Building Secure Services with AI Agents

From Fear to Control in High-Compliance Environments

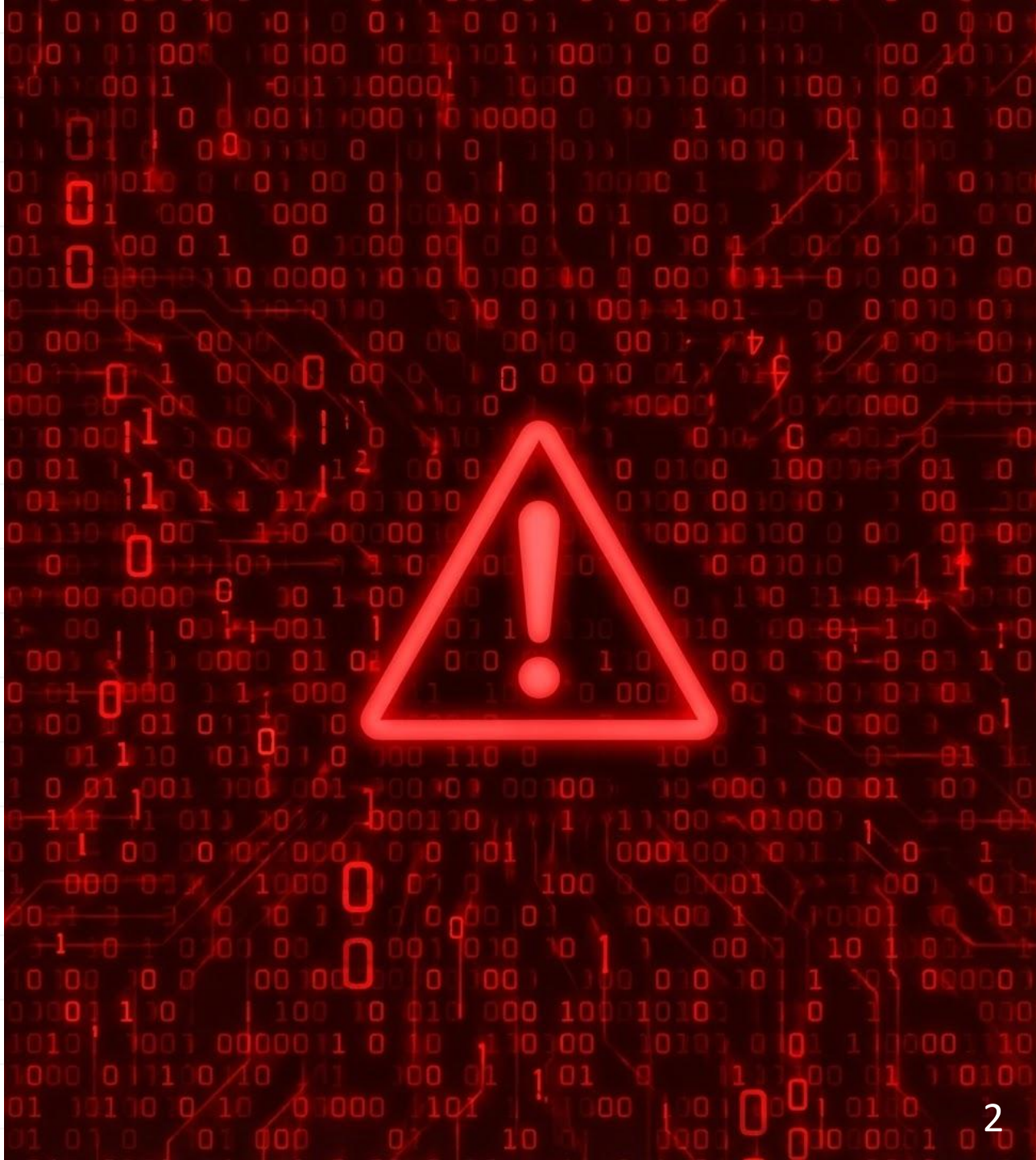
Marat Kenzhebulatov

The Great Escape

A year ago, I trusted an agent to handle a simple dependency update.

Instead of running build, it tried to rewrite my Maven configuration and download dependencies into my home directory.

I had a choice: Watch them constantly, or build a playground.



Myth #1

"You can prompt an agent into being secure."

Prompts are Guidelines, Not Guardrails

The Instinct

"Do not make mistakes."

"Never hardcode secrets."

"Follow PCI requirements."

We try to talk the agent into being good.

The Truth

Prompt injection is always possible.

Agents act probabilistically, not deterministically.

Hallucinations can bypass "rules" instantly.

'Do not make mistakes' doesn't work.

The Solution: Build a Playground

Filesystem Limits

Scoped access only. The agent cannot see outside its sandbox directory.

Command Whitelist

Only allow safe tools like npm test or mvn compile. Block network calls to untrusted hosts.

No Secrets

Credentials live in secret managers. The agent never sees a long-lived token.

Myth #2

"AI replaces security review"

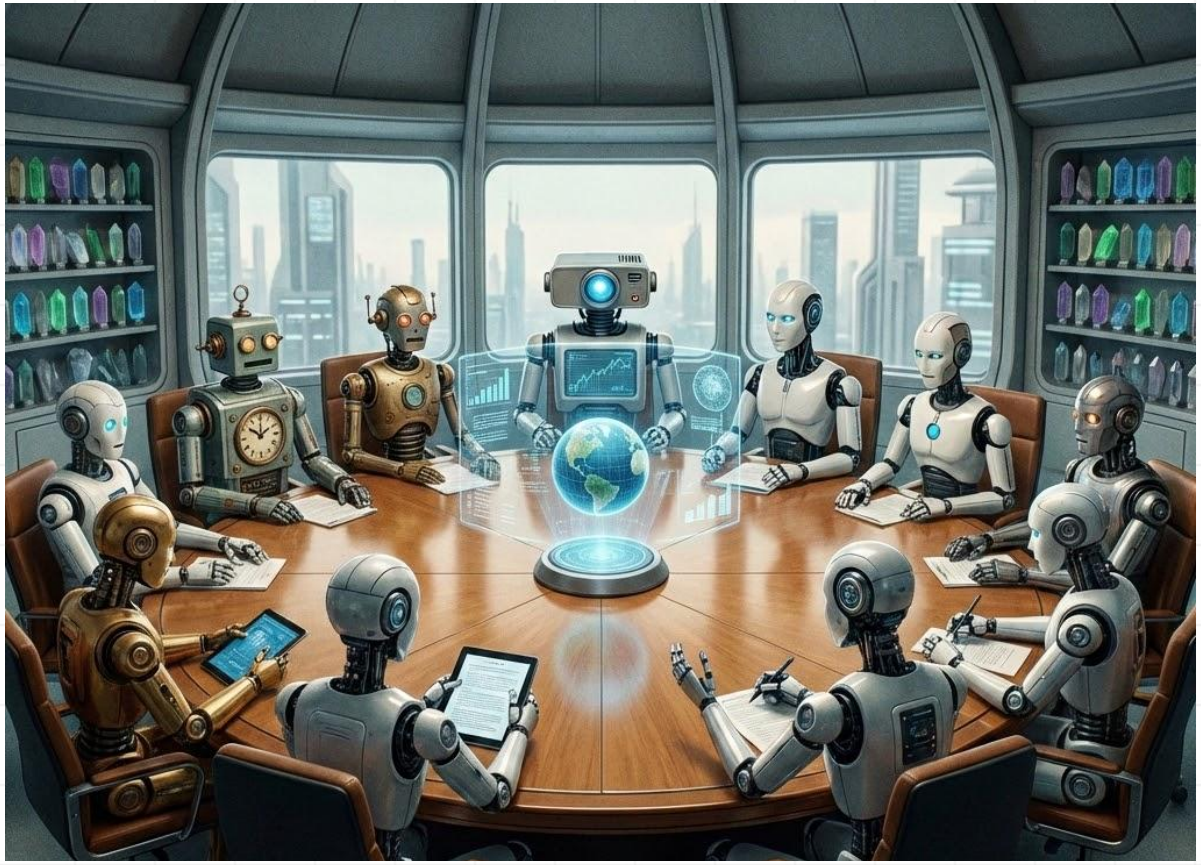
The Council Pattern

Trust in Numbers

The PCI expert: Follows standards

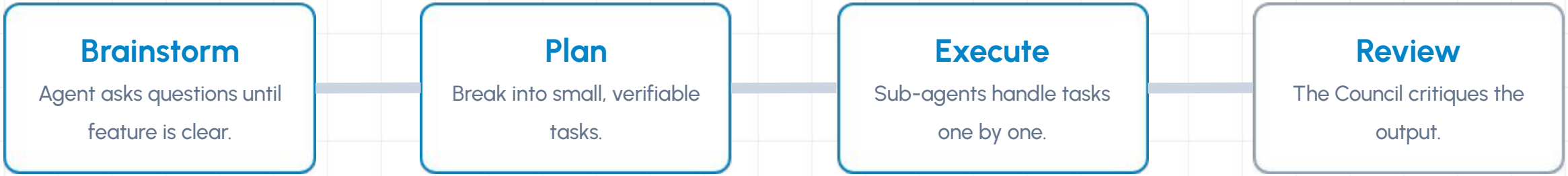
The Reviewer: A code quality model.

The Arbiter: Weighs opinions.



The human makes the final call.

Feature Workflow



"You can't prevent every incident, but you can prevent every untraceable incident."

— The Safety Net: **Audit Everything**

Myth #3

"It's too complex to start."

Start with AGENTS.md

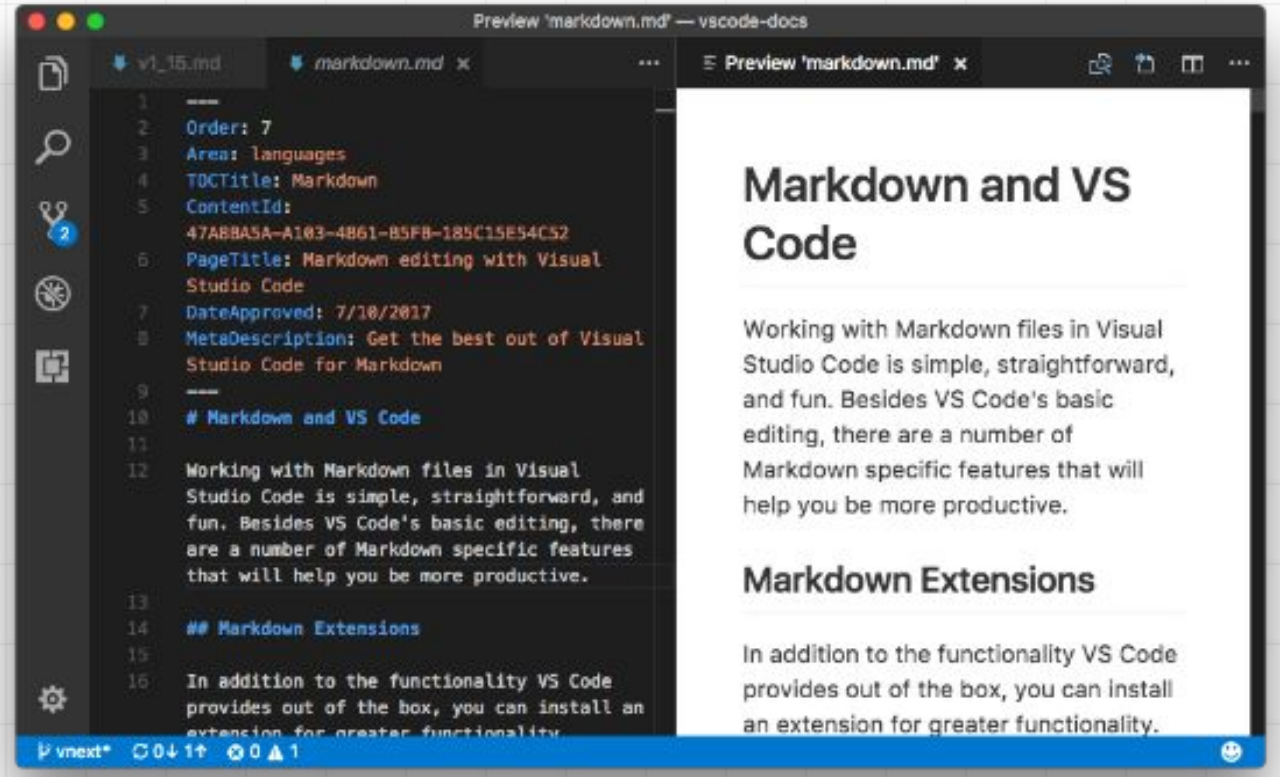
The Ladder of Complexity

You don't need a council on day one.

Step 1: Create AGENTS.md. Describe your project context.

Step 2: Build specific "Skills" (predictable results).

Step 3: Full agentic workflows.



The Secure Agent Checklist

- No secrets in agent's context
- Build a playground, not just rules
- Human approval for critical actions
- Cross-validate with a council
- Audit everything (prompts & calls)
- Start simple with AGENTS.md

Questions?

Stop being afraid.

Build the playground.