

1 - The three words, defined

Generative AI	Makes content from a prompt, then stops. Reactive. One model call.
AI Agent	A goal + tools + freedom to pick its own next step. One working unit.
Agentic AI	Goal-directed behaviour; often several agents coordinated by an orchestrator.

2 - The agent loop (repeats until the goal is met or a stop condition fires)

1. PERCEIVE	take in the request + the latest tool result (detections, transcript)
2. REASON	decide the next step; write a short plan
3. ACT	run a detector, call an API, post a message - the only step that changes things
4. OBSERVE	read the result, feed it back to step 1 - this is what generative AI cannot do
STOP	goal met OR step limit / budget cap / timeout / 'ask a human now'

3 - Should you build an agent? Pick the simplest that works

One answer is enough	-> a single GENERATIVE call (caption, summary, thumbnail)
Steps never change	-> a WORKFLOW: orchestrate tools in code you wrote up front
Next step depends on the last	-> an AGENT: let the model choose the path (costs 5-20x more)

4 - Four video agent patterns (same loop, different goal + tools)

Investigator	archive	answer a question, return evidence clips
Meeting copilot	live call	act during + after a call (ASR, calendar, CRM, docs)
Async reviewer	backlog	tag / moderate / QC overnight, flag only exceptions
Real-time	live stream	watch the stream, trigger an alert or webhook

5 - Governance checkpoints before you ship (engineering context, not legal advice)

- Disclose the AI - EU AI Act Article 50 requires telling people they are dealing with an AI.
- Give every loop a budget cap and a step limit so it cannot spin or spend without bound.
- Keep an audit trail of each perceive-reason-act-observe step for later review.
- Put a human in the loop for any high-stakes autonomous action (surveillance, medical).
- Start with one agent; add more only when one provably cannot hold the whole task.