

EVERYONE · 5 MINUTE READ

What an AI agent is (and is not)

An AI agent works towards a goal by planning steps, using tools and checking results, while a person sets limits and approves what matters.

What it is

A **chatbot** answers once. You ask a question, it replies, and it waits for you. An **AI agent** is given a goal and works towards it in a loop: it plans the steps, uses tools (such as search, files or a web browser), looks at the results, decides what to do next, and keeps going until the goal is met or it needs your help.

Claude, ChatGPT, Microsoft Copilot and Gemini all offer some agent-style features alongside ordinary chat, depending on your plan and settings. The difference is that the agent may take several steps and use tools between your messages.

Think of asking a colleague a question versus giving them an errand ("find three suppliers, compare prices, recommend one"). The errand needs planning and looking around. You still decide whether to accept the recommendation.

An agent is not a person. It does not know your priorities unless told, it can be confidently wrong, and it can repeat a mistake quickly.

Why it matters

- Agents can take on multi-step tasks, not just single answers, which saves more time.
- They can also make more mistakes before anyone notices, because they act between check-ins.
- Knowing what an agent is helps you decide when to use one and when a simple chat is enough.
- It makes clear where a person must approve: before anything is sent, paid, deleted or published.

How to do it

1. **Start** with a task that has a clear goal and a clear finish line, such as "collect and summarise".
2. **Give** the agent only the tools and access it needs for that task.
3. **Set** limits in writing: what it must not do, how long to spend, and when to stop and ask.
4. **Ask** it to show its plan before it begins, and correct the plan if needed.
5. **Check** the results and the sources it used, not just the final summary.
6. **Approve** any action with real-world effect yourself: sending, buying, changing records, deleting.

Example

Three common kinds of agent, with where the human sits:

- **Research agent.** Goal: "Find what public guidance exists on [topic] and summarise the main points with links." It searches, reads several pages, and writes a summary. You open the links to check the key claims.
- **Coding agent.** Goal: "Add a date filter to [this internal report script] and run the tests." It reads the code, makes changes, runs tests and reports back. A developer reviews the changes before they are used.
- **Browser agent.** Goal: "Fill in this internal booking form with the details below." It clicks and types in a web browser. You review the form before it is submitted.

A simple brief for a research agent:

Goal: Summarise publicly available guidance on [topic] for a one-page internal note.
Use: web search only. Do not log in to anything or fill in any forms.
Stop after: about 10 sources, or if you find conflicting guidance you cannot resolve.
Output: 5 to 8 bullet points, each with a link to its source. Mark anything uncertain.
Before you start: show me your plan in 3 to 5 steps.

Common mistakes

- **Treating an agent like a chatbot.** Give it a goal, limits and a finish line.
- **Giving it broad access.** Connect only what the task needs.
- **Skipping the plan.** Ask for the plan first; it is the cheapest point to catch a misunderstanding.
- **Checking only the final answer.** Look at the sources too.
- **Letting it act without approval.** Keep a person approving anything that sends, spends, changes or deletes.

Try it this week

1. Pick a small research question you need answered for work, with no confidential details (2 minutes).
2. Write a brief using the example above: goal, tools, limits, output (5 minutes).
3. If your approved assistant has an agent or research mode, run it and ask for the plan first (5 minutes).
4. Check two of its sources yourself and note whether they say what the summary claims (3 minutes).

Going further

- Agents work in a loop of plan, act, observe, adjust; asking for a short log of each step makes review much easier.
- Agents that read web pages or emails can meet planted instructions (prompt injection); keep them away from sending or changing things without approval.
- Open standards such as MCP (for connecting agents to tools and data) and A2A (Agent2Agent, for agents to talk to each other) are making agents easier to plug into existing systems, which makes access control even more important.

Key words

- **Chatbot:** an AI that answers each message once and then waits.
- **AI agent:** an AI that works towards a goal by planning, using tools and checking results over several steps.
- **Tool:** something the agent can use, such as web search, a file system or a browser.
- **Plan:** the list of steps the agent intends to take, which you can review first.
- **Approval point:** a moment where the agent must stop and wait for a person to say yes.