functions. Chatbots facilitate interaction. Automation workflows execute predefined sequences. Agents, in contrast, are described as orchestrating a process that may involve all three. This orchestration is not unrestricted; it operates within constraints and configurations. However, it introduces a layer of internal decision-making that is less prominent in the other categories.

Another point of comparison relates to **dependency on user input**. Tools and chatbots are typically dependent on user prompts to initiate each step of activity. Automation workflows

are dependent on predefined triggers. Agents are described as operating with reduced dependency on continuous input once a task has been initiated. This does not eliminate user involvement, but it changes its role from directing each step to defining the objective and boundaries.

The distinction is therefore not between dependence and independence, but between different forms of dependence. Agents remain dependent on human-defined goals and constraints, consistent with the concept of dependent autonomy discussed earlier. What changes is the degree to which the system determines the path toward that goal.

It is also observable that these categories are not mutually exclusive in implementation. A single system may incorporate elements of tools, chatbot interfaces, and automation workflows while being described as an agent. This compositional nature contributes to the variability in classification. The same system may be described differently depending on which aspect is emphasised.

This variability reinforces the need to treat these distinctions as **analytical constructs rather than fixed categories**. They provide a way to examine how systems operate, but they do not map cleanly onto all real-world implementations. As systems evolve, the boundaries between categories may continue to shift, further complicating classification.

In early 2026, the term “AI Agent” therefore functions as a point of convergence for multiple capabilities that were previously described separately. It brings together reasoning, action, and coordination in a way that extends beyond the scope of tools, chatbots, or automation alone. At the same time, it remains connected to these categories, drawing on their components and often incorporating them within its operation.

Understanding the distinctions between these categories is not intended to create rigid separations. Rather, it provides a framework for interpreting how the term is used and why different systems may be described differently. By examining where these boundaries appear to hold and where they become less distinct, it becomes possible to approach the concept of “AI Agents” with greater clarity while maintaining the discipline of not overstating certainty.

Chapter 7 — Types of AI Agents Observed in Practice

The identification of different "types" of AI Agents in current discourse does not result in a fixed or universally accepted taxonomy. Across systems, there is no single classification scheme that is consistently applied. Instead, what emerges is a set of recurring patterns—groupings based on how systems are used, how they operate, and the contexts in which they are described. These groupings are neither mutually exclusive nor exhaustive. They overlap, evolve, and vary depending on the perspective from which they are observed.

This chapter therefore approaches "types" not as formal categories, but as **observed patterns of usage**. The aim is to describe how agents are commonly grouped in practice, while preserving the flexibility and variation that characterise the current state of discourse.

One of the most frequently observed groupings is based on **task orientation**. Many systems described as agents are associated with specific domains or functions, such as managing communications, analysing data, or coordinating workflows. In these contexts, agents are framed as systems that carry out defined categories of work rather than as general-purpose entities.

Within this pattern, there is a recurring distinction between **narrowly scoped agents** and **more generalised agents**. Narrowly scoped agents are described as operating within a defined domain, where their capabilities are aligned with a particular type of task. Their behaviour is constrained by the boundaries of that domain, even if they exhibit flexibility within it. More generalised agents, by contrast, are described as capable of handling a wider range of tasks, often by applying similar reasoning and orchestration patterns across different contexts.

This distinction, however, does not form a strict boundary. Systems may exhibit characteristics of both, depending on how they are configured or described. A system may be narrow in deployment but general in underlying capability, or general in description but narrow in practice. As such, the distinction reflects a tendency in how systems are framed rather than a definitive classification.

Another commonly observed grouping is based on the **structure of decision-making and response**. Some agents are described as operating in a relatively immediate and reactive manner, where actions follow closely from perceived inputs within a short cycle. Others are described as engaging in more extended processes, where tasks are broken down into multiple stages and revisited iteratively.

This distinction aligns loosely with earlier conceptual descriptions of reflexive and goal-based systems, though in practice the separation is less formal. Systems may exhibit both reactive and deliberative behaviours within the same task, shifting between immediate response and more extended planning as required. The grouping therefore reflects a spectrum of behaviour rather than discrete types.

A further pattern emerges in the distinction between **single-agent operation** and **multi-agent interaction**. In some contexts, agents are described as operating independently, handling tasks from initiation to completion within a single system. In others, multiple agents are

described as interacting, either cooperatively or in parallel, contributing to different aspects of a broader process.

Multi-agent arrangements are often associated with more complex tasks, where coordination between different roles or functions is required. However, the presence of multiple agents does not redefine the nature of a single agent. Instead, it introduces an additional layer of interaction, where orchestration may occur both within and between systems. The boundary between a single agent and a collection of interacting agents is therefore not always sharply defined.

Another grouping is based on the **relationship between agents and their operating environment**. Some agents are described as embedded within existing systems, functioning as components of larger workflows or platforms. Others are described as more standalone in their operation, interacting with multiple systems from a more independent position.

Embedded agents often operate within defined constraints, closely aligned with the structure of the system in which they are integrated. Standalone agents, by contrast, are described as interacting across systems, drawing on multiple sources of information and action. However, this distinction is again fluid. A system may appear standalone in one context and embedded in another, depending on how it is deployed.

A further observed grouping relates to the **nature of tasks performed**, particularly in terms of cognitive versus procedural emphasis. Some agents are described as focusing on information processing tasks, such as analysing data, summarising content, or generating structured outputs. Others are described as focusing on operational tasks, such as coordinating activities, managing processes, or executing sequences of actions.

This distinction reflects different emphases rather than different categories. Many systems combine both elements, using reasoning to inform action and action to produce new information. The separation between cognitive and operational roles therefore highlights variation in emphasis rather than a strict division.

The **degree of autonomy within tasks** also contributes to how agents are grouped. Some systems are described as operating with minimal intervention once a task is initiated, while others require more frequent input or confirmation. This variation is often interpreted as reflecting different levels of sophistication or capability.

However, as noted in earlier chapters, autonomy functions as a spectrum rather than a defining criterion. Systems with different levels of intervention may still fall within the broader description of agents, provided they exhibit the core characteristics identified previously. The grouping based on autonomy therefore reflects differences in expression rather than differences in classification.

Another pattern of grouping is based on **temporal scope**. Some agents are described as operating within short, bounded tasks that are completed in a single session. Others are described as participating in longer processes, where tasks may extend over time or involve multiple stages of interaction.

This distinction intersects with considerations of memory and continuity. Systems operating over longer durations may incorporate additional mechanisms for maintaining state, while

shorter-duration systems rely primarily on immediate context. The temporal dimension adds another layer of variation, but does not by itself define a separate type.

Across these groupings, a recurring observation is that the same system may be described differently depending on the perspective taken. A system framed as a productivity assistant in one context may be described as a workflow agent in another. A system operating as a single agent may be considered part of a multi-agent arrangement when viewed at a different level of abstraction.

This variability reinforces the idea that "types" of AI Agents are not fixed categories, but **interpretive groupings**. They provide a way to organise observations about how systems behave and how they are described, but they do not impose rigid boundaries. The absence of a stable taxonomy reflects both the evolving nature of the technology and the flexibility of the language used to describe it.

It is also observable that these groupings often reflect practical considerations rather than theoretical ones. Systems are described in terms of the tasks they perform, the environments in which they operate, and the roles they appear to fulfil. These descriptions are shaped by usage and context, rather than by adherence to a predefined classification scheme.

As a result, the term "AI Agent" accommodates a range of system types without resolving them into a single structure. This flexibility allows the term to adapt to new forms of implementation, but it also contributes to the ongoing ambiguity discussed in earlier chapters.

The purpose of identifying these observed groupings is therefore not to establish a definitive taxonomy, but to provide a clearer view of how variation is currently expressed. By recognising these patterns as descriptive rather than prescriptive, it becomes possible to understand the diversity of systems encompassed by the term while maintaining the discipline of not overstating certainty.

The chapters that follow build on this understanding by examining how these variations manifest in real-world contexts and how they interact with the broader limitations and constraints of current implementations.

Chapter 8 — Real-World Patterns of Use (Early 2026 Snapshot)

In early 2026, the use of systems described as "AI Agents" reflects a transition from conceptual framing to practical deployment. This transition is not uniform across contexts. Instead, what emerges is a set of observable patterns that vary by domain, level of complexity, and degree of integration into existing processes. These patterns do not represent a consolidated model of use. They reflect a landscape in which different interpretations of agency are applied to different operational needs.

One of the most consistently observed patterns is the use of agents in **task coordination across multiple steps**. In many contexts, systems are described as handling processes that would otherwise require a sequence of discrete actions. These processes often involve gathering information, performing intermediate transformations, and producing outputs that contribute to an overall objective. The emphasis in such use is not on any single action, but on the continuity of the process.

This pattern is particularly visible where tasks are repetitive but not entirely uniform. Systems are described as managing variations within a general structure, adapting to differences in input while maintaining a consistent objective. The role of the agent in such contexts appears to centre on maintaining coherence across steps, rather than optimising any individual step in isolation.

A second observable pattern is the use of agents in **interface mediation between systems**. In these contexts, agents are described as operating between different sources of information or functionality, translating inputs and outputs in ways that allow processes to continue without direct human coordination at each stage. The agent functions as an intermediary, connecting components that would otherwise require manual interaction.

This pattern does not necessarily involve complex reasoning in all cases. Its significance lies in the continuity it provides across systems. The agent is described as maintaining the flow of a process, ensuring that outputs from one stage become inputs for the next. In this sense, the role is less about decision-making in isolation and more about maintaining structured progression across connected elements.

Another pattern is the use of agents in **contextual information handling**. Systems are described as interpreting inputs that are not fully structured, extracting relevant elements, and organising them in ways that support subsequent actions. This pattern appears in contexts where information must be filtered, prioritised, or transformed before it can be used effectively.

In such cases, the agent is not simply generating outputs, but shaping the context in which further steps occur. The process may involve selecting relevant data, organising it into a usable form, and integrating it into a broader sequence of actions. This reflects a shift from isolated response generation to participation in a process that extends beyond a single interaction.

A further observable pattern involves **conditional progression within workflows**. Systems are described as advancing tasks based on intermediate conditions, adjusting their behaviour

in response to outcomes. This may involve selecting alternative paths, retrying actions, or modifying subsequent steps based on available information.

This pattern aligns with the earlier distinction between predefined and dynamic sequencing. In practice, it often appears as a combination of both. Systems operate within structured workflows but incorporate points of flexibility where decisions are made based on context. The presence of such conditional behaviour contributes to the perception of agency, even where the overall structure remains bounded.

Another pattern relates to the use of agents in **aggregation and synthesis of information across sources**. Systems are described as drawing on multiple inputs, combining them into a coherent output, and presenting results that reflect a broader view than any single source. This pattern is particularly evident in contexts where information is distributed across different locations or formats.

The agent's role in this pattern is not limited to retrieval. It involves selecting, organising, and integrating information in a way that supports the objective of the task. The process often unfolds across multiple steps, with intermediate results informing subsequent actions. This reinforces the characteristic of iterative progression observed in earlier chapters.

A further pattern is the use of agents in **assisted decision processes**. In these contexts, systems are described as contributing to decision-making by providing structured outputs, evaluating options, or presenting alternative pathways. The agent does not determine the final decision but participates in shaping the information on which that decision is based.

This pattern reflects the broader theme of dependent autonomy. The system operates within defined boundaries, contributing to the process without assuming control over outcomes. The distinction between contributing to a decision and making a decision remains consistent across observed uses, reinforcing the role of human oversight in current implementations.

Another observable pattern is the use of agents in **process monitoring and adjustment**. Systems are described as tracking the progression of tasks, identifying deviations, and responding to changes in state. This may involve recognising when expected conditions are not met and adjusting behaviour accordingly.

In such contexts, the agent functions as part of an ongoing process rather than as a discrete responder. Its activity is continuous rather than episodic, reflecting a shift from interaction-based use to process-based use. The system's role is to maintain alignment with an objective over time, rather than to complete isolated tasks.

The pattern of **partial autonomy within constrained environments** is also consistently observed. Systems are described as operating independently within defined limits, handling routine aspects of a task while deferring more significant decisions or actions. This arrangement reflects a balance between automation and oversight, where the agent's activity is bounded by conditions that restrict its scope.

This pattern reinforces the earlier observation that autonomy is local rather than global. The system may act without continuous input, but it does so within parameters that are externally defined. The boundaries of operation are not determined by the system itself, but by its configuration and the context in which it is deployed.

Another pattern emerges in the use of agents for **translation between levels of abstraction**. Systems are described as converting high-level objectives into sequences of actionable steps, or as transforming detailed inputs into more generalised outputs. This translation function bridges the gap between intent and execution, allowing processes to proceed without requiring explicit specification at every stage.

In this role, the agent functions as an intermediary between different representations of a task. It interprets, restructures, and reformulates information in ways that enable further action. This pattern aligns closely with the concept of orchestration, where the system coordinates the relationship between different elements of a process.

Across these patterns, a consistent observation is that the use of agents is rarely isolated. Systems are often embedded within broader processes, interacting with other components, contributing to ongoing workflows, and operating alongside human input. The agent is not typically positioned as a standalone entity replacing existing structures, but as a component within a larger arrangement.

It is also observable that the same system may participate in multiple patterns simultaneously. A system may coordinate tasks, mediate between systems, and synthesise information within a single process. The patterns described here therefore overlap and interact, reflecting the compositional nature of current implementations.

The diversity of these patterns underscores the absence of a single dominant model of use. Instead, usage reflects adaptation to context. Different environments emphasise different aspects of agency, and the same underlying capabilities may be applied in varying ways. This variation is not incidental; it is a defining feature of the current stage of development.

It is therefore appropriate to treat these patterns as a **snapshot of usage in early 2026**, rather than as stable or final forms. They reflect how the concept of an AI Agent is being operationalised across contexts, without implying that these uses will remain unchanged. As with the term itself, the patterns of use remain subject to ongoing interpretation and evolution.

The purpose of identifying these patterns is not to establish a model for adoption, but to provide a clearer view of how the concept is currently applied. By examining these uses in an observational manner, it becomes possible to understand how agency is expressed in practice while maintaining the discipline of not converting description into prescription.

The following chapters extend this examination by considering the limitations and constraints that accompany these patterns of use, further grounding the concept within its present operational context.

Chapter 9 — Risks, Limitations, and Structural Constraints

The observed patterns of use described in the preceding chapter are accompanied by a set of limitations and constraints that shape how systems described as "AI Agents" operate in practice. These constraints do not arise from isolated failures or exceptional cases. They appear as structural features across multiple interpretations, influencing how such systems are deployed, how their outputs are interpreted, and how their scope is bounded.

This chapter does not frame these constraints as warnings or as barriers to use. Instead, it examines them as **conditions of operation**—features that define the current state of agent-like systems in early 2026. These conditions help explain why certain patterns of use recur, why others remain limited, and why the concept of agency remains bounded despite increasing capability.

One of the most consistently observed constraints relates to **interpretation of objectives**. Systems described as agents operate based on representations of goals that are derived from input. This process of interpretation introduces variability. The system's internal representation of a task may not fully align with the intended meaning of the input, particularly when instructions are incomplete, ambiguous, or context-dependent.

This limitation is structural rather than incidental. It reflects the reliance of such systems on patterns within input data rather than on external verification of intent. As a result, the progression of a task may follow a path that is internally consistent but not fully aligned with the originating objective. The system continues to operate within its interpretation unless external intervention occurs.

A related constraint is observed in the **propagation of intermediate outcomes**. Agent-like systems operate through sequences of steps, where each step contributes to the next. This structure enables continuity but also introduces dependency between stages. When an intermediate result is incomplete or misaligned, subsequent steps may proceed on that basis, extending the effect across the process.

This pattern does not depend on the complexity of the task. It arises from the iterative nature of operation. Each step assumes the validity of prior steps, unless mechanisms are present to reassess or interrupt the sequence. The continuity that enables progression also creates a pathway through which deviations can persist.

Another structural constraint concerns the **bounded nature of context awareness**. Systems described as agents maintain awareness within a task, but this awareness is limited to the information available within the operational context. Elements that fall outside this context—such as implicit assumptions, external conditions, or unstated constraints—may not be incorporated into the system's reasoning.

This limitation contributes to behaviour that is consistent within the available context but not necessarily aligned with broader conditions. The system does not extend beyond its accessible inputs unless explicitly provided with additional information. As a result, the scope of reasoning remains bounded by what is represented within the task environment.

A further constraint is the **dependence on structured interaction with external systems**. While agents are described as interacting with tools or environments, these interactions occur within predefined interfaces and permitted actions. The system's ability to affect change is therefore limited by the structure and availability of these interfaces.

This introduces a distinction between theoretical capability and practical operation. A system may be described as capable of certain actions, but its realised behaviour depends on the configuration of its environment. The agent does not independently extend its reach beyond these constraints. Its operation remains anchored to the structures within which it is embedded.

The concept of **dependent autonomy**, introduced in earlier chapters, reflects another consistent constraint. Systems may exhibit autonomy within the execution of a task, but they do not determine the scope, objectives, or boundaries of that task. These elements are externally defined. The system's autonomy is therefore conditional, operating within a framework that remains under human configuration.

This condition is not treated as a temporary limitation within observed usage. It appears as a structural feature of current implementations. The system's role is to carry out processes within defined parameters, rather than to redefine those parameters. This distinction shapes how agency is expressed in practice, maintaining a separation between execution and authority.

Another observable limitation concerns the **handling of extended or complex task sequences**. As the number of steps in a process increases, the system's ability to maintain coherence across those steps may vary. The iterative loop that supports progression also introduces accumulation of intermediate states, which may become more difficult to manage consistently over time.

This pattern reflects the interaction between iterative reasoning and state management. The system operates within a sequence of evolving conditions, each dependent on prior steps. As this sequence extends, maintaining alignment with the original objective becomes more complex. The limitation is therefore not tied to specific tasks, but to the structure of multi-step operation itself.

A further constraint arises in the **transparency of decision pathways**. Systems described as agents often produce outputs that are the result of multiple internal steps, not all of which are explicitly visible. While intermediate reasoning may sometimes be represented, the overall process may not be fully traceable in a structured way.

This affects how outcomes are interpreted. The result may be observable, but the pathway through which it was produced may not be fully accessible or easily reconstructed. This characteristic does not prevent the system from operating, but it influences how its outputs are understood within a broader process.

The **interaction between flexibility and predictability** also presents a structural condition. Agent-like systems are described as capable of adapting their behaviour based on context. This adaptability contributes to their ability to handle variation within tasks. At the same time, it introduces variability in outcomes, as different paths may be taken under similar conditions.

This dual characteristic—flexibility alongside variability—shapes how such systems are integrated into processes. The system's behaviour is not fixed in the way that predefined workflows are. Instead, it operates within a range of possible outcomes, influenced by context and intermediate states. This condition is inherent to the adaptive nature of such systems.

Another observable constraint involves the **alignment between outputs and external expectations**. Systems may generate outputs that are internally coherent and aligned with their interpretation of a task, yet not fully aligned with external standards, conventions, or requirements. This misalignment is not necessarily the result of error in a narrow sense, but of differences between internal reasoning and external context.

This reflects the broader limitation of operating within a defined input-output framework. The system does not independently validate its outputs against external criteria unless such validation is explicitly incorporated into the process. As a result, outputs may require external interpretation or adjustment within the broader workflow.

The **handling of exceptions and novel conditions** represents another structural constraint. Systems operate based on patterns derived from prior data and within defined operational structures. When encountering conditions that fall outside these patterns or structures, their behaviour may become less consistent. The system may attempt to proceed based on available patterns or may defer to external input.

This limitation does not manifest uniformly. Some systems incorporate mechanisms to recognise and respond to such conditions, while others continue within the existing process. The constraint lies in the bounded nature of pattern-based reasoning, which does not inherently extend to all possible scenarios.

Across these constraints, a consistent theme is the interaction between **capability and boundary**. Systems described as agents exhibit capabilities that extend beyond single-step responses, yet those capabilities are consistently framed within limits that shape their operation. These limits are not external impositions alone; they arise from the structure of how such systems function.

It is also observable that these constraints do not prevent the emergence of the patterns described in earlier chapters. Rather, they coexist with them, influencing how those patterns are realised. The same characteristics that enable multi-step progression, adaptation, and interaction also introduce dependencies, variability, and bounded scope.

In this sense, the limitations described here are not separate from the concept of agency. They are part of its current expression. Understanding these structural conditions provides a more complete view of how the term "AI Agent" is applied in practice, not by restricting its meaning, but by situating it within the realities of its operation.

The chapters that follow continue this examination by exploring how these constraints intersect with broader patterns of interpretation and how they contribute to the ongoing ambiguity and variation associated with the term.

Chapter 10 — Autonomy, Control, and Human Oversight

The concept of autonomy occupies a central position in how "AI Agents" are described, yet its meaning varies significantly across contexts. In common usage, autonomy often carries an implicit association with independence, self-direction, or reduced reliance on external input. However, across observed systems in early 2026, autonomy does not appear as a binary state. It is more consistently represented as a **spectrum of behaviour**, bounded by conditions that define how and where a system can act.

This distinction between perception and structure is important. While the term "autonomous" is frequently applied to agent-like systems, its practical meaning is narrower. Systems are described as capable of operating without continuous instruction within a task, but not as capable of determining their own objectives, redefining their scope, or assuming responsibility for outcomes. Autonomy, in this sense, is localised and conditional rather than generalised or absolute.

A useful way to examine this is through the distinction between **execution and authority**. Systems described as agents may execute sequences of actions, select from available options, and adjust behaviour based on intermediate results. These activities give the appearance of independence within a task. However, the authority to define what the task is, what constraints apply, and what outcomes are acceptable remains external to the system.

This separation between execution and authority is consistently observed. The system may determine how to carry out a task, but it does not determine why the task exists or whether it should be undertaken. The originating objective, the permissible range of actions, and the conditions for completion are defined outside the system. As a result, autonomy is exercised within boundaries rather than across them.

The concept of **dependent autonomy** provides a useful framing for this condition. Systems operate with a degree of procedural independence, but this independence is dependent on human-defined parameters. The system's behaviour is shaped by its configuration, the inputs it receives, and the constraints within which it operates. It does not extend beyond these parameters without external intervention.

This dependence is evident in the role of **goal definition**. Across observed uses, agents do not originate their own objectives. They operate in response to tasks that are defined externally, whether through direct input or through structured processes. The system may interpret and refine the representation of a goal, but it does not establish that goal independently. The starting point of activity remains external.

A related dimension of control concerns the **boundaries of permitted action**. Systems interact with external environments through defined interfaces and within specified limits. These limits determine what actions are available, what data can be accessed, and what changes can be made. The system does not expand these boundaries; it operates within them.

This introduces a distinction between **capability and permission**. A system may be capable of performing a range of actions in principle, but its realised behaviour is constrained by what it is permitted to do in a given context. The configuration of these permissions is external to

the system's internal reasoning process. As such, autonomy is shaped not only by what the system can do, but by what it is allowed to do.

Another dimension of autonomy relates to **continuity of operation within a task**. Once initiated, agent-like systems are described as capable of progressing through multiple steps without requiring instruction at each stage. This continuity is one of the defining features that differentiates agents from purely reactive systems. It reflects the system's ability to maintain a trajectory toward an objective.

However, this continuity does not imply independence from oversight. In many observed patterns, the system's operation includes points at which external input may be required, either to confirm actions, provide additional information, or resolve ambiguity. These points of interaction do not negate autonomy within the task, but they illustrate its bounded nature.

The role of **human oversight** therefore remains structurally central. Oversight does not always manifest as continuous intervention. It may take the form of initial configuration, definition of constraints, or periodic review of outcomes. The presence of oversight reflects the separation between system action and human responsibility.

This separation is particularly evident in the distinction between **system action and outcome accountability**. Systems may carry out actions that contribute to a process, but the responsibility for those actions remains external. The system does not assume accountability in the way that a human actor would. This distinction reinforces the boundary between execution and authority, extending it into the domain of responsibility.

Another aspect of autonomy concerns the system's ability to **adapt within constraints**. Agent-like systems are described as adjusting their behaviour based on context, selecting different paths toward an objective depending on intermediate conditions. This adaptability contributes to their flexibility and is often associated with autonomy.

At the same time, this adaptability operates within a defined range. The system does not generate entirely new categories of action or redefine the structure of the task. It selects from available options, guided by its internal processes and external constraints. The adaptability is therefore bounded, reflecting variation within limits rather than unrestricted freedom.

The spectrum of autonomy can also be observed in the **degree of intervention required during operation**. Some systems are described as operating with minimal input once a task begins, while others require more frequent interaction. This variation reflects differences in how autonomy is expressed, rather than differences in its presence or absence.

Even in cases where intervention is limited, the underlying structure remains dependent. The system's operation continues to rely on externally defined goals, permissions, and evaluation criteria. Reduced interaction does not equate to independence; it reflects a shift in how control is exercised rather than a transfer of control itself.

Another dimension of autonomy relates to **handling of uncertainty and exceptions**. Systems may encounter conditions that are not fully specified within the task or that fall outside expected patterns. In such cases, the system's behaviour may involve attempting to proceed based on available information or deferring to external input.

This behaviour illustrates the limits of autonomy. The system does not extend its reasoning into areas where its internal structures do not provide sufficient guidance. Instead, it either operates within known patterns or relies on external clarification. The boundary of autonomy is therefore defined not only by constraints, but by the scope of the system's internal representations.

Across these dimensions, autonomy appears as a **layered characteristic**, composed of multiple interacting elements: execution capability, adaptability, continuity, and bounded decision-making. Each of these elements contributes to how autonomy is perceived, but none of them independently establishes it as a complete or unrestricted state.

The framing of autonomy as a spectrum allows for variation without implying progression toward full independence. Systems may occupy different positions within this spectrum depending on their configuration, context, and mode of operation. However, the overall structure remains consistent: autonomy is exercised within limits that are externally defined and maintained.

It is therefore appropriate to interpret autonomy in the context of AI Agents as **bounded and conditional**. It reflects the system's ability to carry out processes within a defined scope, not its ability to transcend that scope. The distinction between acting within a framework and defining that framework remains central.

This understanding also clarifies the role of human oversight. Oversight is not positioned as a temporary measure that diminishes over time, but as an integral component of how such systems operate. It defines the boundaries within which autonomy is expressed and maintains the separation between system action and human authority.

In early 2026, the concept of autonomy in relation to AI Agents can therefore be described as structured rather than absolute. It is present in the system's ability to progress through tasks, adapt within constraints, and operate with limited intervention. At the same time, it remains dependent on external definition, bounded by configuration, and separated from authority and responsibility.

The chapters that follow continue to examine how this bounded autonomy interacts with broader patterns of use and interpretation, further situating the concept within its current operational context.

Chapter 11 — Ambiguity, Semantic Dilution, and Variability of the Term "AI Agent"

The term "AI Agent," as it appears across current discourse, is characterised by a level of ambiguity that is not incidental but structural. This ambiguity does not arise solely from incomplete understanding or inconsistent usage. It reflects the interaction between evolving system capabilities, flexible language, and differing interpretive frameworks. As a result, the term functions less as a fixed definition and more as a **variable construct**, whose meaning depends on context, emphasis, and perspective.

This variability is evident in how different systems describe the same or similar capabilities. Across observed outputs, there is no singular threshold that consistently determines when a system is classified as an agent. Instead, the term is applied across a range of systems that share certain characteristics but differ in structure, scope, and operation. These differences are not always resolved through explicit criteria. Rather, they coexist within the broader usage of the term.

A central aspect of this variability is the phenomenon of **semantic dilution**. As the term "AI Agent" has gained prominence, it has been applied to an increasing number of systems, some of which exhibit only partial alignment with more structured interpretations of agency. Features such as tool use, multi-step execution, or limited task automation are sometimes presented as sufficient to justify the label, even where other characteristics—such as iterative reasoning or internal sequencing—are less evident.

This expansion of meaning does not occur uniformly. In some contexts, the term retains a relatively narrow interpretation, associated with systems that demonstrate coordinated reasoning and action. In others, it is used more broadly to describe systems that extend beyond single-response behaviour. The result is a widening of the conceptual boundary, where the term encompasses both more structured and more loosely defined systems.

Semantic dilution, in this sense, is not merely a matter of imprecision. It reflects the adaptability of language in response to technological change. As systems acquire new capabilities, the language used to describe them adjusts to accommodate these changes. However, this adjustment does not always preserve distinctions that were previously clearer. Instead, the term absorbs multiple meanings, some of which overlap and some of which diverge.

Another dimension of ambiguity arises from the **overlapping nature of adjacent concepts**. As discussed in earlier chapters, the boundaries between agents, tools, chatbots, and automation workflows are not fixed. These categories share underlying components, and systems often incorporate elements of more than one category. As a result, classification depends on which aspects of a system are emphasised.

This overlap contributes to variability in how the term is applied. A system that is described as an agent in one context may be described as a tool or a workflow in another. The difference is not necessarily due to a change in the system itself, but to a shift in perspective. The term "AI Agent" therefore operates within a **relational framework**, where its meaning is shaped by comparison with adjacent concepts.

Ambiguity is also reinforced by differences in **levels of abstraction**. In some interpretations, the term is used to describe a specific architectural pattern involving iterative reasoning and action. In others, it is used more generally to describe any system that appears to carry out tasks with limited input. These different levels of abstraction coexist without a clear boundary between them.

At a higher level of abstraction, the term functions as a conceptual descriptor, capturing the idea of systems that "act" rather than merely "respond." At a lower level, it may refer to particular structural characteristics or operational behaviours. The movement between these levels contributes to the fluidity of the term, allowing it to be applied in both broad and narrow senses.

Another source of variability lies in the **weight assigned to different characteristics**. As identified in earlier chapters, certain features—such as goal-directed progression, iterative operation, and interaction with an environment—are commonly associated with agency. However, the relative importance of these features is not consistently defined.

Some interpretations emphasise planning and reasoning as central, while others focus on action and execution. Some treat memory as a defining element, while others consider it optional. These differences in emphasis lead to different thresholds for classification. The term "AI Agent" therefore reflects not only the presence of certain characteristics, but also how those characteristics are prioritised.

The role of **language and metaphor** further contributes to ambiguity. Descriptions of agents often draw on analogies to human roles, such as assistants, workers, or collaborators. These analogies provide intuitive entry points for understanding, but they also introduce interpretive flexibility. Systems that exhibit limited forms of behaviour associated with these roles may nonetheless be described using the same terminology.

This use of metaphor does not necessarily clarify the concept. Instead, it can broaden the range of systems to which the term is applied, reinforcing semantic dilution. The analogy becomes a bridge between technical capability and everyday understanding, but it does not define a precise boundary.

Ambiguity is also sustained by the **absence of formal standardisation**. There is no universally accepted definition of "AI Agent" that is enforced across systems or contexts. While certain patterns of agreement exist, they do not constitute a formal framework. This absence allows for flexibility, but it also permits variation in interpretation to persist.

This condition is not necessarily temporary. The variability of the term reflects the current stage of development, where capabilities are expanding and interpretations are evolving in parallel. The lack of standardisation allows the term to accommodate new forms of implementation without requiring immediate redefinition.

At the same time, this flexibility introduces challenges in maintaining conceptual clarity. Without stable boundaries, the term may be used in ways that obscure differences between systems. This does not invalidate its use, but it requires attention to context when interpreting its meaning.

Across these dimensions, ambiguity appears as a **feature of the term rather than a flaw**. It allows the concept to remain adaptable, accommodating a range of systems and interpretations. However, this adaptability comes with the consequence that meaning must be inferred from context rather than assumed from the term itself.

The variability of "AI Agent" can therefore be understood as the result of multiple interacting factors: evolving capabilities, overlapping categories, differing levels of abstraction, variable emphasis on characteristics, and flexible use of language. These factors do not converge into a single definition. Instead, they produce a field of meaning that is structured but not fixed.

It is therefore appropriate to treat the term as **intentionally open and context-dependent**. Its meaning is shaped by how it is used, by what is being described, and by the perspective from which it is interpreted. Attempts to impose a single, stable definition risk overlooking the diversity of usage that characterises the current discourse.

This chapter does not resolve the ambiguity associated with the term. It situates it. By recognising ambiguity as structural and semantic dilution as an observable pattern, it becomes possible to engage with the concept without requiring it to conform to a fixed boundary. This approach maintains clarity while preserving the flexibility that defines the term's current state.

The chapters that follow continue this examination by exploring how these patterns of ambiguity and variability interact with the near-term trajectory of the concept, further grounding the term within its evolving context.

Chapter 12 — Near-Term Trajectory (0–36 Months)

The trajectory of systems described as "AI Agents" in early 2026 can be observed through emerging patterns of development, deployment, and interpretation. These patterns do not form a single, unified direction. Rather, they indicate multiple tendencies that appear to be developing in parallel. This chapter examines those tendencies within a defined time horizon of approximately 0–36 months, based on observable signals present in current usage and discourse.

The purpose of this examination is not to project a fixed future state. It is to identify directions of movement that are suggested by present conditions, while retaining the variability and uncertainty that characterise the current landscape.

One observable direction is the **continued expansion of multi-step task execution within bounded contexts**. Systems described as agents are increasingly associated with the ability to sustain processes across multiple stages, particularly in environments where tasks follow recurring patterns. This expansion appears to be occurring incrementally, with systems handling broader variations of tasks within defined domains rather than extending into unrestricted generality.

The emphasis in this trajectory is not on removing constraints, but on extending capability within them. Systems are described as becoming more consistent in maintaining progression across steps, handling intermediate variation, and coordinating interactions with external components. However, this development remains situated within controlled environments where inputs, actions, and outcomes are structured.

Another observable tendency is the **increased integration of agent-like behaviour into existing systems and workflows**. Rather than appearing as standalone entities, agents are frequently described as components within larger processes. Their role is to contribute to continuity, coordination, and transformation within those processes.

This pattern suggests a trajectory in which agent-like capabilities become more embedded within operational contexts. The system's identity as an "agent" may become less visible as a separate classification and more integrated into broader system behaviour. At the same time, the underlying characteristics associated with agency—such as iterative progression and orchestration—remain present within these integrations.

A further direction is the **refinement of orchestration within defined boundaries**. As discussed in earlier chapters, orchestration—the coordination of reasoning, action, and sequencing—appears as a central feature of agency. Observed developments suggest a focus on improving how this orchestration operates within constrained environments, rather than expanding it into unbounded autonomy.

This refinement includes more consistent handling of intermediate states, clearer structuring of task progression, and more stable interaction between components. The trajectory here reflects an effort to stabilise behaviour within the existing framework of dependent autonomy, rather than to redefine that framework.

Another observable pattern is the **continued presence of human oversight as a structural component of operation**. Across current implementations, oversight appears not as a temporary condition, but as an integral part of how systems are configured and used. This includes the definition of goals, the setting of boundaries, and the interpretation of outcomes.

Within the near-term horizon, there are indications that this relationship between system operation and human oversight remains stable. While the form of interaction may vary—ranging from direct intervention to more periodic review—the underlying structure of dependence does not show signs of removal. Instead, it appears to be incorporated more explicitly into how systems are described and understood.

A related tendency is the **clarification of boundaries between capability and authority**. As systems are more widely used, distinctions between what a system can do and what it is permitted to do are becoming more visible. This distinction contributes to how autonomy is interpreted and how systems are positioned within processes.

The trajectory here suggests a continued emphasis on maintaining this separation. Systems may expand in their ability to generate actions or propose outcomes, but the authority to define scope, approve actions, and assume responsibility remains external. This reinforces the concept of dependent autonomy as a persistent structural condition.

Another observable direction involves the **ongoing variability in how the term "AI Agent" is applied**. As discussed in the previous chapter, the term exhibits semantic flexibility and is used across a range of systems with differing characteristics. Within the near-term horizon, there is no clear indication that this variability converges into a single, standardised definition.

Instead, the term appears likely to continue functioning as a flexible descriptor, accommodating different interpretations depending on context. Some contexts may move toward more structured usage, while others retain broader application. The trajectory, therefore, reflects continued coexistence of multiple interpretations rather than consolidation into a single meaning.

The **interaction between agent-like systems and adjacent concepts** also appears to be evolving. As systems incorporate elements of tools, chatbots, and automation workflows, the distinctions between these categories may become less rigid in practice. Systems may increasingly combine these elements in ways that make classification dependent on perspective rather than on clear separation.

This does not eliminate the analytical distinctions described in earlier chapters. Rather, it suggests that those distinctions may be applied more as interpretive frameworks than as fixed categories. The trajectory here reflects increasing compositional complexity, where systems integrate multiple capabilities within a single process.

Another observable tendency is the **incremental extension of task complexity within controlled environments**. Systems are described as handling more varied inputs, adapting to a wider range of intermediate conditions, and maintaining coherence across longer sequences. This extension appears to occur within bounded contexts, where the structure of tasks remains defined.

The emphasis on controlled environments suggests that development is oriented toward improving reliability and consistency within known conditions, rather than extending into open-ended or unstructured domains. This reflects the broader pattern of capability expansion within constraints.

A further direction is the **continued alignment between system outputs and structured processes**. As agents are integrated into workflows, their outputs are increasingly expected to align with the requirements of those workflows. This includes producing results that can be incorporated into subsequent steps without requiring extensive reinterpretation.

This alignment does not eliminate variability in outputs, but it introduces a tendency toward structuring outputs in ways that support continuity. The trajectory suggests an ongoing interaction between flexibility in reasoning and consistency in integration.

Across these observed directions, a consistent theme is the **coexistence of expansion and constraint**. Systems described as agents are extending their capabilities in terms of multi-step progression, coordination, and adaptability. At the same time, these capabilities remain bounded by structural conditions, including defined environments, external oversight, and limited authority.

This coexistence shapes the near-term trajectory. Development does not appear to be moving toward unbounded autonomy or unrestricted generality within the observed time horizon. Instead, it reflects a pattern of **incremental extension within established boundaries**, where improvements are made to how systems operate within their existing frameworks.

It is also important to recognise that these tendencies are not uniform across all contexts. Different domains, use cases, and configurations exhibit different patterns of development. Some contexts emphasise integration, others focus on coordination, and others on information handling. The trajectory is therefore not singular, but composed of multiple parallel developments.

The variability described in earlier chapters remains present within this trajectory. Differences in interpretation, classification, and emphasis continue to influence how systems are described and understood. This variability introduces uncertainty into how these patterns may evolve over time.

For this reason, the trajectory described here is best understood as **directional rather than deterministic**. It reflects observable movements based on current signals, without implying that these movements will converge into a single outcome. The patterns identified may persist, evolve, or diverge depending on how systems and interpretations continue to develop.

Within the defined 0–36 month horizon, the concept of “AI Agents” therefore appears to be stabilising in some respects—particularly in its association with multi-step, goal-directed processes—while remaining fluid in others, especially in its terminology and classification. This combination of emerging structure and continued variability characterises the near-term trajectory.

The purpose of this chapter is not to define where the concept will arrive, but to situate where it appears to be moving based on current observations. By maintaining a time-bound and

evidence-based perspective, it becomes possible to examine these developments without extending them into speculative or prescriptive conclusions.

The chapters that follow build on this perspective by drawing together the observations made throughout the guide, further clarifying how the concept of “AI Agents” can be understood within its present context.

Chapter 13 — Interpretive Synthesis (Non-Prescriptive)

The preceding chapters have examined the term “AI Agent” through multiple lenses, each addressing a different aspect of how the concept is currently understood and applied. These lenses include definitional patterns, distinguishing characteristics, relationships to adjacent concepts, observed usage, structural constraints, autonomy, and variability of interpretation. Taken together, they form a layered view of the term as it appears in early 2026. This chapter brings those observations into alignment, not to resolve them into a single definition, but to clarify how they coexist.

A central observation emerging across the chapters is that the term “AI Agent” operates as a **convergence point for multiple characteristics**, rather than as a discrete or sharply bounded category. The characteristics identified as necessary—goal-directed progression, iterative operation, internal sequencing, interaction with an environment, and task-level continuity—form a functional core. Around this core, additional capabilities such as memory, learning, or extended integration appear as variable extensions. This layered structure allows the term to accommodate systems of differing complexity while maintaining a recognisable centre.

At the same time, the boundaries of this structure remain **permeable rather than fixed**. As discussed in earlier chapters, the distinction between agents and adjacent concepts such as tools, chatbots, and automation workflows is not defined by a single criterion. Instead, it emerges from the interaction of multiple characteristics, particularly the role of orchestration in determining how actions are sequenced and connected to an objective. This interaction creates a boundary that is observable but not absolute, allowing for overlap and hybrid forms.

The concept of **orchestration** appears consistently as a unifying element across interpretations. It provides a way of understanding how reasoning and action are coordinated within a process, distinguishing agent-like behaviour from systems that execute predefined or externally directed steps. However, this distinction does not eliminate variability. Systems may exhibit orchestration to different degrees, and its presence does not always lead to uniform classification. The concept therefore functions as an anchor within a broader field of variation.

Another recurring observation is the role of **dependent autonomy**. Autonomy, as applied to agents, is not presented as an independent state. It is consistently bounded by externally defined goals, constraints, and permissions. The system may determine how to progress within a task, but it does not determine the scope or authority of that task. This separation between execution and authority reinforces the idea that autonomy is conditional, operating within a framework that remains externally controlled.

This framing of autonomy aligns with the observed patterns of **human oversight**. Oversight is not positioned as an external constraint imposed on an otherwise independent system. It is embedded within the structure of how such systems operate. The definition of objectives, the configuration of permissible actions, and the interpretation of outcomes remain outside the system. This arrangement reflects a stable relationship between system behaviour and human responsibility, rather than a transitional stage.

The examination of real-world patterns of use further reinforces the idea that the term "AI Agent" is applied across **diverse contexts and functions**. Systems are described as coordinating tasks, mediating between components, handling contextual information, and contributing to decision processes. These uses do not converge into a single model. Instead, they illustrate how the concept adapts to different operational environments, with different aspects of agency becoming more prominent depending on context.

This variability in use is closely connected to the phenomenon of **semantic dilution**. As the term has expanded, it has come to encompass systems that exhibit only some of the characteristics associated with agency. This expansion reflects both the adaptability of language and the absence of formal standardisation. The result is a term that is widely used but variably interpreted, with meaning shaped by context rather than fixed definition.

The absence of a stable definition does not imply the absence of structure. Across systems, there is observable **partial convergence** around certain ideas, particularly the association of agents with multi-step, goal-directed processes. However, this convergence does not extend to agreement on thresholds or boundaries. Instead, it coexists with divergence in how characteristics are weighted, how categories are distinguished, and how the term is applied.

The interplay between convergence and divergence is a defining feature of the concept. It allows the term to function across a range of contexts while maintaining a degree of recognisability. At the same time, it introduces ambiguity that cannot be fully resolved without imposing constraints that are not present in current usage.

The examination of risks and limitations further situates the concept within its operational context. The constraints identified—interpretation of objectives, propagation of intermediate outcomes, bounded context awareness, dependence on structured environments, and variability in handling extended processes—do not stand apart from the concept of agency. They are part of how it is expressed in practice. The same structures that enable multi-step progression and adaptability also introduce these conditions.

Within the near-term horizon, the trajectory of agent-like systems reflects a pattern of **incremental extension within established boundaries**. Capabilities are observed to expand in terms of coordination, integration, and adaptability, while remaining situated within controlled environments and dependent frameworks. This trajectory does not resolve the ambiguity of the term. Instead, it reinforces the coexistence of expansion and constraint.

Taken together, these observations suggest that the term "AI Agent" functions as a **context-dependent construct**, shaped by how systems are described, how their capabilities are interpreted, and how their roles are understood within broader processes. Its meaning is not contained within a single definition, but distributed across a set of interacting elements.

This distribution of meaning has several implications for interpretation. First, it indicates that the term should be understood in relation to the specific context in which it is used, rather than assumed to carry a uniform meaning. Second, it suggests that distinctions between agents and adjacent concepts are best treated as analytical tools rather than fixed categories. Third, it highlights the importance of recognising both the core characteristics that support convergence and the variability that sustains divergence.

At the same time, it is important to note that this synthesis does not seek to stabilise the term. The observations presented throughout the guide do not converge into a single, definitive framework. They remain open, reflecting the current state of discourse in which meaning is negotiated rather than settled.

The purpose of this chapter is therefore not to conclude the discussion, but to integrate its components. By aligning the different perspectives examined in earlier chapters, it provides a structured view of how the concept is currently understood, while preserving the ambiguity and variability that define it.

In this sense, the interpretive synthesis reflects the central position of the term “AI Agent” in early 2026: structured yet open, bounded yet flexible, and consistently used yet variably understood.

Chapter 14 — Closing Boundaries and Open Questions

The examination of "AI Agents" across the preceding chapters has established a set of observable boundaries that frame how the term is currently used and understood. These boundaries do not form a closed perimeter. They define areas of relative clarity while leaving other areas open to interpretation. This chapter revisits those boundaries, not to fix them in place, but to articulate their current contours and to identify where questions remain unresolved.

A first boundary can be observed in the distinction between **goal-directed, multi-step systems** and systems that operate through isolated or reactive responses. Across interpretations, this distinction appears consistently as a point of convergence. Systems described as agents are associated with processes that extend over time, where intermediate steps contribute to an ongoing objective. This boundary provides a functional centre to the concept.

At the same time, this boundary does not determine a precise threshold. Questions remain regarding how many steps are required, how persistence is measured, and how variation in task structure affects classification. The presence of multi-step progression establishes a direction, but not a fixed point of inclusion or exclusion.

A second boundary is formed around the concept of **orchestration**. The distinction between systems that determine their own sequence of actions and those that follow predefined paths provides a structural marker within the broader landscape. Orchestration introduces a layer of internal coordination that differentiates agent-like behaviour from execution alone.

However, this boundary is not absolute. Hybrid systems incorporate both predefined and dynamic elements, making it possible for orchestration to exist in partial or constrained forms. This raises questions about how much internal sequencing is required before a system is considered to exhibit agency, and whether degrees of orchestration should be treated as differences in kind or in degree.

The boundary between **capability and permission** also remains central. Systems may be designed to act upon external environments, yet operate within constraints that limit their execution. The distinction between what a system can do and what it is allowed to do shapes how agency is interpreted. This boundary clarifies that authority remains external, even where capability is present.

At the same time, it leaves open the question of how capability should be evaluated in the absence of execution. If a system consistently prepares actions but does not perform them directly, the relationship between intention and action becomes a point of interpretation rather than a fixed criterion.

The concept of **dependent autonomy** provides another boundary that appears stable in current discourse. Autonomy is observed within the execution of tasks, but not in the definition of those tasks. The separation between execution and authority establishes a structural limit on how autonomy is expressed.

Yet this boundary also introduces open questions. The degree to which autonomy can vary within tasks, the extent to which systems can adapt within constraints, and the relationship between autonomy and oversight remain areas of interpretation. The spectrum of autonomy is observable, but its endpoints are not defined with precision.

Another boundary is present in the distinction between **task-level continuity and broader contextual awareness**. Systems maintain continuity within a task, allowing them to progress across multiple steps. However, their awareness remains bounded by the information available within that task. This boundary highlights the difference between local coherence and broader contextual integration.

This distinction raises questions about how continuity should be extended across tasks, how context is defined, and how systems might operate when boundaries between tasks become less distinct. These questions remain open within current usage, reflecting the variability of system design and interpretation.

The boundary between **necessary and optional characteristics** further defines the structure of the concept. Certain features appear consistently as forming a functional core, while others vary across implementations. This distinction allows for flexibility in how the term is applied, accommodating systems with differing capabilities.

At the same time, it introduces uncertainty regarding how these characteristics interact. Questions remain about whether additional features—such as memory or extended integration—alter the nature of agency or simply extend its expression. The boundary between core and extension remains observable but not fixed.

The variability in how the term is applied introduces another boundary, one shaped by **context and perspective**. The same system may be described differently depending on the lens through which it is viewed. This variability reflects the absence of formal standardisation and the influence of interpretation in shaping meaning.

This condition raises questions about the role of context in defining terms. If classification depends on perspective, then the boundary between categories becomes relational rather than absolute. This does not eliminate the possibility of structure, but it situates that structure within a framework that is inherently flexible.

The phenomenon of **semantic dilution** further complicates the boundaries of the term. As the label “AI Agent” is applied to a wider range of systems, its meaning expands to include partial or simplified forms of agency. This expansion reflects the adaptability of language, but it also introduces ambiguity into how the term is understood.

This raises questions about how meaning evolves as usage expands. Whether semantic dilution leads to greater inclusivity or to loss of clarity remains an open consideration. The term continues to function despite this variability, suggesting that its utility does not depend on strict definition.

The **relationship between the term and its underlying characteristics** also remains open. The term “AI Agent” is often used as a shorthand for a set of capabilities, yet those capabilities are not always present in full. This creates a distinction between the label and the structure it is intended to represent.

Questions arise regarding how closely the label must align with the underlying characteristics to remain meaningful. The absence of a fixed threshold allows for flexibility, but it also requires interpretation in each instance of use.

Across these boundaries, a consistent observation is that the concept of "AI Agent" is both **bounded and open**. Boundaries exist in the form of recurring patterns and distinctions, yet these boundaries do not close the concept. They define areas of relative agreement while leaving space for variation and reinterpretation.

The open questions that emerge from this structure are not incidental. They are part of how the concept functions in its current state. These questions include:

- how thresholds of agency are determined in the absence of fixed criteria
- how degrees of orchestration and autonomy are interpreted across contexts
- how the relationship between capability and execution is understood
- how semantic dilution affects the usefulness of the term
- how classification adapts as systems incorporate multiple overlapping characteristics

These questions do not point toward immediate resolution. They reflect the ongoing interaction between capability, language, and interpretation. As such, they remain part of the concept rather than external to it.

This chapter does not seek to resolve these questions or to close the boundaries that have been described. Instead, it situates them as part of the current landscape. The concept of "AI Agent" remains in a state where structure and openness coexist, where boundaries guide interpretation without fixing it, and where meaning is shaped by context rather than determined in advance.

In early 2026, the term can therefore be understood as a **bounded but open construct**, defined by patterns of use, shaped by evolving capabilities, and sustained by variability in interpretation. The boundaries identified throughout the guide provide a framework for understanding, while the open questions preserve the flexibility that characterises the term.

The discussion concludes not with a definition, but with a position: that clarity can be achieved through structured observation, even where final resolution remains out of scope.

FUTURE EDITION UPDATES & USER SUBMISSIONS

The concept of “AI Agents” is evolving, and its interpretation continues to change across contexts.

Future editions of this guide may:

• incorporate emerging patterns of usage
• reflect shifts in interpretation and terminology
• refine structural observations

AISF welcomes input from readers who identify:

• additional patterns of interpretation
• variations not reflected in this guide
• areas requiring further clarification

All submissions will be reviewed for alignment with AISF’s principles of neutrality, clarity, and non-prescriptive analysis.

ACKNOWLEDGEMENTS

This guide is informed by structured analysis of outputs from multiple globally deployed AI systems, alongside broader observations of how the term “AI Agent” is used across technical, commercial, and public discourse.

AISF acknowledges the contributions of researchers, practitioners, developers, organisations, and users whose interactions with AI systems shape the evolving understanding of this concept.

All content has been synthesised and presented under AISF editorial governance.

About

AI Sourced Facts (AISF) Pte. Ltd.

AISF is a Singapore-headquartered institution dedicated to structured reasoning, responsible AI navigation, and governance-informed adoption of artificial intelligence systems.

AISF operates with a capability-first, vendor-neutral posture. Its publications do not rank platforms, endorse providers, or promote specific technologies. Instead, AISF develops structured frameworks that help individuals, professionals, and institutions reason clearly before integrating AI into operational, strategic, or educational environments.

AISF's work spans whitepapers, applied insight books, education instruments, governance architectures, and structured research initiatives. These outputs are informed by cross-system AI research methodologies and reflect globally observed usage patterns at the time of publication. Human accountability remains central across all AISF frameworks.

AISF does not provide regulatory, legal, financial, investment, or compliance advice. Its publications are designed to support structured thinking, proportionate governance, and disciplined evaluation of AI capabilities prior to deployment or reliance.

As artificial intelligence systems continue to evolve, AISF's focus remains constant: clarity before integration, governance proportionate to capability, and long-term institutional resilience in the age of AI.

BACK COVER

The term “AI Agent” is widely used, yet rarely defined with consistency.

This guide provides a structured and time-bound clarification of how the term is currently understood across global AI discourse.

It examines:

• what AI Agents are described as
• what they are not
• how their characteristics are interpreted
• where boundaries form and where they remain open

Rather than imposing a single definition, this guide maps the patterns, variations, and ambiguities that shape current understanding.

It is intended for:

• individuals seeking clarity in AI terminology
• professionals navigating evolving AI concepts
• organisations interpreting AI capabilities
• educators and learners building structured understanding

AISF publications are developed using cross-system research methodologies and reflect globally observed AI usage patterns at the time of publication.

Artificial intelligence systems assist.
Responsibility remains human.

www.aisourcedfacts.com