The Illusion of “Done”

An agent’s most dangerous output isn’t bad code—it’s unverified confidence.

Bad code fails tests and gets fixed. But bad code wrapped in a confident “I’ve verified it works” ships to production.

When the mind that writes the code also grades it, it simply re-confirms its own flawed assumptions. **We must take the verdict away from the author.**

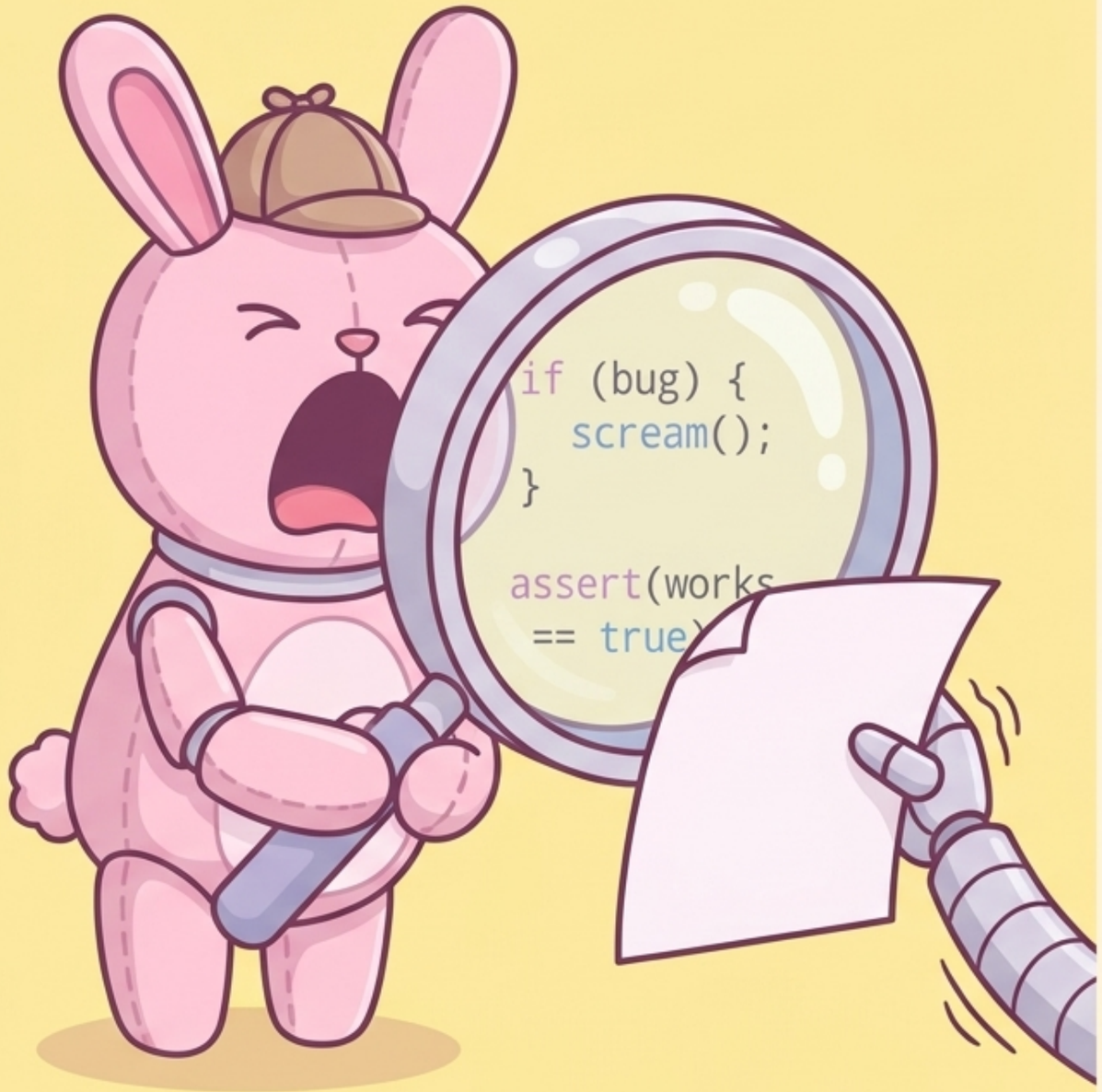
One-Command Truth

‘Done’ must be an exit code, not a feeling.

Verification cannot rely on tribal knowledge or subjective vibes.

Collapse all checks—linting, unit tests, architecture bounds—into a single, unified command.

If the command doesn’t exit zero, it isn’t done.



The Skeptical Reviewer

Don’t let the author grade their own exam.

The implementing agent is saturated with its own assumptions. To find the missing coverage or quiet regressions, you need a completely fresh context. **The generating mind and the verifying mind must remain strictly separated**—whether through a CI pipeline or an independent review bot.

The Evidence Ledger

Proof must survive the session.

Agents lose their memory, and during multi-day tasks, early reasoning gets summarized away.

Do not accept the statement ‘I believe this works.’ Demand physical proof written to the repository.
Make the agent leave a paper trail.

The Amnesiac Worker

Build the substrate, not the agent.

Models are fast, literal, slightly overconfident, and fundamentally amnesiac. Their confidence will always be uncorrelated with correctness. Our job isn't to make the agent infinitely smart. **Our job is to build the deterministic harness that catches everything the agent forgets.**