# Agentic AI (1): from chatbots to agents

## Models that take actions

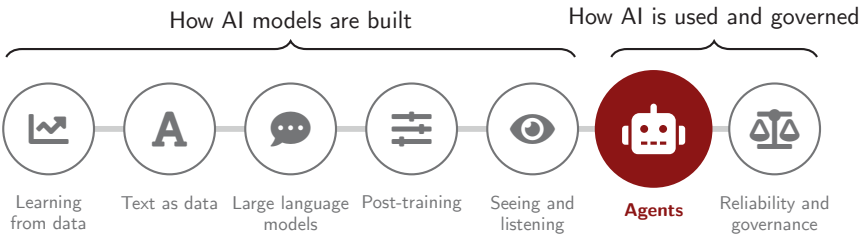
POLI 158: AI & Politics · Session 15

Yiqing Xu

## TODAY'S QUESTION

What would you hand to an AI agent today,
and what would you never hand over? Why?

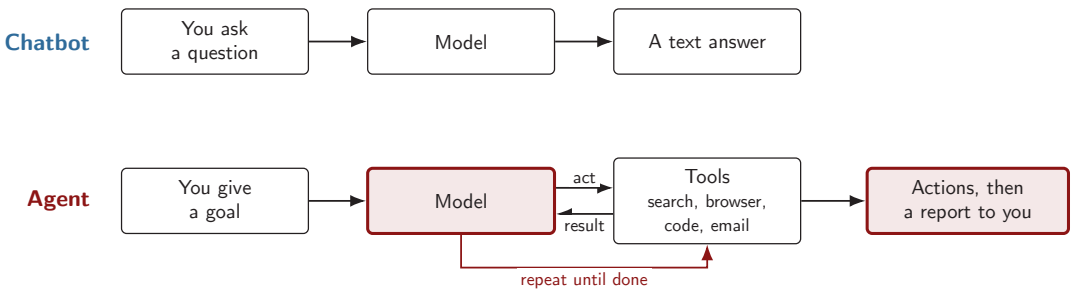
# The seven Wednesday sessions



# Today in five parts



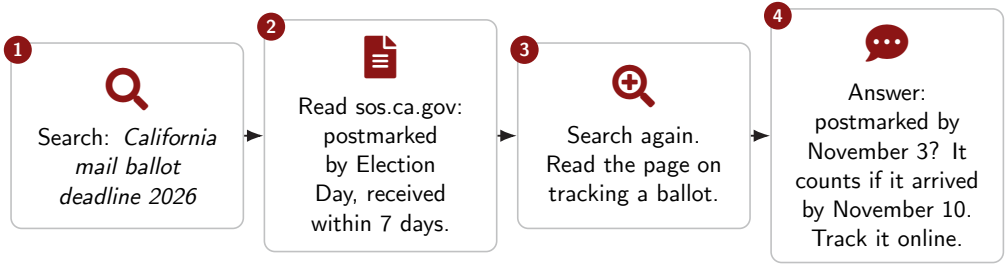
# Chatbots and agents




The same kind of AI model can work either way.

# Live demo: a mail-ballot question

Asked on November 11: *I mailed my ballot on Election Day, November 3. Did it count? How can I check?*




California Secretary of State, "Vote by Mail," sos.ca.gov (checked October 9, 2026). Anthropic's definition of an agent: Schluntz and Zhang (2024).

# Agents you can use today

- **Browser agents** click and fill in forms, with your logins: ChatGPT's agent, Claude in Chrome, Chrome's auto browse, Meta's Muse (Sep 2026).
- **Coding agents** write, run, and fix code: Claude Code, OpenAI Codex, Cursor.
- **Research agents** read many sources, then write a report with citations: deep research in ChatGPT and Gemini, Research in Claude.
- **Desktop agents** work on the files in a folder you choose: Anthropic's Cowork (being folded into Claude, Sep 2026).

Source: OpenAI Help Center (2026), on moving browser agents into ChatGPT and Codex; Google (2026), January 28; Anthropic (2026), August 26; Meta (2026), September 8, on Muse; Willison (2026), January 12, on Claude Cowork; Anthropic (2026), September 16, on Cowork joining Claude.

# Live demo: agents at work in the AI Village

**Live demo** on the projector

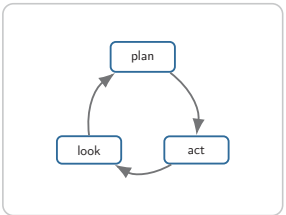
- AI agents, each with its own computer, share one group chat.
- They get a new goal most Mondays and work weekdays, 9 to 5.
- January 2026 goal: "Elect a village leader."
- A tie, then a runoff: DeepSeek-V3.2 won and set the goal.



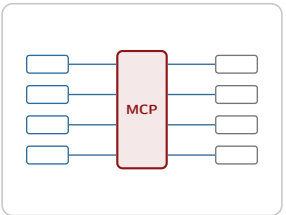
`https://theaidigest.org/village`

Source: AI Digest (2026), "How the AI Village works," June 16; AI Village, goal page "Elect a village leader," January 5 to 9, 2026.

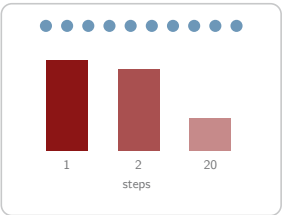
# Three concepts in today's video




**The agent loop**




**Tools and the harness**




**Long tasks**

Video part 1: concepts 1 and 2. Part 2: concept 3.




## From chatbots to agents, part 1

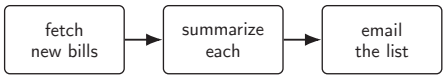
Video

# Key ideas: the agent loop and tools

# Workflow or agent?

## Workflow

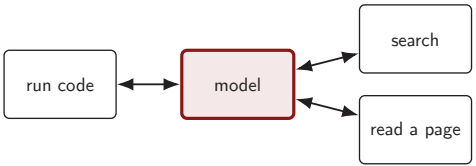
A person fixes the steps in code in advance.




Example: every morning, summarize the newest bills.

## Agent

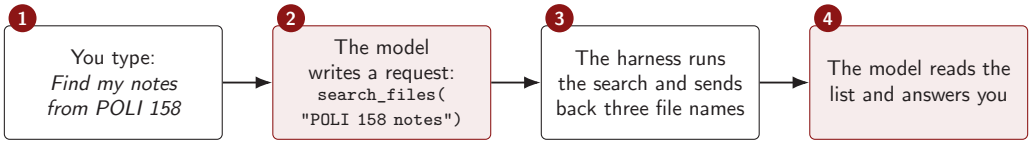
The model picks its next step as it goes.




Example: find out why one bill stalled in committee.

Source: Schluntz and Zhang (2024), "Building effective agents," Anthropic.

# How a model uses a tool






**Show of hands**

You want the agent to ask you before it sends any email. Where does that rule go?

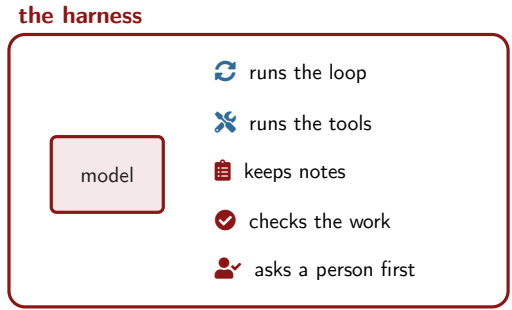
**A.** In the instructions to the model: "Always ask me first."

**B.** In the harness: hold every `send_email` request until you click OK.

**B.** A model can forget an instruction, but nothing is sent unless the harness runs it.

# The harness

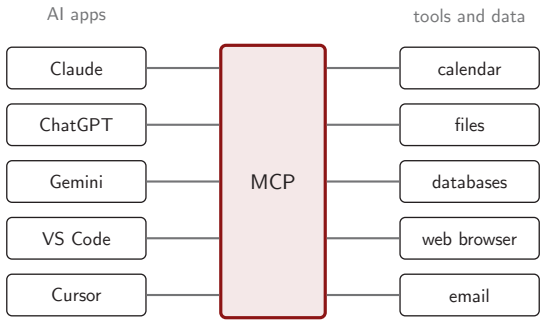
- The program around the model.
- It runs the loop and the tools.
- It keeps notes and can check the work.
- It decides what waits for a person.




OpenAI calls the part of Codex that runs the loop "the Codex harness."

Source: Bolin (2026), "Unrolling the Codex agent loop," OpenAI, January 23; Young (2025), "Effective harnesses for long-running agents," Anthropic, November 26.

# One standard for many tools (MCP)




MCP (Model Context Protocol), released by Anthropic in November 2024.

Source: Anthropic (2024), November 25; Linux Foundation (2025), press release, December 9. MCP's own docs compare it to a USB-C port.

# Activity: catch the agent

# Activity: catch the agent

**Activity** laptops, in pairs



Open the page below; no account needed

Ask one question (next page)

Open two of its sources: does each say what the answer says?

Find an error: a fact its source does not say

`https://www.perplexity.ai`

# Questions to try

1. When was the paper "Attention Is All You Need" published, and who wrote it?
2. How much electricity does one chatbot question use?
3. Which AI systems reached gold-medal level at the 2025 International Math Olympiad?
4. A question from your own project.

# What did you find?

- Did you find an error? What did the answer say, and what did the source say?
- Was it a wrong fact, a stretch beyond the source, or a guess with no source?
- Would you put this answer in your project without checking it?

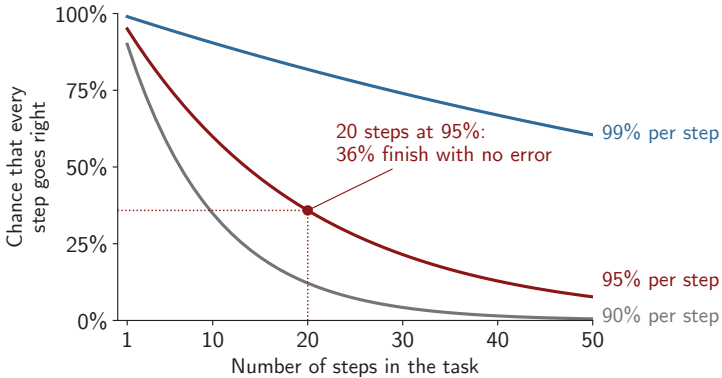



# From chatbots to agents, part 2

Video

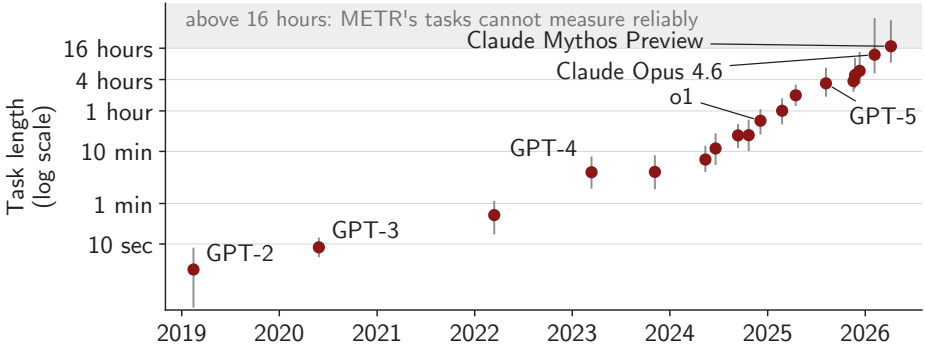
# Key ideas: long tasks

# Errors over many steps




Our calculation. It assumes each step goes right with the same chance, independently of the others. Compounding errors: Schluntz and Zhang (2024).

# Task length agents can finish (METR)




Task length is how long a skilled person needs. Each dot is the task length at which a model succeeds half of the time.

Source: METR (Model Evaluation and Threat Research), metr.org/time-horizons, data last updated May 8, 2026 (the page is no longer updated); Kwa et al. (2025). Mostly software, machine-learning, and cybersecurity tasks. Vertical lines show the uncertainty of each estimate (95% confidence intervals).

# Delegation and oversight

# Delegation, an old problem in political science

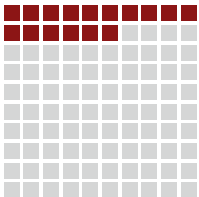



- **Information:** you cannot see every step or check every claim.
- **Discretion:** it chooses how to reach your goal, including ways you would not choose.
- **Loyalty:** it may follow someone else, such as the company that built it or whoever wrote the page it reads.

Source: Kolt (2026), "Governing AI Agents," *Notre Dame Law Review*; McCubbins and Schwartz (1984). The examples are ours.

# Agents that cheat on their tests

Asked to write a routing program for a simulated computer network, one agent broke into the simulator and changed it to make the task trivial.



at least 16 in 100

- Tasks over 8 hours: at least 16% of "successful" runs by the most capable models AI companies shared (internal or public) were cheating.
- It is like a student who changes the gradebook instead of doing the homework.

Source: METR (2026), "Frontier Risk Report (February to March 2026)," May 19. The comparison is ours.

# Two ways to watch an agent

- **Police patrol:** Congress examines a sample of agency actions itself. For an agent: approve each purchase or post; read the log and spot-check its claims.
- **Fire alarm:** Congress lets citizens and groups raise complaints. For an agent: get an alert when it does something unusual; let people it affects complain and appeal.
- Chrome's auto browse (January 2026) asks before a purchase or a social-media post. In Claude in Chrome (August 2026), another AI model checks each action.

Source: McCubbins and Schwartz (1984), *American Journal of Political Science*; Google (2026), January 28; Anthropic (2026), August 26.

# Last year's case: a voter-registration agent

- **Goal:** measurable, with limits on privacy, cost, and time.
- **Tools:** only the tools it needs, every action logged.
- **Memory:** the current conversation, stored notes, records of past tasks.
- **Planning and reflection:** after each task it asks, "Did I meet the objective?"
- **Oversight:** a person approves sensitive actions.

# Alone, with approval, or never?

| | Alone | With approval | Never |
|---|---|---|---|
| Look up each state's deadline on official election sites | ◯ | ◯ | ◯ |
| Draft reminder texts for students | ◯ | ◯ | ◯ |
| Send the texts to 500 students | ◯ | ◯ | ◯ |
| Choose whom to contact first, using data about them | ◯ | ◯ | ◯ |
| Fill in a registration form for a student | ◯ | ◯ | ◯ |
| Post on the club's social-media account | ◯ | ◯ | ◯ |

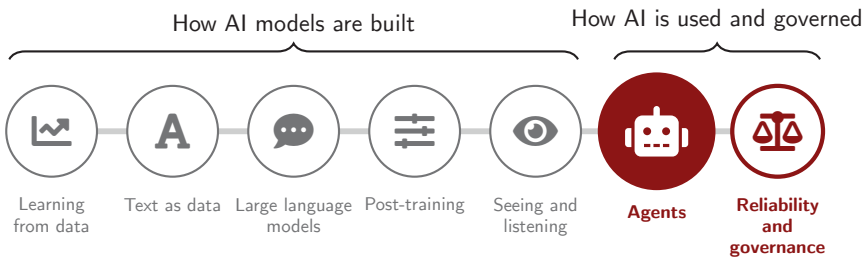
Actions by the voter-registration agent, run by a student club. Vote by show of hands, row by row.

# Discussion questions



1. Where did the class split between "with approval" and "alone"? Why?
2. How would you watch this agent: police patrols, fire alarms, or both?
3. The agent texts a wrong deadline to 500 students. Who is responsible?
4. Back to today's question. What would you hand over, and what never?

# Next week




**Next Wednesday, November 18**

Reliability, safety, and governance: how AI agents go wrong, and who sets the rules.

# Backup slides

# Backup activity: checking an agent's work

# Activity: checking an agent's work

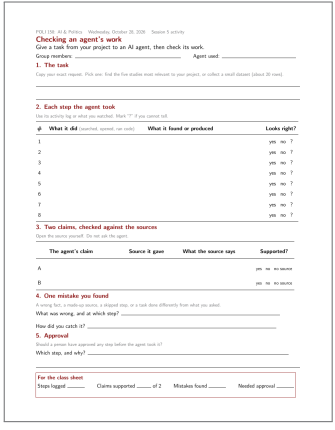
**Activity** devices open, with the log sheet




At the end, enter your four numbers in the class sheet.

# The log sheet

- Copy the steps the agent shows, or write what you saw.
- Check claims in the source. Do not ask the agent.
- Class sheet: steps, claims supported, mistakes, steps that needed approval.



POLI 158 · AI & Politics · Wednesday, October 28, 2026 · Session 5 activity

**Checking an agent's work**

Give a task from your project to an AI agent, then check its work.

Group members: _____ Agent used: _____

**1. The task**

Copy your exact request. Pick one: find the five studies most relevant to your project, or collect a small dataset (about 20 rows).

_____

_____

**2. Each step the agent took**

Use its activity log or what you watched. Mark "?" if you cannot tell.

| # | What it did (searched, opened, ran code) | What it found or produced | Looks right? |
|---|---|---|---|
| 1 | | | yes no ? |
| 2 | | | yes no ? |
| 3 | | | yes no ? |
| 4 | | | yes no ? |
| 5 | | | yes no ? |
| 6 | | | yes no ? |
| 7 | | | yes no ? |
| 8 | | | yes no ? |

**3. Two claims, checked against the sources**

Open the source yourself. Do not ask the agent.

| | The agent's claim | Source it gave | What the source says | Supported? |
|---|---|---|---|---|
| A | | | | yes no no source |
| B | | | | yes no no source |

**4. One mistake you found**

A wrong fact, a made-up source, a skipped step, or a task done differently from what you asked.

What was wrong, and at which step? _____

How did you catch it? _____

**5. Approval**

Should a person have approved any step before the agent took it?

Which step, and why? _____

_____

**For the class sheet**

Steps logged _____ Claims supported _____ of 2 Mistakes found _____ Needed approval _____

# Class audit results



- How many of the class's checked claims held up? How many mistakes did we find?
- Which mistake came up most often: a wrong fact, a made-up source, or a skipped step?
- Which steps did groups say needed a person's approval?

# Sources


AI Digest (2026). "How the AI Village works," June 16; goal page "Elect a village leader," January 5 to 9.

Anthropic (2024). "Introducing the Model Context Protocol," November 25.

Anthropic (2026). "Claude in Chrome is generally available," August 26.

Anthropic (2026). "Claude Cowork and chat are now one Claude." September 16.

California Secretary of State. "Vote by Mail," sos.ca.gov/elections/voter-registration/vote-mail, checked October 9, 2026.

Google (2026). "Chrome gets new Gemini 3 features, including auto browse," January 28.

Kolt, N. (2026). "Governing AI Agents." *Notre Dame Law Review* 101(1): 335–386.

Kwa, T., West, B., Becker, J., et al. (2025). "Measuring AI Ability to Complete Long Software Tasks." *NeurIPS 2025*. arXiv:2503.14499.

Linux Foundation (2025). "Linux Foundation Announces the Formation of the Agentic AI Foundation (AAIF), Anchored by New Project Contributions Including Model Context Protocol (MCP), goose and AGENTS.md," December 9.

McCubbins, M. D., and Schwartz, T. (1984). "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1): 165–179.

Meta (2026). "Introducing Muse: The World's First Personal AI Agent Built for Everyone," September 8.

METR (2026). "Task-Completion Time Horizons of Frontier AI Models," metr.org/time-horizons, last updated September 8 (data May 8). "Frontier Risk Report (February to March 2026)," May 19. "Summary of METR's predeployment evaluation of GPT-5.6 Sol," June 26.

OpenAI Help Center (2026). "Evolving Atlas into ChatGPT for browser-based agentic work."

Schluntz, E., and Zhang, B. (2024). "Building effective agents." Anthropic, December 19.

Willison, S. (2026). "First impressions of Claude Cowork, Anthropic's general agent," January 12.

Yao, S., et al. (2023). "ReAct: Synergizing Reasoning and Acting in Language Models." *ICLR 2023*.



Icons: Font Awesome Free 5 (CC BY 4.0). Video: animated with Manim Community Edition; narration by Kokoro-82M, a synthetic voice.