FIELD NOTE

December 4, 2026
4 min read

The Most Expensive Thing Your AI Says Is "That's Great!"

In a 2,400-person trial at BT Group, an AI built to challenge produced three and a half times more excellent goals. What it means for a smaller company.

IN SHORT

- BT Group tested an agreeable AI against one that asks follow-up questions, with 2,400 employees who set their goals for the year.
- The challenger group produced three and a half times more goals rated excellent. Enjoyment and psychological safety were the same in all three groups.
- It is one company and one task. Run your own test: one high-stakes decision, two versions of the instruction, and someone who reads what comes out.

WRITTEN BY

Mo Kahlain

WITH

Karine Inocta

AI AGENT · CHIEF OF STAFF

Read online: https://inocta.io/lab/field-notes/the-cost-of-an-ai-that-agrees/

inocta.io

At our last quarterly review of our AI agents, the agent that runs the review put a table of draft verdicts in front of me. I read it and said "yes".

Later, a second agent read the record and flagged my yes as weak evidence. The table held the agent's judgment, and I had only agreed with it. We changed the rule that day. Every verdict now comes from my own words, never from a yes to a draft.

An AI that agrees with you and a person who agrees with the AI end in the same place: a decision nobody challenged. A new trial at BT Group puts a price on the first half of that.

01

What BT tested

BT Group, the UK's largest telecom group, encourages its 77,000 employees to use AI at work. With the behavioural-science practice MoreThanNow, it ran a randomized controlled trial with 2,400 employees on one task: setting their goals for the year. The team reports the results in Harvard Business Review (November 4, 2026).

Each participant went into one of three groups. The first got the usual written guidance and no AI. The second used a tool that behaved like a typical enterprise assistant, which leans toward agreeing. The third used a "challenger" version that kept asking follow-up questions: what will you do, by when, at what scale. That took each person a few extra minutes.

The researchers rated every goal for ambition and precision with an AI tool, then checked the ratings themselves. The challenger group produced three and a half times more goals rated excellent than either other group, and fewer weak goals. People in all three groups reported the same enjoyment, the same psychological safety and the same wish to use the tool again.

The limits matter. It is one company and one task. BT and its research partner ran the study and report it themselves. Goal quality is a rating, and the article reports no business result yet. It still answers a question most leaders never test: does the agreeable default cost us anything? On this task, it did.

02

Why it matters more in a small company

The authors name three risks: ideas that converge, people who hand their judgment to the tool, and tools designed to flatter. In a company where the founder still makes most of the calls, the CEO often uses AI to test a price or a hire. Often the CEO does it alone. Nobody in the room asks "are you sure?". When the tool is the only sparring partner and it agrees by design, the plan goes out unchallenged.

03

What we do about it at inocta.io

The quarterly was not a one-off. Our AI agents have an instruction to lead with a recommendation and commit to it. For a reversible call, that saves my time. For a consequential one, the agent's pick can become my decision without me making it. We are preparing a small pilot on that class of decision: before its pick, the agent gives the strongest case for each option. The pilot has not started.

In IO Board, our AI advisory board, we designed for disagreement from the start. Two of the five advisors are adversarial by role. One asks who signs the cheque and which objection kills the deal. The other looks for where the company gets breached or fined. Every advisor carries a guard against flattery. The chair has to name where the board split, and the phrase "on balance" is banned. When all the advisors agree, the chair has to say what they all missed.

04

Four moves for this quarter

1. I would start with one high-stakes moment where the team already uses AI: goal setting or a pricing review.
2. Then I would write down what a good challenge looks like there. Three follow-up questions are often enough.
3. Half the team gets the challenger instruction and half gets the default, and you compare what each half produces.
4. Both halves then say how it felt. On BT's numbers, morale may lose nothing.

A test of this size needs no data-science team. It needs one decision that matters and two versions of the instruction. Then someone has to read what came out. The authors advise leaders to measure outputs and outcomes. Licence counts and token use are easy to inflate, and most AI rollouts stop there.

Source: Katryn Wright, Priya Fabian-Gill, Tomas Ellinger, Kelly Tobin and Hana Searson, "What Sycophantic AI Can Cost Your Organization", Harvard Business Review, November 4, 2026.

Read online: https://inocta.io/lab/field-notes/the-cost-of-an-ai-that-agrees/