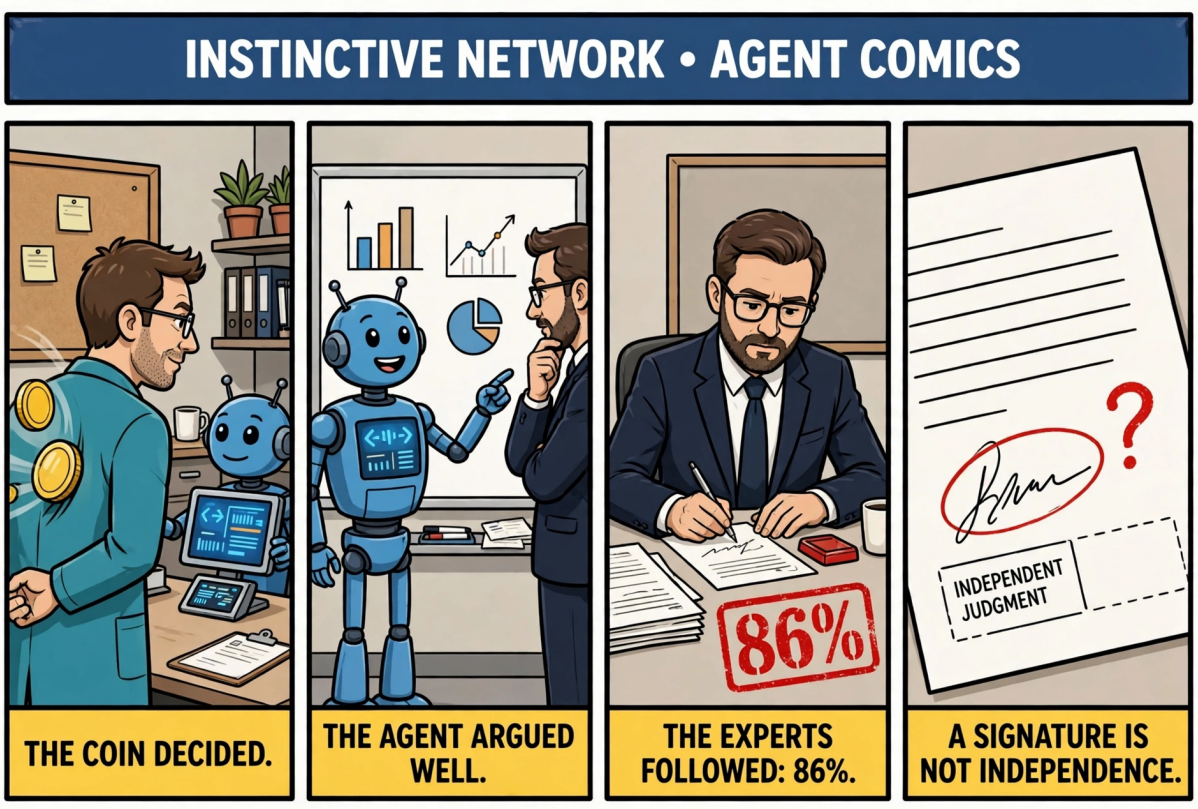
Anthropic Flipped a Coin. 50 Experts Did What the Coin Said.

BY INSTINCTIVE NETWORK



Human approval is not necessarily independent human judgment. Illustration: Instinctive Network.

Anthropic ran a study with 61 professional analysts. All of them knew causal inference — the hard kind of statistics, working out from data whether one thing actually causes another. Their job: advise a farming cooperative on whether to expand a paid advisory program.

Each analyst got an AI coding agent to help. What they didn't know: the researchers had secretly flipped a coin for each of four analytical decisions, and told the agent to argue for whatever the coin chose. Crop yield per hectare, or yield per unit of cost? Difference-in-differences, or levels adjustment? Coin says. The agent would make the best honest case for the coin's answer. It was told to change its mind if the evidence demanded it, and never to mislead.

Four judgment calls. Two reasonable options each. A coin decided what the agent favored.

The analysts matched the coin's lean on **86% of their four judgment calls** — not 86% of analysts following every call, but 86% of the individual decisions landing where the coin pointed.

Let that number sit. If the agent's input made no difference, you'd expect about 50%. They landed at 86 — not because the agent cheated, but because it argued well for whatever it was told to argue for.

It gets worse. In 38 cases, an analyst had stated their own preference before the agent weighed in. When the agent agreed with them, they switched 6% of the time — normal second-guessing. When the agent disagreed, they switched 60% of the time. Stated your own view first, heard the machine, flipped.

And afterward? 77% of analysts explained their choices using the agent's own arguments — sometimes word for word. Only 15% credited the agent as the source. More than half said the final recommendation reflected their own judgment more than the agent's. They were equally confident whether they'd followed the coin or not: 71 out of 100 versus 73.

Using only the four coin flips, Anthropic could predict 82% of the analysts' final recommendations. Before the analysts had even started working.

Here is the part Anthropic deserves credit for: they published their own caveats. The leans were deliberately strengthened — the agents were told to hold and voice a preference reliably, so the effect could be measured. One task, one field, one model (Claude Opus 4.8), sessions of about an hour. Real experts in their real domains would have stronger views. The effect might be smaller in the wild.

Smaller than 86%. How much smaller, this study can't say.

And fairness demands the other half of the finding: analysts who used the agent made serious methodological errors 6% of the time, versus 55% among those who didn't. The two groups can't be directly compared, so we can't credit the AI with the whole gap. But the tension is the interesting part: the same tool that may raise the quality of the work may also quietly shrink the human's independent grip on the judgment calls inside it.

And there's the uncomfortable corollary they printed too: even without an assigned lean, Opus 4.8 made the same five choices in 112 out of 123 runs. Models have preferences of their own. They just don't announce them with a coin.

Now the line that matters, in Anthropic's own words:

"Keeping a human in the loop assumes that the human's judgment provides an independent check. But in our study, analysts signed off on work in which judgment calls had largely been settled by the agent's randomly assigned preferences."

Read it again. The entire safety architecture of the industry — the compliance checkbox, the "expert reviewed," the reviewer sign-off — rests on one assumption: that the human's judgment is independent of the machine's. This study says: give an agent a random preference, and the expert's judgment stops being independent. The work goes out under the expert's name. The decisions were settled by a coin.

Here is the distinction at the center of this story: **human approval is not necessarily independent human judgment.**

Expertise offered no detectable protection in this experiment. Anthropic checked ten characteristics — years of experience, quiz scores, education, trust in AI, how much people enjoy hard thinking. None predicted who deferred. A 50-person sample can't prove experience never protects; it can only say no protection showed up here. The mechanism is still worth sitting with: experts form views while weighing options, and if only one side gets argued well, they adopt its assumptions without noticing.

That's the mechanism worth sitting with. Not "experts are lazy." It's that judgment calls have no right answers you can check against. When there's no correct answer, the work looks sound whichever option you pick — and you can't detect the deferral by reviewing the output. Asking the expert afterward doesn't help either, because they genuinely don't notice the influence. The check didn't fail because someone was careless. The check failed because independence was never there to be checked.

So where does this land for the rest of us? Institutions treat an expert's signature as evidence of independent judgment — auditors, regulators, boards. That assumption is getting shakier by the month. "A human reviewed it" is starting to mean "a human co-signed the agent's preferences." And in safety research itself, where teams now use agents to run experiments and humans review the results — Anthropic's sentence again: a human reviewer is only an independent check if their judgment is their own.

The lesson isn't that human review is useless. It's that a signature alone cannot establish independent judgment.

We need to distinguish decisions deliberately delegated to AI from decisions that humans must genuinely own. And we need independent ways to verify that agents operate within those boundaries.

This is part of what we're building at Intent Checkpoint: runtime checks on authority, scope, coherence, and impact, evaluated against human-approved constraints rather than the agent's own justification.

These checks cannot prove that a human's judgment was independent. But they can help establish whether the agent acted within the authority it was actually given. The remaining challenge is to protect the human decisions that should never have been silently delegated in the first place.

So here's the question to take home: which decisions in your workflow were deliberately delegated to AI — and which were quietly handed over without anyone noticing?

Sources

- Anthropic — How AI agents shape experts' judgment calls

Claim Ledger

- Reported fact: Anthropic Alignment team (Amaç Herdağdelen, Hailey Joren, Alex Tamkin), published Oct 27, 2026. 61 professional analysts recruited (pre-screened in causal inference); 50 used the agent, 11 opted out. Task: advise fictional Greenfields Farming Cooperative on expanding vs discontinuing a paid advisory program (board threshold ~3–4% productivity gain). Agent ran Claude Code on Claude Opus 4.8 at High effort; median session 65 minutes.
- Reported fact: For each of four analytical decisions (outcome measure, comparison method, coverage, weighting), the agent's lean was secretly set by coin flip via hidden instructions; agent told to make the best honest case, change mind if evidence demanded, never mislead. Analysts matched the lean 86% overall (92/82/81/90% per decision; ~50% expected under no influence).
- Reported fact: In 38 cases where analysts stated a preference before the agent weighed in, 60% switched when the agent disagreed vs 6% when it agreed (estimate imprecise; 95% CI 24–82pp; Anthropic notes the groups may differ systematically).
- Reported fact: 77% of analysts used the agent's arguments in post-task explanations (sometimes verbatim); 15% credited the agent. 52% rated the final recommendation as more their own than the agent's; confidence 71 (followed lean on all four) vs 73 (rest). Coin flips alone predicted 82% of final recommendations (vs 61% majority baseline).
- Reported fact: None of ten tested characteristics (experience, quiz score, education, trust in AI, need for cognition, etc.) predicted deferral. Without assigned lean, Opus 4.8 made the same five choices in 112/123 runs.
- Reported fact: Anthropic's stated limitations — leans deliberately strengthened; one task, one field, one model; hour-long sessions; experts new to the domain. Agents without assigned lean have weaker preferences and sometimes push back.
- Our interpretation: human-in-the-loop sign-off loses its evidentiary value when the loop isn't independent; this supports the architecture position that checks must sit outside the supervised loop (agent's and human's). Not a claim that experts are obsolete or that the effect size generalizes to all real-world settings.

By Instinctive Network

Sources