AI Agents Explained

An AI agent is a software system that uses an AI model to work toward a goal, by deciding its own steps and using tools to carry them out. Unlike a simple chatbot that only answers, an agent can act: it can search, read files, send messages or change records. Definitions vary, so this video uses a simple, common view. We follow Hadi, an operations manager at an online shop.

1. What an Agent Is

The key difference is the goal. With a chatbot, we ask a question and receive an answer. With an agent, we give a goal, such as finding late payments, and the agent chooses the steps itself. It keeps going until it decides that the goal is reached or that it needs help.

Example: Asking a friend for restaurant ideas gets a list, like a chatbot. Asking a helper to organise a birthday dinner means that the helper chooses and completes the steps, such as finding a place, booking a table and messaging the guests, like an agent.

At work: Hadi gives an agent the goal of finding all invoices that are 30 days overdue and reminding the customers. The agent chooses the steps itself, and Hadi reviews the result at the end.

2. Tools and the Loop

Agents work in a loop. First, the model plans the next step from the goal and from what it already knows. Then it acts by calling a tool, such as a search, a database query or an email system. Next, it observes the result and decides what to do next, and the loop repeats until the goal is met. The tools and the information that the agent can see give it real abilities, so they should be chosen with care.

Example: A cook follows the same loop: plan the next step, add an ingredient, taste the dish and adjust. The loop continues until the dish tastes right.

At work: Hadi's agent plans to find late invoices and queries the invoice database. It observes a list of overdue accounts and drafts a reminder email for each one. It then checks the replies and plans the next step.

3. Levels of Autonomy

Agents differ in how much freedom they have, and one common way to see this is as levels. At the lowest level, the agent only suggests and a person does the work. At the middle levels, the agent acts but asks for approval before important steps, or acts alone on small tasks and

This document is a property of Asaseeyat Academy - © Copy Right Protected.

reports afterwards. At the highest level, the agent works alone for long periods, which is rare and needs strong controls.

Example: A new intern first suggests replies and a manager sends them. Later, the intern sends small messages alone but asks before anything important.

At work: Hadi starts the invoice agent at the approval level, so it drafts the reminders and he approves each one. After a few weeks of good results, it sends reminders for small amounts alone and still asks him before contacting big customers.

4. Risks and Controls

Agents bring new risks. They can make mistakes with real consequences, such as sending a wrong email or deleting a record. They can also be tricked by instructions hidden in the content that they read, which is called prompt injection, or they can reveal data if they have too much access. Good controls include giving the least access needed, asking for approval before risky or irreversible actions, limiting time and cost, and keeping a log of every action. Frameworks such as the NIST AI Risk Management Framework help organisations govern, map, measure and manage these risks.

Example: Before giving a house key to a helper, we decide which rooms the helper may enter, ask for a check before any spending and keep the receipts. The helper can still be useful, with limits and a record.

At work: Hadi's agent can read invoices and send reminders, but it cannot issue refunds or change prices. It asks for approval before any email to a big customer, and every action is written in a log. Hadi first tests the agent on a small group of customers.

Conclusion

In summary, an AI agent pursues a goal by looping through plan, act and observe, using tools to take real actions. Four steps put it to work: know what an agent is, understand the tools and the loop, choose the level of autonomy and set controls for the risks. Used with clear limits and human oversight, agents can save time on multi-step work without losing control.

Sources

- NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) – https://www.nist.gov/itl/ai-risk-management-framework
- Anthropic (2024). Building effective agents – https://www.anthropic.com/engineering/building-effective-agents
- OpenAI (2025). A practical guide to building agents – https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf

This document is a property of Asaseeyat Academy - © Copy Right Protected.