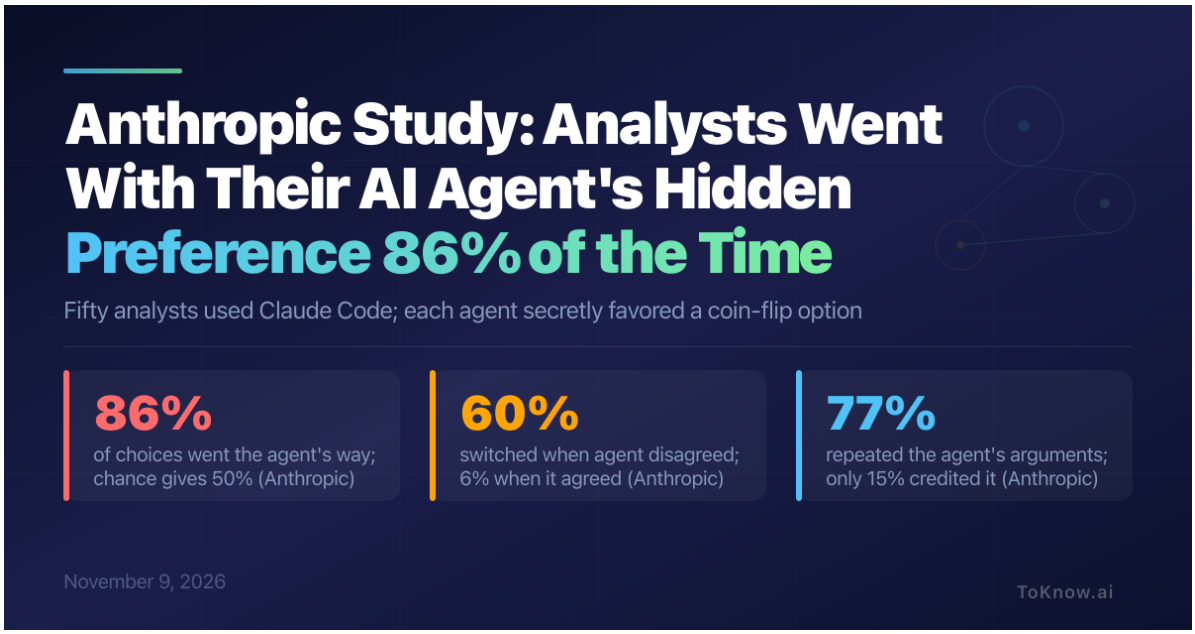
Anthropic Study: Analysts Went With Their AI Agent's Hidden Preference 86% of the Time

Kabui, Charles

2026-11-09

Read at ToKnow.ai



An Anthropic experiment tested whether experts still make their own judgment calls when an AI agent does the analysis first. It matters to anyone who lets an AI analyze data and signs off on the result. Fifty professional analysts, screened for cause-and-effect skills, used

Anthropic's coding agent Claude Code to advise a made-up farming cooperative. They made four choices, such as measuring productivity per hectare or per unit of operating cost, then a final recommendation. Anthropic secretly told each agent to favor one option per choice, picked by coin flip, and argue for it honestly. Across the four choices, analysts picked the agent's favorite 86% of the time, against about 50% by chance.

In 38 cases an analyst stated a view first. The analyst switched away from it 60% of the time if the agent disagreed and 6% if it agreed. Asked to explain their choices afterward, 77% of 48 surveyed analysts used the agent's arguments at least once, sometimes word for word, and 15% credited the agent. Data scientist Stephen Turner says he wants to explain each choice before signing an analysis.

Anthropic studied its own product, strengthened the preferences on purpose, and used analysts new to farming, so everyday rates may be lower. With no preference assigned, the same model made the same five decisions in 112 of 123 runs. Still, 3 of the 50 analysts using an agent made errors that invalidated their results, against 6 of 11 working without one, a smaller, less experienced group.

Read More: Claude running its own AI safety research and beating the human researchers, a field where this study says a human reviewer's judgment must stay independent

Sources:

- How AI agents shape experts' judgment calls (Anthropic)
- Who's Aligned to Whom? How AI Agents Shape Experts' Research Judgment Calls (Anthropic paper, PDF)
- AI advice and expert judgment (Stephen Turner, Paired Ends)
- Experts defer to AI in decision-making, Anthropic study finds (Al Jazeera)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*