AI Agents 101

Bounded Agency Over Autonomous Agents

Practical guidance for teams navigating AI hype

The Problem

Everyone is talking about AI agents. Few agree on what the terms mean.

When someone says "let's use an AI agent for this" — does everyone in the room picture the same thing?

Usually not.

The Salesforce Lesson

- **September 2025:** Cut ~4,000 support staff citing AI product "Agentforce"
- **December 2025:** Leadership had been "more confident than they should have been"
- Service gaps, quality issues, reliability problems followed

They sell the AI product. They still got caught by the gap between expectations and reality.

What We'll Cover

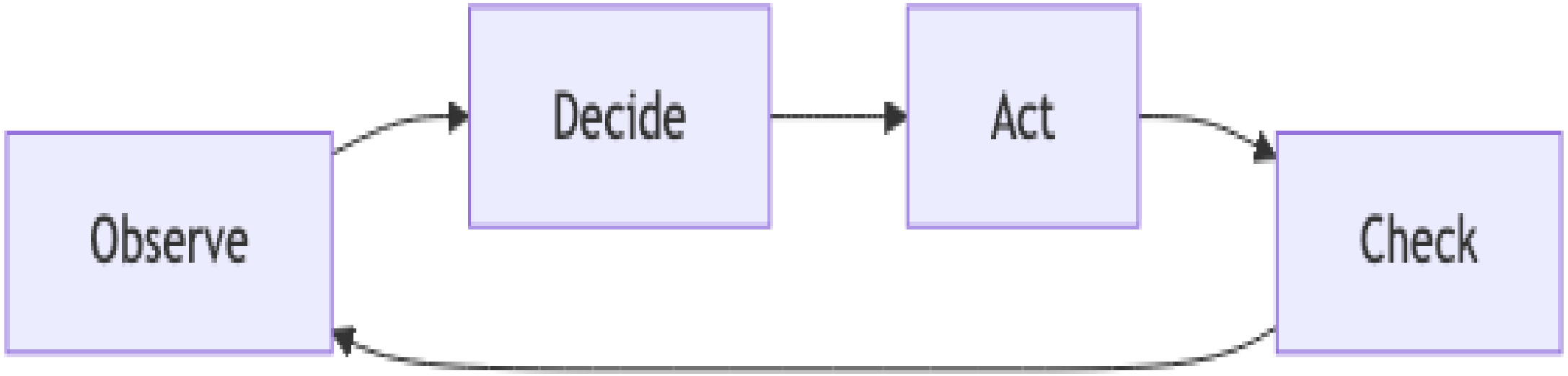
1. **Vocabulary** — shared definitions for the terms that matter
2. **Live Demo** — see the concepts in a simple workflow
3. **Case Study** — bounded agency at scale (briefly)
4. **Design Principles** — what to remember when the details fade

Core Vocabulary

Term	What It Is	Key Question
Workflow	Structure, sequence of steps	What's the process?
Agent	Entity that does work	What's doing the work?
Agency	Granted decision-making authority	What decisions can it make?
Agentic	Behavior where agency is exercised	How much can it adapt?
Tool	Single discrete operation	What can it do?

The Loop

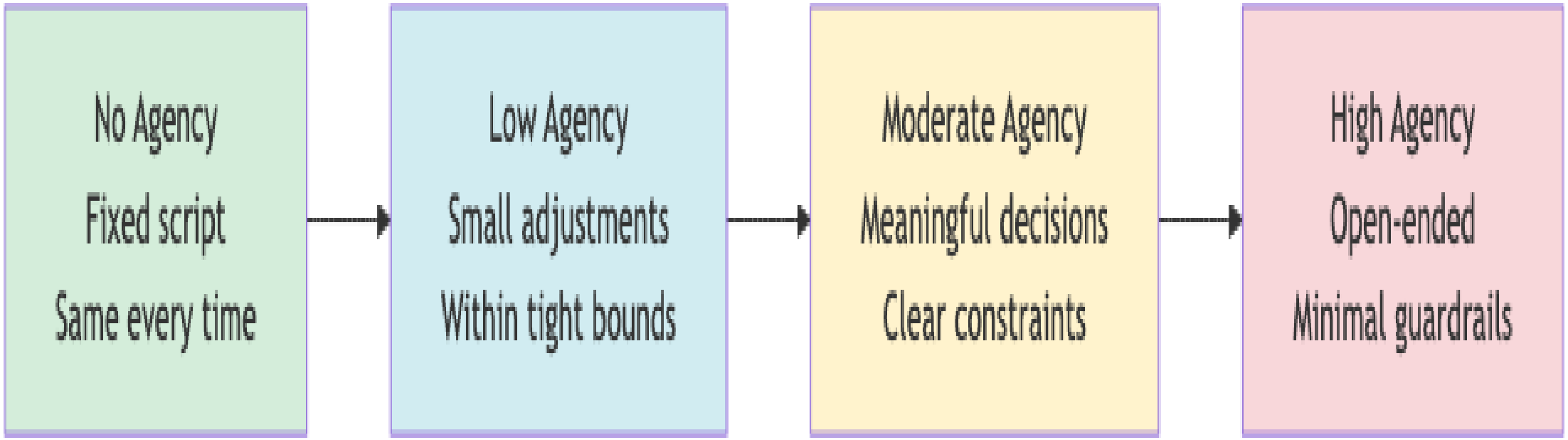
The heartbeat of agent behavior:



Observe → Decide → Act → Check → (repeat)

Every framework, every tool — this pattern remains constant.

Agentic Is a Dial, Not a Switch



Move right only when you have to. Stay left when you can.

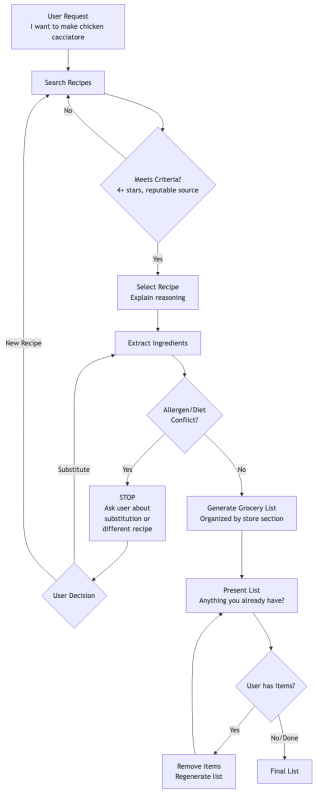
The Loop in Practice

A recipe assistant — simple enough to understand, complex enough to be useful.

Phase	What Happens
OBSERVE	User says "I want to make chicken cacciatore"
DECIDE	Which recipe meets criteria? (4+ stars, reputable source)
ACT	Search → Select → Extract ingredients → Generate list
CHECK	Allergen conflict? User has items? Goal met?

The loop repeats until done — or until it needs to stop and ask.

Live Demo: Recipe Workflow



View full diagram

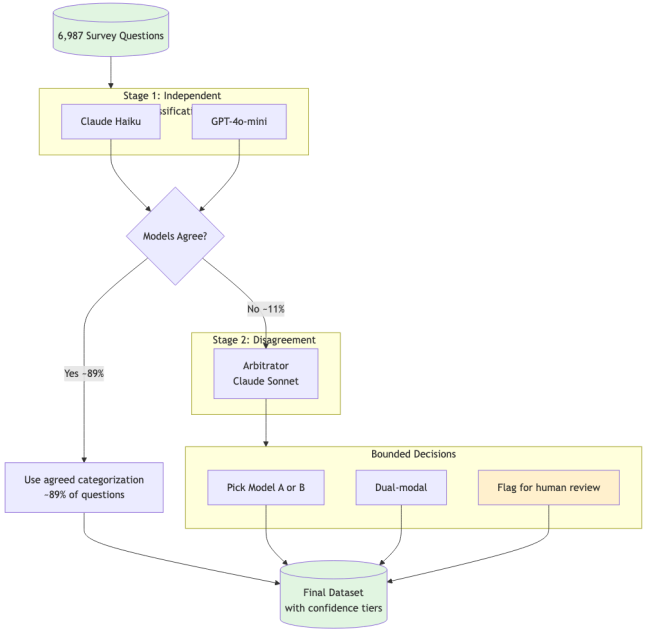
What Makes It Agentic?

- **Decision points with criteria** — which recipe? (4+ stars, reputable source)
- **Conditional branching** — allergen check is a hard stop
- **Iteration** — "anything you already have?" creates a refinement loop
- **Transparency** — "explain why you chose it"

This is a **workflow with bounded agency**, not an autonomous agent.

Case Study: Multi-Survey Concept Mapper

6,987 questions across 46 surveys → mapped to official taxonomy



View full diagram

Why Bounded Agency?

Autonomous Agent Approach	This Workflow
One model decides everything	Cross-validation catches errors
Confidence is implicit	Confidence tiers are explicit
Edge cases get guessed	Edge cases get flagged
Failures are silent	Failures are surfaced

Result: 99.5% success rate, ~$15 total cost, complete audit trail.

Design Principles

#	Principle	One-Liner
1	Good judgment upfront	Design quality bounds output quality
2	Autonomy is governance	Less autonomy is often better
3	Most problems don't need agents	Simple solutions beat complex ones
4	Specification is the skill	Clarity beats capability
5	Design for uncertainty	Plan for failure, not just success
6	Digestible chunks	Focused beats sprawling

The Stakes: What Ignoring These Causes

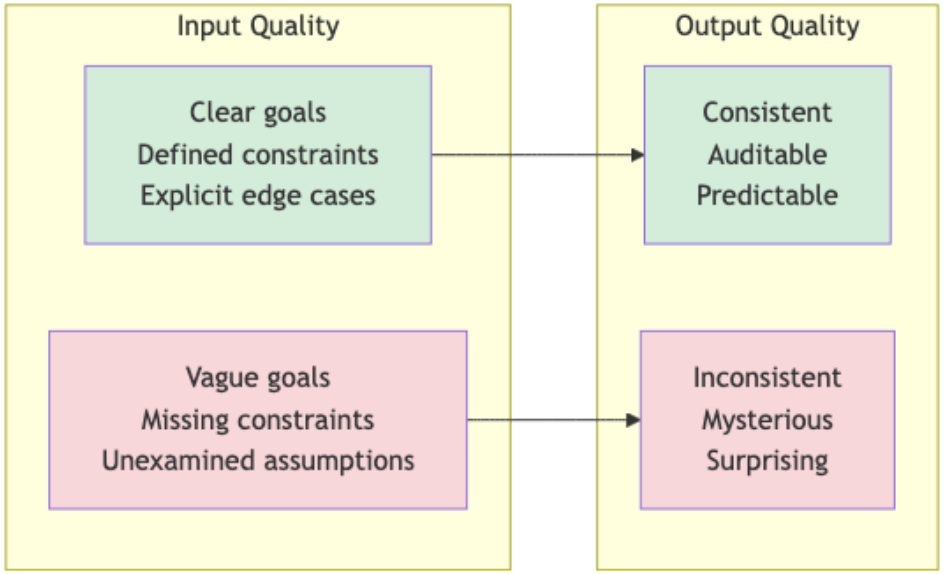
Microsoft's AI Red Team documented failure modes in agentic systems.

Nearly every one traces back to violating these principles:

Ignore This...	...And Get This
Good judgment upfront	Misalignment, hallucinations, misinterpretation
Autonomy as governance	Actions outside scope, user harm
Simple solutions first	Attack surface, knowledge loss
Clear specifications	Wrong permissions, accountability gaps
Designing for uncertainty	Cascading failures, denial of service

These principles aren't caution. They're how you build systems that work.

The Meta-Principle



AI amplifies your process. Good process + AI = faster good outcomes.

The Skill Is Specification

When reviewing agent behavior, ask:

1. Did it stay within its job description and data boundaries?
2. Is the action correct in context?
3. Is the rationale understandable?
4. **Would I sign my name under this action?**

The Bottom Line

You don't need expensive tools or deep technical skills.

You need:

- Clear thinking
- Well-structured prompts
- Appropriate skepticism

Start simple. Add complexity only when justified. Design for uncertainty. Keep humans in the loop where it matters.

Resources

Student Handout: Vocabulary, templates, copy-paste recipe prompt

Course Notes: Full speaker notes and reference material

Code: github.com/brockwebb/federal-survey-concept-mapper *(case study source)*

Further Reading:

- Microsoft AI Red Team, "Taxonomy of Failure Modes in Agentic AI Systems" (2025)
- OpenAI, "A Practical Guide to Building Agents" (2025)
- Anthropic, "Building Effective Agents" (2024)

Questions?

Backup: Key Terms Reference

Reference slides for selected vocabulary

Artificial Intelligence (AI)

A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

https://csrc.nist.gov/glossary/term/artificial_intelligence

Context Window

The maximum span of text (measured in tokens) that a language model can process at once—including both input and output. Functions as the model's working memory.

https://en.wikipedia.org/wiki/Large_language_model

Hallucination

A common but imprecise term for AI confabulation (see above). Widely used in industry and media.

Key point: Hallucination implies perceiving something that isn't there. LLMs don't perceive—they generate. Confabulation (generating plausible content to fill gaps) is the accurate descriptor. The term "hallucination" persists because it's convenient and dramatic, not because it's correct.

https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

Few-Shot Prompting

Providing a small number of examples (typically 2-5) in your prompt to demonstrate the desired pattern before asking the model to perform on new input.

https://arxiv.org/abs/2005.14165

Retrieval-Augmented Generation (RAG)

A generative AI approach where a model is paired with an information retrieval system. Retrieved information is incorporated into the prompt to ground the response in actual sources.

Key point: Reduces hallucination. Enables AI to answer questions about your documents.

https://csrc.nist.gov/glossary/term/retrieval_augmented_generation

Bias (in AI)

Systematic errors in AI outputs that can result in unfair outcomes, such as privileging one group over others.

https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence