What is an agent?

Rowan Udell
Cloud Security Consultant

It's software

...just not like other software

Traditional software *(deterministic)*: Same input, same output, every time

An agent *(non-deterministic)*: Same input, different path, different result

- Not a bug: it's how reasoning works
- The model decides the path, not the code

Is it an agent?

Chatbot

- Responds to input
- Takes no external actions
- Observes no results
- **Not** an agent: no loop - question in, answer out

Is it an agent?

Thermostat

- Has a goal (maintain temperature)
- Takes action: heat on or off - one rule, always the same
- Observes results (reads the temperature back)
- **Not** an agent: one rule, not reasoning - it can't decide, it just triggers

Is it an agent?

OpenClaw

- Has a goal (whatever you give it)
- Takes actions: browser, calendar, files
- Observes results
- Decides what to do next
- **Agent**: goal, action, observe, decide - all four

An agent...

- Has a **goal**
- Takes **actions**
- Observes **results**
- Decides **what to do next**: repeat or stop

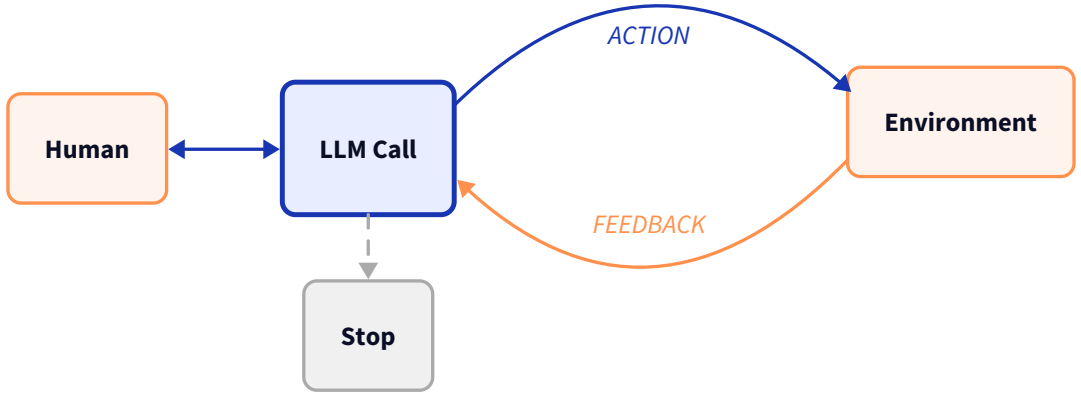
Not an agent: it's a chatbot

Input in. One LLM call. Output out.

e.g. ChatGPT: you ask, it answers.

It's an agent

Act on the environment. Observe feedback. Decide what to do next. Repeat until done.



The model is the brain

The model reads the situation, reasons through it, and decides what to do next.

- Reads: the goal, every prior action, every result, available tools
- Thinks: what worked? what failed? what fits next?
- Decides: act, or stop

Tools are the hands

The model names the tool and inputs. The loop runs it and returns the result.

```
// Model outputs a structured tool call:
{ "tool": "book_restaurant", "name": "Gino's", "date": "Saturday", "time": "7pm", "party": 2 }

// Loop runs it. Model receives:
{ "result": "Confirmed. Table for 2 at 7pm Saturday. Ref #4821." }
```

- Common tools: search the web, book a table, edit a file, write code, send an email

Stopping condition

Every loop needs a way to **stop**

Vague

- "Plan my trip"
- "Write a better email"
- "Research the topic"

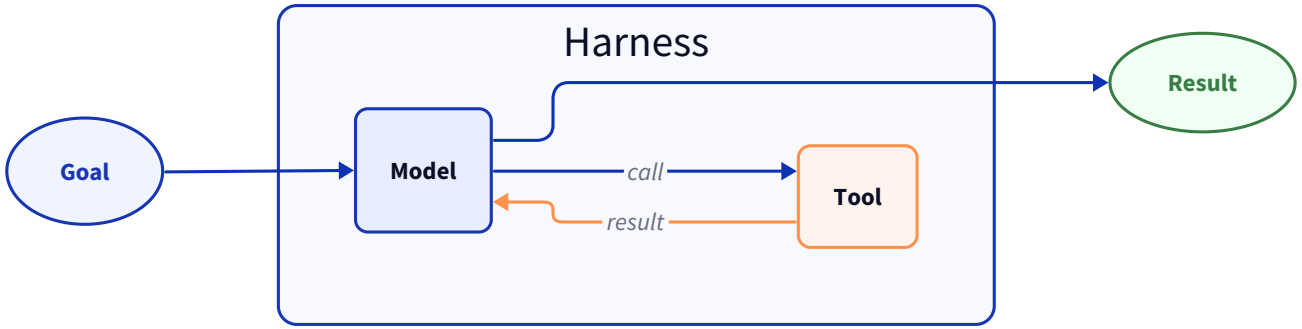
Concrete

- "Book cheapest return flight, under $600"
- "Rewrite to be direct, under 100 words"
- "Summarise in 3 bullet points"

Demo

terminal

The plumbing is the hard part



What you just saw is the loop: one model call, one tool, repeat until done

- Controls what the model can see and do
- Production adds: errors, context, memory, retries
- That demo: an afternoon. Production: weeks of plumbing.

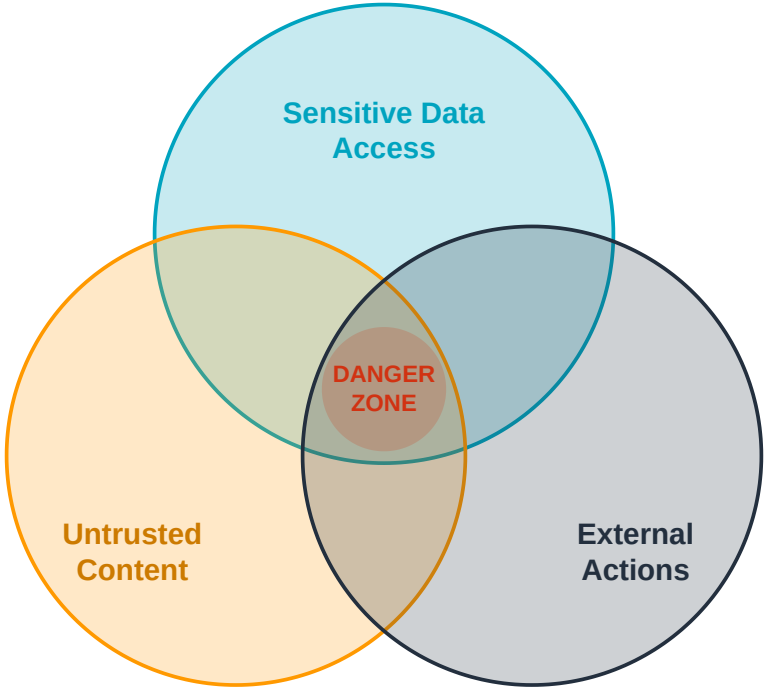
Security is an issue

The Lethal Trifecta

Simon Willison, 2025

- **Inputs** it receives
- **Data** it can access
- **Tools** it can use to make changes

All three together: dangerous.



A shared mental model

Now you agree on what an agent is...

- How do you trust it?
- What is it allowed to do?
- How do you hold it accountable?

Questions? Or connect on LinkedIn