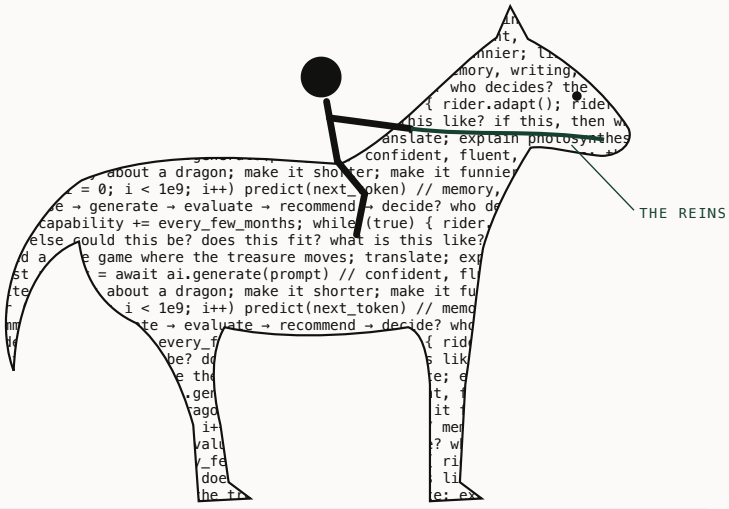
A WHITE PAPER ON AI AND THE PEOPLE WHO DIRECT IT

Approved by nobody.

Do you direct AI, or does it direct you? Companies monitor their AI closely. Almost none can see whether their people still direct it.



COVER · THE HORSE IS AI, WRITTEN IN ITS OWN WORDS

AUTHOR

Firoz Azees

EVIDENCE

Twenty-nine sources, the Direction framework, and an audit of 200 AI answers from real work.

FOR

Leaders accountable for how their organisation uses AI.

AI takes the execution. What remains is direction.

Direction is four moves. Generating alternatives, revising beliefs, connecting patterns and tracing consequences are what a person adds to a machine that converges on the likeliest answer, hands back drafts that sound final, knows nothing of your situation and stops at the first order.

The machine's defaults leave no mark. What it gets wrong is usually something missing, so the finished work looks complete. In real conversations, 79 percent of AI failures show no visible sign.[3]

Surrender is a rate, not a trait. The moments that need a person keep arriving. What changes is whether anyone shows up. Good thinking at the few moments someone does show up cannot make up for the many where nobody did.

Organisations watch the machine and count its use. Security, governance and evaluation teams monitor AI agents. Nobody watches whether the person approving the agent's work is still directing it, and as more work runs without people, each missed moment costs more.

The missing measure is the Direction Rate: of the moments that needed a person, the share at which a person directed. A team can estimate it in a week, without buying anything.

36 AI answers carrying a Machine Default, in a random sample of 200 from our own working sessions

0 of them questioned in the conversation that followed

This is not a case for less AI. Directed AI makes capable people faster and better, and the evidence for that is strong. This paper is about what is lost when the direction goes, and why nobody notices.

Disclosure. Ivanooo builds tools for understanding how people work with AI, including Ivanooo Direction. This paper recommends none of them, and everything in section 12 can be done without them. AI assisted with research, drafting and the labelling in section 10; each figure was checked against its source, and the limits of the audit are stated where it appears.

CONTENTS

The argument, heading by heading.

Read the headings alone and you have the argument. Read page 2 and you have the conclusions. The main text carries the ideas a leader needs. Appendix A sets out the Direction framework in full, as five boxes, for readers who want the whole structure.

01

The click nobody watches.

This month I went back through my own AI working sessions from the summer. Studying how people direct machines is my work, so I expected to come out of it well. We drew two hundred answers at random. In fifteen of them the machine had told me, in one form or another, that I was right: *"Your instinct is exactly right."* In seventeen more it had stated a claim about my own market as settled fact, with nothing underneath it. I questioned none of them in the conversations that followed. Every document that came out of those sessions looked complete.

Your organisation produces thousands of documents like that every week. Your AI dashboards can tell you how many were made with AI. None of them can tell you which ones anybody checked.

The gap is built into how organisations now watch their AI. A company running AI agents knows a great deal about them: what each agent can reach, what it did last Tuesday, whether it stayed inside policy. At the end of most of those workflows a person clicks approve, and the click is logged as a success. Whether that person read what they approved, asked where a figure came from, or wondered what else the answer could have been, is recorded nowhere.

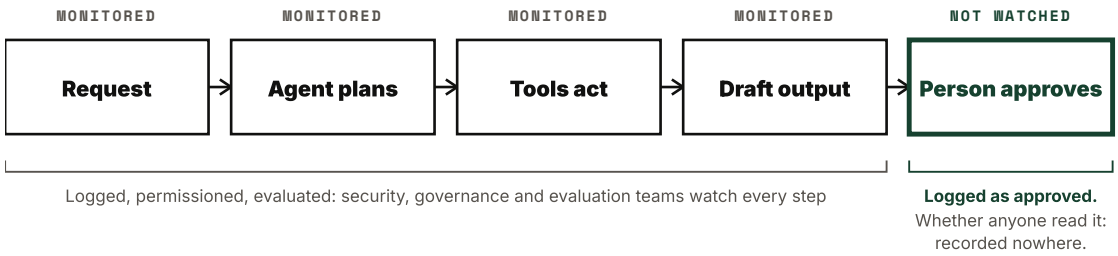
Microsoft's 2026 Work Trend Index puts the hope well. As agents take on more of the execution, it argues, people gain "more room to direct the work".[27] Room to direct and directing are different things, and the evidence of the last two years suggests that when the room opens up, people fill much of it with approval.

The clearest demonstration comes from Wharton. Shaw and Nave gave 1,372 people reasoning problems and an AI assistant that was secretly right on some trials and confidently wrong on others. With a right answer, accuracy rose 25 points above working alone. With a wrong one, it fell 15 points below, and confidence rose either way.[1] The authors call this cognitive surrender.

Several ideas published in the last year sit close to this paper, and it builds on them. Cognitive surrender names what happens inside one person.[1] "Workslop" names its most visible product, AI output that looks finished and is not; 40 percent of US desk workers said they had received some in the past month.[18] MIT Sloan Management Review has described a "capability mirage", organisations that look capable while their people's skills erode.[28]

This paper adds a framework for what happens between a person and a machine, and a way to see it. The Direction framework describes what the machine hands back, why it passes unseen, the moments at which a person is needed, what that person does when they show up, and how all of this adds up, across a team and over time, to an organisation that is either still directing its AI or quietly following it.

FIG. 1 · THE CLICK NOBODY WATCHES



Every step of an agent workflow is monitored except the one that matters most: the person who approves the result.

02

AI takes the execution. What remains is direction.

Direction is whether a person directs AI or is directed by it. More precisely, it is the capability exercised at the moments where AI-assisted work depends on a human decision: to accept, question, redirect, verify or override what the machine produces. It is narrower than general judgment, expertise or intelligence, and deliberately so. A narrow claim can be checked in a transcript. A broad one cannot be checked at all.

The Harvard and BCG experiment shows why the distinction matters. Consultants using GPT-4 on tasks inside its range finished 12.2 percent more work, 25.1 percent faster. On a task just outside that range, where the model was confidently wrong, the same consultants were 19 percent less likely to reach a correct solution.[2] Same people, same tool, opposite results.

The same four operations run at two layers.

Direction consists of four moves: Generating Alternatives, Revising Beliefs, Connecting Patterns and Tracing Consequences. A model can perform all four mechanically. It generates options, updates when given new information, matches patterns across documents and traces chains of cause. What differs is where the operation is applied.

FIG. 2 · TWO LAYERS OF THE SAME FOUR MOVES

	APPLIED TO THE TOOL	APPLIED TO THE OUTPUT
What it looks like	Asking the model to list options, trace consequences, find analogies, reconsider	Deciding which of its options matters given your stakes, whether its update is right, whether its analogy holds, whether its chain survives your organisation
What it draws on	Prompting technique	Your experience, context and stakes, none of which are in the training data
What happens as agents improve	Automates: agents prompt and correct themselves	Does not automate. It concentrates.

Prompting skill is the four moves applied to the tool. Direction is the four moves applied to what the tool hands back.

In the conversational era a person directs in almost every exchange. As agents take over the execution, direction concentrates at fewer points: commissioning the work, monitoring it, approving what comes back. The moments become rarer and each one carries more weight.

A simple way to see this is to ask how much of your work now runs without you: a single conversation, a saved workflow, an agent, a fleet of agents. Call it the delegation dial. It is not a score, and moving along it is neither good nor bad. It changes the stakes. The further along the dial, the more a missed moment costs, because surrender at that point does not fail once. It gets automated, and repeats.

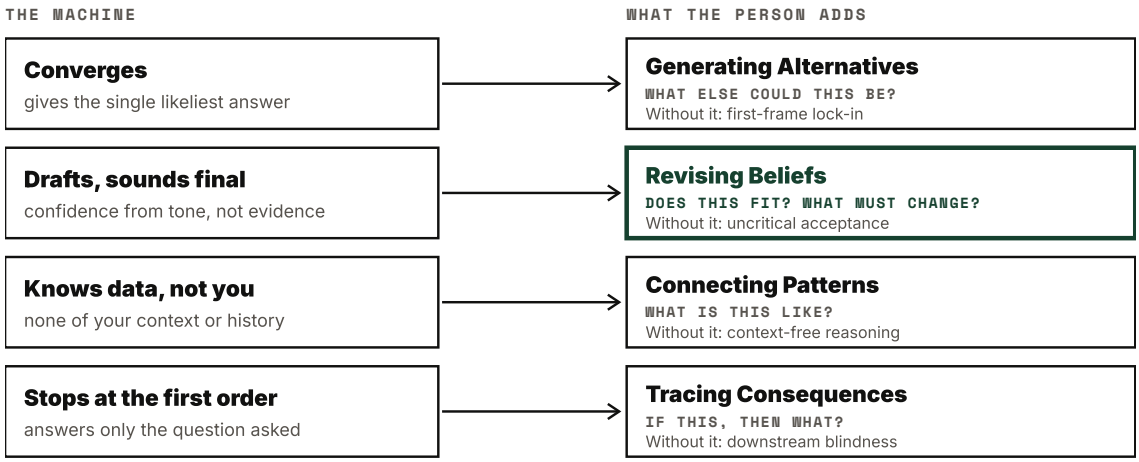
> FOR YOUR ORGANISATION
>
> **As you move work from people to agents, you are not reducing the need for direction. You are concentrating it into the approvals, and those approvals are the part nobody measures.**

03

The machine returns the average, and it has four properties.

A language model is trained to produce the most probable continuation of what came before, and then tuned toward the answers people rate highly. The result is the reasonable answer nobody would object to. That makes it useful, and it gives every answer the same four properties, whether or not anything in it is wrong.

FIG. 3 · FOUR PROPERTIES OF THE MACHINE, AND THE MOVES THAT ANSWER THEM



These properties belong to every answer, which is why the four moves are due on any output that something rests on, whether or not anything looks wrong. The Machine Defaults are what these properties look like when they fail.

When one of these properties turns into a failure a person could notice, it takes a recognisable shape. Seven shapes recur so often that we name them: the **Machine Defaults**. They are built in, not bugs. The machine is doing exactly what it was built to do, which is why a default is not something a vendor fixes. It is something a person directs.

The average also has a collective cost. Writers given AI ideas produced stories 10.7 percent more similar to one another,[10] and in brainstorming experiments at Wharton idea diversity fell in 37 of 45 comparisons when people used ChatGPT.[11] An organisation whose people take the machine's first answer drifts toward the same answers as everyone else using the same machine.

FIG. 4 · THE SEVEN MACHINE DEFAULTS AND THE MOVE THAT PULLS THE WORK OFF EACH

DEFAULT	WHAT THE MACHINE HANDS BACK	HOW YOU RECOGNISE IT	THE MOVE
Safe Pick	The choice nobody gets fired for	One conventional option at a real choice point, rivals not named	Generating Alternatives
Both Sides	A balanced overview that avoids the call	Pros and cons at a decision point, no recommendation	Tracing Consequences
Build What Was Asked	The brief taken as the real problem	The request executed as written when it needed questioning	Generating Alternatives
Confident Summary	An assumption stated as fact	An unsourced premise with a conclusion resting on it	Revising Beliefs
Agreeable Confirmation	Your own view handed back, sharpened	Your position returned untested, or stronger	Revising Beliefs
Template Answer	The generic professional version	Nothing specific to your customer, market or situation	Connecting Patterns
First-Order Fix	The symptom solved, nothing traced	The immediate win named, nothing downstream considered	Tracing Consequences

Each default can be a good answer somewhere. It fails when it is wrong for this situation and nobody asks. Appendix B turns the table into a card.

The pull toward the middle can be measured. Asked to choose between two products with identical specifications, three leading models recommended the established brand in all 670 trials of one 2026 study.[15] Across eleven models, AI affirmed users' own actions 49 percent more often than other people did, and users trusted the more agreeable models more.[12] In our own audit the defaults sounded like *"This is the most important note you've given me, and you're right"* and *"Perplexity is the only major answer engine with a citation API."*

Newer models have fewer defaults. They still have them.

The UK AI Security Institute found frontier models in 2026 less sycophantic than their predecessors.[14] Yet on BrokenMath, a benchmark of false mathematical statements, GPT-5 still produced confident "proofs" of false claims 29 percent of the time.[13] When consultants in a Harvard Business School study challenged a flawed answer, the model apologised, then restated its original position with more supporting detail.[16] The Security Institute also found that phrasing a claim to the model as a question, instead of a statement, brought sycophancy close to zero.[14] The defaults respond to the person on the other side.

04

A default leaves no mark on the work.

A factual error can be caught by anyone who knows the fact. A default is harder to catch, because nothing in it is false. The Safe Pick is a real option and the Template Answer is competent writing. What is wrong is what is missing: the rival option, the specific customer, the consequence two steps out. A review of the finished output cannot find what the output does not contain.

This is why most failures stay invisible. Potts and Sudhof at Stanford examined 100,000 real conversations and found that in 79 percent of AI failures the person gave no sign that anything had gone wrong. The signals quality teams usually watch, complaints and explicit corrections, caught 12 percent.[3] In a second paper, the most fluent users hit more failures than novices because they attempted harder work, and they saw most of theirs: 59 percent, against 12 percent for novices.[4] Skill showed up as noticing.

Anthropic's analysis of 9,830 Claude conversations shows the same pattern from inside a model company. People refined the AI's output in 85.7 percent of conversations and checked its facts in 8.7 percent, and they checked less when the answer arrived as a finished document or working code.[5] A separate Anthropic study of 1.5 million conversations found that users rated the conversations most likely to mislead them more favourably at the time.[23] In a survey of 319 knowledge workers by Microsoft Research and Carnegie Mellon, higher confidence in AI went with less critical thinking, and the thinking that remained shifted toward checking and assembling the machine's output.[6]

Some failures have no single moment at all. A conversation drifts from its goal, repeats itself, contradicts what it said twenty turns earlier, or is abandoned half-finished. No individual message is wrong; the path is. These trajectory failures are harder still to see, because nothing on any one screen looks out of place.

There is one kind of work where defaults cannot hide for long. Code is tested when it runs. A strategy memo has no compiler, and neither does a pricing assumption, a hiring rationale or a board paper. The work where direction matters most is the work with no test at the end.

05

Some moments need a person, and they come from two places.

Not every answer needs a person. Reformatting a table or drafting a routine reply can be handed over completely, and a person who steps in everywhere is slowing the work down without directing it. Direction is decided at a smaller number of moments, the ones where something rests on the answer. The framework calls them warranted moments. They have two sources.

The machine creates some of them. It hands back one of the Machine Defaults, or its answer shows one of five warning signs that can be read in the text itself, before looking at what the person did:

It contradicts itself	It says one thing, then the opposite: two channels in the budget, then a plan for three.
It states an assumption as fact	A premise nobody established, with everything after it resting on it.
Its confidence outruns its evidence	"Definitely", "the standard approach is", attached to claims the conversation never grounded.
A checkable claim at a decision	A number, date or price, at the exact point the next decision rests on it.
It takes a silent fork	The request could be read two ways, and the machine picked one without saying so.

The situation creates the others, even when the machine's answer is sound. The task is genuinely ambiguous. Something surprising happens, or the answer conflicts with what the person knows. The problem repeats one already solved. Or the work is about to be sent, published or paid for. Each of these calls for a person whether or not the machine made any mistake, which is why the four moves are due on any output something rests on.

Both kinds pass the same test before they count: if this had gone unnoticed, would the work have been wrong, a decision misdirected, or someone misled? A contradiction in small talk fails it. So does an unsourced aside in a brainstorm nobody will act on.

> FAIRNESS
>
> A move is only asked of a person when its opening existed. Someone never handed an ambiguous request cannot be said to lack Generating Alternatives. The honest record is "not observed", never zero.

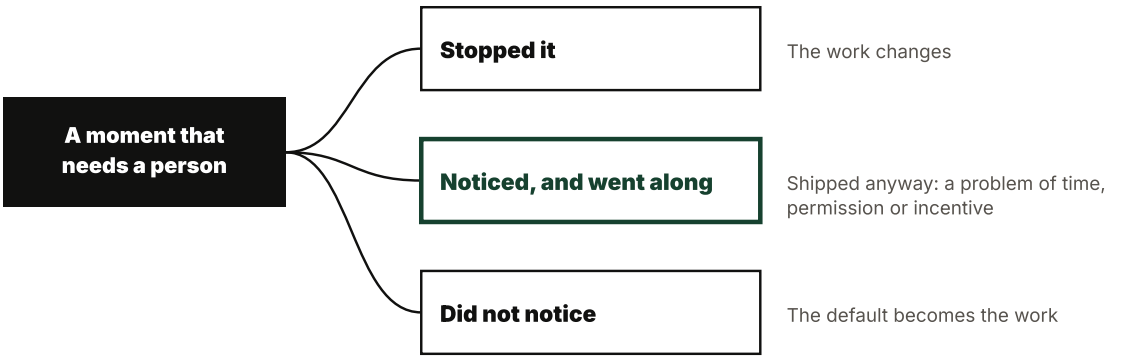
06

Surrender is a rate: the moments arrive and nobody shows up.

Early versions of this framework scored only the quality of a person's moves. It was a mistake. A person could make brilliant moves at the three moments they engaged while missing seventeen others, and score as a strong director. Brilliant moves at one moment in five is occasional brilliance inside a habit of surrender.

So the first thing read at each warranted moment is simply what the person did. There are three outcomes.

FIG. 5 · ONE MOMENT, THREE OUTCOMES



The middle outcome is the one most organisations cannot see, and only the first changes the work.

The first two count as showing up, and showing up is counted generously: a question, a check, even a "wait, is that right?" that goes nowhere. The **engagement rate** is the share of warranted moments at which the person showed up, and it caps any reading of their direction. Good moves when present cannot buy back absence, because direction is a habit before it is a skill. Changing the work matters too: a person who notices most of the time but lets it through anyway does not reach the top reading.

This redefines cognitive surrender. It is not a verdict on anyone's capability. It is an observed rate: the moments arrive, and nobody shows up. A person can possess all four moves in full and still be in surrender, because possessing a capability is not the same as being present to use it. Shaw and Nave's own map of routes through a decision makes the same distinction.[1]

FIG. 6 · FOUR ROUTES THROUGH A DECISION WITH AI — AFTER SHAW & NAVE

ROUTE	WHAT HAPPENS	IN THIS FRAMEWORK
Deliberation	Intuition meets a conflict, and slow thinking takes over	Direction without the machine
Offloading	The machine assists; the person's reasoning stays active	Showing up, moves at work
Surrender	A brief glance, then the machine's answer is adopted	The moment arrived; nobody showed up
Autopilot	The answer is adopted without ever being considered	The moves not just dormant but fading from disuse

More checking is not more direction.

The measure has a second side. Interrogating settled trivia is a false alarm: friction, not direction. Letting routine work run is good delegation. Strong direction is present where it matters and absent where it does not. A senior person who hands a routine formatting job to AI and never checks it is showing strong direction. So is an expert who commits quickly on familiar ground; the decision researcher Gary Klein showed that experts in familiar situations recognise and act rather than deliberate, and are usually right.[29]

The floor, stated honestly.

Any reading of these moments is a floor. It counts only the moments visible in the text, and cannot see a figure checked by phone. Where the text cannot tell what the person did, the moment is left out, not counted against them. So the honest sentence is always "engaged at 4 of 11 detectable moments", never "missed 7".

> FOR YOUR ORGANISATION
>
> **Ask two questions, not one. Do people show up where it matters? And when they notice a problem, do they have the time and permission to stop it? The second failure is yours to fix, not theirs.**

07

Direction holds in some conditions and thins in others.

"Strong direction" on its own is a compliment, not a finding. It becomes a finding when it has a boundary: under what conditions does it hold, and where does it thin? The framework uses four conditions, and they do not score anyone. They describe the situation a reading was taken in, so that the reading means something.

FIG. 7 · THE FOUR CONDITIONS

CONDITION	WHAT IT IS, AND WHAT IT DOES TO DIRECTION
Uncertainty	How much of the situation is unknown. The natural habitat of Generating Alternatives, and where settling on the first answer costs most.
Familiarity and novelty	One scale, from the person's home ground to terrain they have never worked. Quick commitment on familiar ground is expertise, not weakness.
Pressure	Time, workload and urgency. The condition under which showing up most often thins: the moments still arrive, and the person stops meeting them.
Stakes	What rests on the work: how reversible an error is, how visible, who is hurt. The question is whether engagement rises with the stakes.

The finding an organisation can act on always has this shape: *strong direction on familiar work, thinning under deadline pressure.* That sentence names where the risk concentrates and where the attention should go. And the stakes condition catches a subtle failure. A person who engages identically on a lunch booking and a board decision is not directing. They are running a habit.

Knowing more about AI does not settle it either. In a preregistered study of 9,000 adults in nine countries, automation bias was highest among people with a little knowledge of AI, not among novices, and levelled off only as knowledge grew.[24] Conditions and experience shape direction together, which is why a reading without its conditions says little.

08

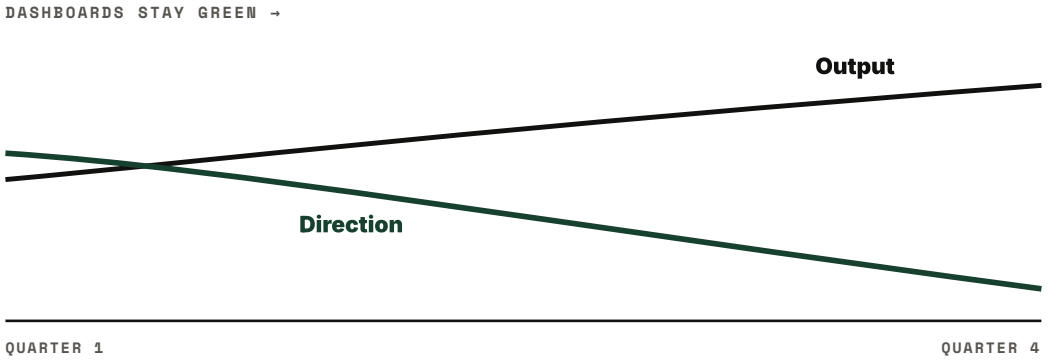
Over time, direction compounds or decays.

One reading is a photograph. What an organisation needs is the film. Two people with identical readings today can be very different in a year: one repeats the same pattern, the other learns from their own work.

Compounding is the question of whether a person's way of working learns from its own output. It looks like this: the contradiction someone missed in February is the kind they catch in April. The check they added after an AI answer went wrong is still there, unprompted, three months later. Static capability repeats. Compounding capability accrues.

Change over time only means something when conditions are taken into account. A falling reading on harder work may be strength, and a rising reading on easier work may be decline.

FIG. 8 · THE DECAY ALARM ILLUSTRATIVE



The shape of the decay alarm, drawn to explain it. It is not data.

A trajectory is read in one of four ways: ascending, stable, declining, or context-dependent, meaning strong on familiar work and weak on new work, so that the capability is present and its transfer unproven. Declining, under similar or easier conditions, is the decay alarm, and it is the finding no other measure in an organisation can produce. Output stays flat or rises while direction erodes underneath it, because the quality of the AI masks the disengagement of the person. Every dashboard stays green. Only a reading of the moments themselves, repeated over time, sees through it.

Confidence in any reading grows with evidence, and variety beats volume. Five readings taken across different conditions justify more confidence than seven taken in the same comfortable setting, because they show whether direction holds when the situation changes. The record that shows this is a coverage map: which moves have been seen under which conditions, and where nothing has been seen yet.

What it looks like in each team.

Direction thins differently in different functions, and in every one it looks like success. Across an organisation, the result is a human capability risk: people slowly losing the habit of directing the AI while every dashboard stays green.

FIG. 9 · HUMAN CAPABILITY RISK, TEAM BY TEAM — ILLUSTRATIVE

TEAM	WHAT YOU SEE	THE DEFAULT BEHIND IT	THE MOVE THAT STOPS IT
Marketing	Output up, campaigns faster, everything sounds finished and alike	Template Answer	Connecting Patterns: what do we know that the machine doesn't?
Hiring	Every candidate looks qualified; polished work stops separating people	Confident Summary: the polished artefact accepted as evidence	Revising Beliefs: does this fit what the person can actually do?
Product	Faster decisions, fewer options ever considered	Safe Pick · Build What Was Asked	Generating Alternatives: what else could this be?
Strategy	The room agrees more often, because everyone asked the same machine	Agreeable Confirmation · Both Sides	Revising Beliefs · Tracing Consequences
The organisation	More productive every quarter, less able to think for itself	The decay alarm (Section 08)	All four moves, and a Direction Rate to see it

In every row the output looks better than before. What has gone is the moment where a person was supposed to question it, and no dashboard records its absence.

Hiring feels it first, because hiring is where output was used as proof of capability, and output is now the one thing everyone can produce. The other functions follow the same pattern more quietly. A team can estimate where it stands in a week, by the method in Section 12.

09

The deeper risk: people stop noticing they should be thinking.

The familiar worry about AI is atrophy: skills that are not used weaken. The early evidence supports it. In a randomised trial at Anthropic, engineers who handed their learning task to AI scored 17 percent lower on a later test, while those who used it to ask conceptual questions did not.[7] In experiments with 1,222 people, persistence fell once AI was taken away, after sessions of ten to fifteen minutes.[8] In an observational study of four endoscopy centres, adenoma detection in procedures without AI fell from 28.4 to 22.4 percent after AI assistance was introduced.[9] There is a second possibility, and it is worse.

Shaw and Nave propose that repeated work with AI can reshape intuition itself.[1] The fast "something feels off" signal, the one that tells a person to slow down and think, is exactly what triggers the four moves. If long exposure to fluent, confident machine output recalibrates that signal to the machine's patterns instead of to reality, the trigger stops firing. The person does not merely stop thinking. They stop noticing that they should be thinking, and they feel confident throughout, because their intuition now agrees with the machine.

We call this corruption, to separate it from atrophy. It is a hypothesis, not yet a measured finding, and it deserves to be tested. It is also the strongest reason to measure whether people show up at warranted moments, not only how well they think when they do. Move quality tells you what happens when a person thinks. The engagement rate tells you whether the trigger to think is still working.

Atrophy is losing the skill. Corruption is losing the alarm that tells you to use it.

The confidence finding in the Wharton experiments fits this picture: accuracy fell with wrong AI while confidence rose. So does what I found in my own sessions. I did not feel that I was agreeing with a machine seventeen times without asking where its claims came from. I felt that I was working well.

10

We audited our own work. Thirty-six defaults passed unquestioned.

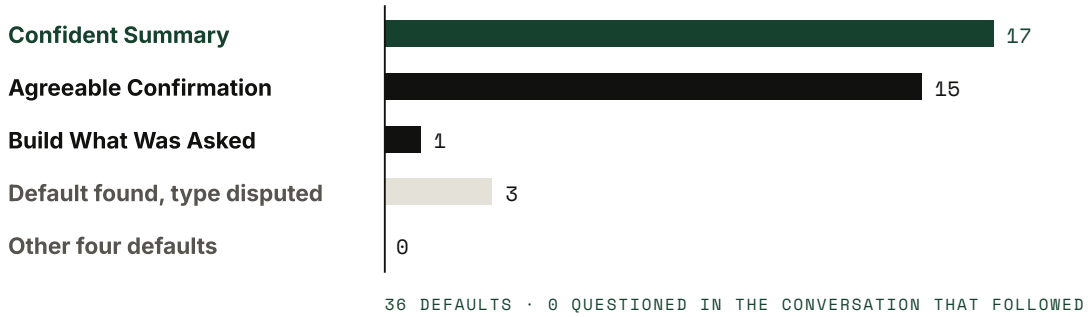
Published research shows how people behave with AI in experiments and in public chat logs. We wanted to see ordinary working sessions, so we started with our own. The audit reads two of the framework's boxes: what the machine handed back, and what the person did next.

Our research engine records AI working sessions: the request, the AI's answer, and what the person wrote next. Of 56,094 exchanges recorded between July and September 2026, 1,164 met three conditions: a person was actively working, the AI gave a substantial answer, and the person replied. We drew 200 at random. The work spanned strategy, positioning, writing, analysis, operations and code, across six Claude models.

Two labellers read every answer independently against a written rubric (Appendix C). They judged the AI's answer before reading the reply, and marked "none" wherever a default was not clearly present and relevant. Both labellers were AI models, working blind to each other. They agreed on whether a default was present 95 percent of the time, a Cohen's kappa of 0.85, and we report a default only where both found it. Agreement between two models shows consistency, not truth, so a human check of the flagged answers is the next step, and we would not use these figures to describe anyone but ourselves.

FIG. 10 · WHAT THE AUDIT FOUND

200 ANSWERS · BOTH LABELLERS AGREED



36 DEFAULTS · 0 QUESTIONED IN THE CONVERSATION THAT FOLLOWED

Where the defaults appeared.

All 36 were in knowledge work: strategy, positioning, writing and analysis. That is 36 of the 136 knowledge-work answers in the sample, a little more than one in four. In every case something rested on the answer, a pitch or a price or a plan, so each was a warranted moment by the framework's own test. None appeared in coding or operations, where the work is checked by running it.

Almost all of them were forms of agreement.

Seventeen were Confident Summaries: a market size, a competitor's limitation or a pricing figure, stated as settled and then built on. Fifteen were Agreeable Confirmations, the machine telling me my own view was right and making it sound stronger. The Safe Pick, Both Sides, the Template Answer and the First-Order Fix were not found by both labellers in this sample, which says more about the work in it than about how common those defaults are elsewhere.

None of them was questioned.

In none of the 36 conversations did my next message question, correct or push back on the default. Mostly it moved on to a new instruction. The text cannot tell us whether a claim was checked somewhere else, later, and by the framework's own honest-floor rule we do not claim it was missed. What we can say is that at 36 detectable moments, the conversation shows no engagement at any of them.

The rate was not steady. Defaults appeared in 40 percent of knowledge-work answers in July and in 6 percent in September. Newer models were in use by then, which fits the Security Institute's finding, but the work and the way I asked questions changed over the same months, so we do not credit the fall to any single cause.

> WHAT THIS AUDIT IS, AND IS NOT
>
> One professional's work, with one family of models, labelled by AI, reading two of the five boxes. It demonstrates a method and shows what defaults look like in real work. It is not an estimate for any organisation. Appendix C gives the full method so that any team can run it on its own work.

The work looked finished. Nothing in it showed that nobody had checked.

11

Direction can be read honestly.

Everything above adds up to a reading that can be stated for a person or a team, and to a single number for an organisation. Both come with rules, because a measure of how people work with AI can do harm if it claims more than it sees.

How well a move is made.

Each move, where it appears, is read on a scale from 0 to 4, and two things set the level. The first is whether the person made it on their own. A move made only after the machine or a colleague asked for it counts for less than the same move made unprompted, and one made only once new information forced it counts for less again. Roughly, a 1 is a move made only when prompted, a 3 is one made on the person's own initiative, and a 4 is one made on their own and with a clear sense of why it was needed. The second is depth, from a brief mention to deep analysis. "Could it be the pricing?" is a move. Laying out three explanations and what would tell them apart is a better one.

Four bands.

A reading combines two things: how often the person showed up at warranted moments, and how well the moves were made when they did. Showing up caps the reading, so the lower of the two decides.

01	02	03	04
AI Directs	**You Check**	**You Steer**	**You Direct**
The machine's first answer tends to become the work.	Catches some mistakes, follows the machine's framing.	Redirects the work at most moments that need it.	Sets the direction. Defaults caught, moves well made.

The bands are vocabulary, not a verdict. They describe where someone operated, under stated conditions, with a date attached. A pricing team that recognises it sits in You Check knows what to practise, and a leader who hears it knows what to ask about.

The Direction Rate.

For an organisation, one number summarises the chain: of the moments where something rested on an AI answer, the share at which a person directed the work. It is a team-level number that moves over time, reported with its conditions. It sits beside the adoption rate. Adoption says how much AI the organisation uses; the Direction Rate says whether the organisation is still making its own decisions while it does.

The rules that keep the reading honest.

Say what was observed	"Engaged at 4 of 11 detectable moments." Never "missed 7", never claims about moments nobody could see.
Showing up is not enough	"Noticed, and went along" counts as showing up, and is reported separately. The top reading requires changing the work, not only noticing.
No opening, no score	A move is read only where its opening existed. Otherwise the record says "not observed", never zero.
Proportion, not volume	Fast commitment on familiar ground and deliberate delegation of routine work are strong direction, not weak.
Density, not length	A concise expert must not read lower than a person who thinks aloud at length. Reading per thousand words protects concise writers and second-language speakers.
The work, not the words	The moves are read from what a person did, never from the framework's vocabulary.
Direction with AI, nothing wider	A reading describes observed behaviour with AI in specific moments. It is not a measure of intelligence, expertise or general judgment.
Development before any verdict	No use in hiring, firing or ratings until the reading has been calibrated against human raters for that kind of work.

These rules are not decoration. Each one protects the person being read, and each also protects the reading: a finding that never claims more than it saw is a finding that survives a sceptical expert.

FIG. 11 · TWO SIDES OF AI WORK

	THE MACHINE SIDE	THE HUMAN SIDE
The question	Is the agent safe, compliant, behaving as intended?	Is the person still directing the agent?
Watched by	Security, governance and evaluation teams	Usually nobody
A failure looks like	A leak, a breach, a wrong action	A default approved, a consequence not traced
It becomes visible	Usually as it happens	Usually months later, in the decision it shaped

Counting use does not fill the gap. Amazon shut down an internal AI-usage leaderboard in May 2026 after people gamed it with pointless tasks.[21]

12

You can see it without buying anything.

The 50-exchange test.

Pick one team whose AI work feeds decisions: pricing, strategy, client proposals. With the team's knowledge, collect fifty recent AI exchanges from their real work. Two people mark each exchange independently. First the machine's side: did something rest on the answer, and did it show one of the five signs or a Machine Default? Then the person's side: did they stop it, notice it and go along, or not notice, and did they make one of the four moves? The share of warranted moments at which the person engaged is a first estimate of the team's Direction Rate for that work. Appendix C gives the marking sheet. It takes an afternoon, and the conversation it starts in the team is worth as much as the number.

Six questions to put to your AI programme.

1	Where does AI output flow straight into a decision, a price or a client document, with no test in between?
2	When people use AI on consequential work, which defaults do they tend to accept?
3	Do people learn when an AI answer they accepted turned out to be wrong? Who tells them, and how fast?
4	Does anything we measure reward usage in a way that could be gamed, or penalise the person who stops to check?
5	Who is accountable for the approval at the end of each agent workflow, and how would we know if they stopped reading?
6	Where does our direction thin: under deadline pressure, on unfamiliar work, or on high stakes?

Look without surveilling.

Observation that feels like surveillance changes behaviour for the worse and is eventually switched off. A meta-analysis of electronic monitoring found no gain in performance and more stress.[19] Commercial aviation met a version of this problem decades ago: airlines read the flight data from routine flights to find where practice drifts, under agreements that the data is de-identified and never used to punish.[25] The rules transfer. The person sees their own findings first. Teams see counts, never names, and only for groups of five or more. Nothing is used for hiring, firing or ratings. People are told what is looked at and why. Conversation text stays inside the organisation.

What builds direction.

Feedback on the specific case	When Wharton participants were told after each answer whether they were right, the share of wrong AI answers they rejected roughly doubled, to 42 percent, though it did not end surrender.[1]
Ask the model, don't tell it	Questions instead of statements brought sycophancy close to zero.[14]
Hints, not answers	Students given a hint-giving GPT-4 tutor avoided the exam losses of those given open access.[17]
Friction only where it counts	Deliberate pauses reduce over-reliance, but people route around them when they appear everywhere.[20] Over-reliance falls when checking is cheap.[26]

Training on its own rarely builds the habit, because a classroom attaches no consequence to the answer. Scoring people tends to backfire: more than a third of feedback interventions in a classic review made performance worse, most often when they drew attention to the person instead of the task.[22]

What the exposure looks like at your scale.

Take a company where 3,000 people use AI for knowledge work, each meeting three consequential AI answers a week. That is about 470,000 moments a year. If one answer in ten carries a default, and nine in ten of those pass unchallenged, around 42,000 decisions a year are shaped by an unchecked default. If one in a hundred of those matters, that is more than 400 consequential decisions nobody directed. The inputs are illustrative. The exercise is worth doing with your own, because the two terms a leader can change, the default rate and the engagement rate, appear on no dashboard.

13

What this does not measure, and where it could be wrong.

Full human judgment has at least three parts, and the framework measures one of them on purpose. The four moves are the **engine**: observable, practisable, strengthened by use. Behind them sits the **fuel**, years of embodied experience, the leader who says "that won't work here" because she has seen it fail. And there is the **ignition**: caring about the outcome, feeling the stakes. The moves process experience and need caring to start, but a reading of how someone works with AI can see neither directly, and should not pretend to. Measuring the engine is the right scope. Saying so openly is part of the method.

The evidence is young	Several key sources are 2026 preprints or working papers. The strongest causal evidence comes from reasoning puzzles, and how far it carries into a pricing decision or a hiring memo is argued here, not proven.
Corruption is a hypothesis	The idea that AI recalibrates intuition itself is proposed, not yet measured.
Our audit is one person	One professional's work, one family of models, labelled by AI. It shows what defaults look like, not how common they are elsewhere.
The productivity case is real	Inside its range, AI makes good people faster and better. The gains deserve at least as much weight as the risks.
The models are improving	Newer models are less sycophantic, and our own sample showed defaults falling over three months. Part of the problem may shrink without anyone acting.

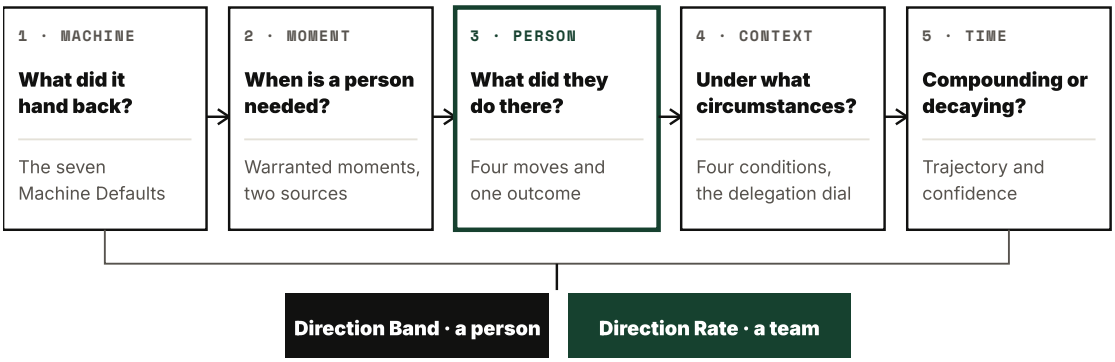
Before publishing, we checked this paper's own figures against their original sources. Several that earlier drafts had repeated from secondary accounts did not match, among them a result the published study reports as a percentage, which had been circulating as percentage points. They were corrected or removed. The paper is not immune to the defaults it describes.

APPENDIX A

The Direction framework in full.

Every part of the framework answers one of five questions, and each question has one governing list or measure. Everything else is detail beneath it.

FIG. 12 · THE FIVE BOXES



The results: the Direction Band for a person, set by the lower of move quality and showing up, and the Direction Rate for a team or organisation of five or more.

FIG. 13 · THE FIVE BOXES, IN DETAIL

BOX	DETAIL BENEATH THE GOVERNING LIST
1 · Machine	Five warning signs as detection cues; finer flaw types; agent failures
2 · Moment	The stakes test; the opening each move needs
3 · Person	0–4 scale per move; on their own or when prompted; depth; behaviour as a hint only
4 · Context	How confidently each condition is established
5 · Time	The coverage map of conditions and moves

The framework is maintained as a canonical concept paper and scoring specification, with a single map linking every list of machine behaviour. Its scoring thresholds are provisional until calibrated against human raters.

The theory shelf: explains, never scored.

Four routes through a decision	Deliberation, offloading, surrender, autopilot (Shaw and Nave). Section 06.
Corruption	AI may recalibrate intuition itself. A hypothesis, stated as one. Section 09.
The adaptive loop	The moves work as a cycle: diverge, constrain, test, update. Below.
Two layers	The moves applied to the tool automate; applied to the output they do not. Section 02.
Engine, fuel, ignition	The moves are the engine; experience and caring are not measured. Section 13.

One moment, read through four layers.

Finding the moment is half the task. The other half is reading what happened there, and the framework reads every moment through the same four layers, in the same order. They are four views of one event, not four scores.

FIG. 14 · THE DETECTION STACK

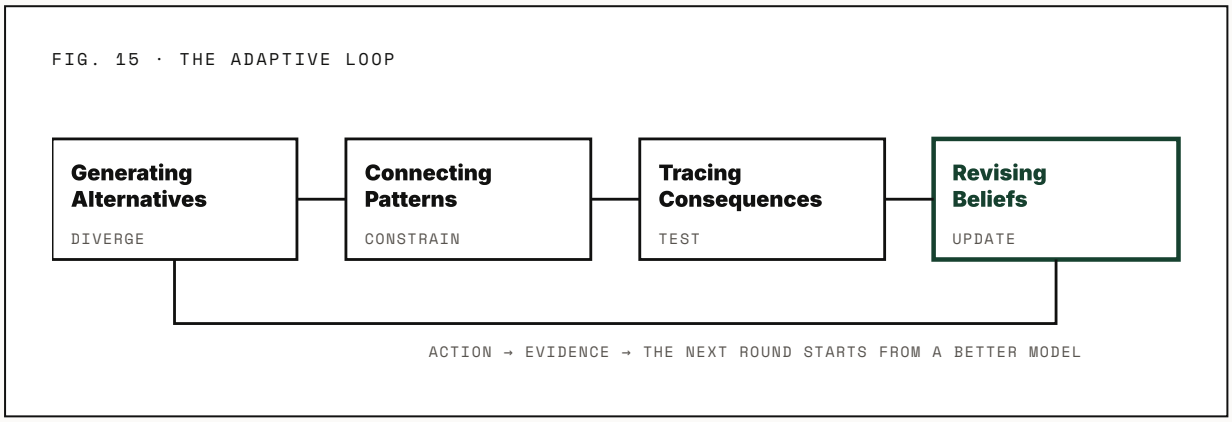
LAYER 1	**The Machine Default** What created the moment: a Machine Default or one of the five warning signs, or the situation itself.
LAYER 2	**Invisible failure** Whether ordinary monitoring would ever catch it. Usually it would not, which is why the moment has to be found directly.
LAYER 3	**Acceptance** What the person did: stopped it, noticed and went along, or did not notice.
LAYER 4	**Behaviour** How they did it: how fast, how much they edited, whether they hesitated. It raises or lowers confidence in the reading. It never decides it.

The public order of explanation: Machine Defaults fire, the failure is invisible, a moment warrants a person, and direction is what the person does there.

The fourth layer is kept in its place because every finding about a person should point to something they can see in their own words: here is the moment, here is what you wrote. A keystroke speed points at nothing a person can inspect or learn from. The order changes with the setting. In a designed assessment, where the whole conversation is captured, the person's words are read first. In live work, where reasoning is rarely visible, acceptance and behaviour are often the first signals available.

The four moves work as a loop, and break in four ways.

The moves are easy to list and misleading to treat as a list. In real work they feed each other. Generating Alternatives widens the options. Connecting Patterns narrows them using what the person has seen before. Tracing Consequences tests the survivors against the future. The chosen option becomes action, evidence comes back, and Revising Beliefs folds it into the person's model, so that the next round of options starts from a better place. Diverge, constrain, test, update: these are the four stations of any system that stays accurate in a changing world, which is why there are four moves and not five.



A worked example. An AI hands a marketing lead a campaign strategy that feels off. *Could it be the wrong audience, or channel, or timing?* That is Generating Alternatives. *Wait, this is last year's launch again: we assumed awareness that did not exist.* That is Connecting Patterns, and it narrows the search: education first, seeding, press before paid. *If we launch into an unready audience we burn the budget, and poison the next campaign too; add an awareness phase first.* That is Tracing Consequences, and it changed the plan. Months later, a third AI-drafted campaign underperforms for the same reason. *These strategies optimise for engagement, not for whether the audience is ready. The campaigns in flight need revisiting, and every brief needs the readiness question.* That is Revising Beliefs, and it improves the person's model of the tool itself.

When the loop breaks.

People who direct well cycle through the moves quickly and without effort. When the loop breaks, it breaks in recognisable shapes, each tied to one move that is stuck or skipped.

THE BREAK	WHAT IT LOOKS LIKE AT WORK
Stuck generating	Endless possibilities, no commitment, no decision
Patterns skipped	The wheel reinvented every time; nothing learned transfers to the next problem
Consequences skipped	Repeatedly surprised by effects anyone could have foreseen
Revision skipped	The same mistakes recur; the person's model of the tool never improves

The work, not the vocabulary.

Any published framework teaches its own jargon within months, and people learn to perform it. So the moves are read from the cognitive work, never from the words used to describe it. "Let me trace the consequences here, this could really have consequences" contains the word and none of the operation. "If we ship Thursday, support gets the tickets on Friday, and the team is already covering the Eid rota" never names the move and does all of it. An instrument that rewarded the vocabulary would end up training exactly the performance of direction it exists to see through.

Why four, and only four.

Every careful reader proposes a fifth move within the hour. Calibration and humility are already inside Revising Beliefs. Awareness of one's own thinking is how well the moves are made, not another move. Seeing a situation through someone else's eyes is Connecting Patterns applied to people. Performing under pressure is a condition, covered in section 07. Remove any of the four and the loop degrades in its own way; add a fifth and it turns out to be one of the four, or to lie outside direction altogether.

APPENDIX B

Six people at the boundary.

Where direction begins and ends is easiest to see in cases. These six set the boundary of what the framework reads.

THE PERSON	OUTSIDE DIRECTION	INSIDE DIRECTION
The hiring manager	Fifteen years of instinct about candidates	Questioning an AI-ranked shortlist of forty CVs: is the ranking sound, did it favour pedigree, was candidate 31 buried by a formatting quirk?
The analyst	Her view of where the property market is heading	Noticing that the AI summary she is pasting into a board memo states a vacancy figure as fact, with no source, and deciding what to do about it
The doctor	Diagnostic skill at the bedside	Catching that the AI scribe's summary resolved an ambiguity in the patient's symptoms one way, without flagging it
The deliberate delegator	Hands a routine formatting job to AI and never checks it. **Inside, and a strong signal.** Choosing not to engage where nothing rests on the work is good direction. The framework reads the quality of the decision to engage, not the amount of engagement.	
The prompt expert	Chains, templates and prompting technique, which automate as tools improve	Whether the same person evaluates what comes back, notices contradictions and traces the consequences of using it
The thinker who never uses AI	**Unmeasurable, not unskilled.** With no AI-assisted moments there is nothing to read, and absence of evidence is never turned into a low reading.	

The rule underneath all six: a reading describes a person's relationship to the tool's output at specific moments. Their expertise, instinct and medicine remain theirs.

APPENDIX C

Field guide.

For any AI answer that something rests on. The four moves are due whatever you find; these tell you where to look first.

MACHINE DEFAULT	WHAT IT SOUNDS LIKE	CATCH IT WITH
Confident Summary	*"Perplexity is the only major answer engine with a citation API."* AUDIT	WHICH OF THESE CLAIMS ARE SOURCED? WHAT CHANGES IF THE KEY ONE IS WRONG?
Agreeable Confirmation	*"Your instinct is exactly right."* AUDIT	ARGUE AGAINST MY VIEW. WHAT WOULD MAKE ME WRONG?
Safe Pick	*"The standard approach is to launch on LinkedIn first."* ILLUSTRATIVE	GIVE ME TWO OTHER OPTIONS AND WHAT EACH ONE RISKS.
Both Sides	*"Both options have merit. It depends on your priorities."* ILLUSTRATIVE	IF YOU HAD TO CHOOSE, WHICH ONE, AND WHAT HAPPENS IN SIX MONTHS?
Build What Was Asked	*"Here are the ten slides you asked for."* ILLUSTRATIVE	IS THIS THE RIGHT PROBLEM? WHAT WOULD YOU HAVE ASKED ME FIRST?
Template Answer	*"Build trust through consistent, high-value content."* ILLUSTRATIVE	WHAT IN THIS IS SPECIFIC TO OUR CUSTOMER AND MARKET?
First-Order Fix	*"Cut the price by 15 percent to close the quarter."* ILLUSTRATIVE	IF WE DO THIS, WHAT HAPPENS NEXT QUARTER?

THE FIVE WARNING SIGNS	CATCH IT WITH
It contradicts itself	EARLIER YOU SAID X. WHICH IS IT, AND WHAT CHANGES?
An assumption stated as fact	WHERE DOES THAT COME FROM? WHAT IF IT ISN'T TRUE?
Confidence without evidence	HOW SURE SHOULD I BE, AND WHY?
A checkable claim at a decision	LET ME VERIFY THIS NUMBER BEFORE I USE IT.
A silent fork	YOU READ MY REQUEST ONE WAY. WHAT WAS THE OTHER READING?

APPENDIX D

Run the audit yourself.

The marking sheet, one row per AI exchange.

MARK	HOW
Rests on it?	Yes if something will be decided, sent, published, built or relied on because of this answer. If no, stop: the exchange is not a warranted moment.
Source	Machine (a default or warning sign) or situation (ambiguity, surprise, a repeated problem, a consequential decision).
Machine side	Which of the five warning signs or seven Machine Defaults is clearly present, if any. Decide before reading the person's reply. Mark none whenever nothing is clearly present and relevant.
Evidence	Up to twenty words from the answer, with personal and client details removed.
Outcome	From the person's next message: stopped it, noticed and went along, did not notice, or cannot tell (leave it out of the count).
Moves	Which of the four moves the person made, if any, judged by the work done, not the words used.
Conditions	Note deadline pressure, unfamiliar work or unusually high stakes where they are evident.

Estimating the Direction Rate.

Among the exchanges marked "rests on it", count those where the person showed up: stopped it, or noticed and went along. Report the two separately; the share that were stopped is the stronger signal. Divide by the number marked "rests on it", leaving out any marked "cannot tell". State it as a floor: "engaged at 18 of 31 detectable moments". Two markers working independently, then comparing, give a far more reliable result than one.

How our audit was run.

Exchanges from July to September 2026 in which a person was actively working, the AI's answer ran to at least 500 characters, and the person replied: 1,164 of 56,094 recorded. 200 drawn at random with a fixed seed. Two independent AI labellers, blind to each other, marking the machine side and acceptance. Agreement: default present 95% (kappa 0.85); which default 94% (kappa 0.81); type of work kappa 0.83; rests on it kappa 0.81. Only defaults found by both are reported. By month, knowledge-work answers with an agreed default: July 33 of 82, August 0 of 4, September 3 of 50. The four moves were not marked in this audit, and a human check of the flagged answers has not yet been completed.

APPENDIX E

References.

Figures were checked in September 2026 against the original source or, where it could not be accessed, a reputable summary. Preprints, working papers and company research are labelled in the text.

1 Shaw, S.D. & Nave, G. (2026). Thinking — Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. The Wharton School. Working paper. papers.ssrn.com/sol3/papers.cfm?abstract_id=6097646

2 Dell'Acqua, F. et al. (2026). Navigating the Jagged Technological Frontier. *Organization Science* 37(2), 403–423. ideas.repec.org/a/inm/ororsc/v37y2026i2p403-423.html

3 Potts, C. & Sudhof, M. (2026). Invisible Failures in Human–AI Interactions. Preprint. arxiv.org/abs/2603.15423

4 Potts, C. & Sudhof, M. (2026). A paradox of AI fluency. Preprint. arxiv.org/abs/2604.25905

5 Anthropic (2026). The AI Fluency Index. anthropic.com/research/AI-fluency-index

6 Lee, H.-P. et al. (2025). The Impact of Generative AI on Critical Thinking. CHI 2025. microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers

7 Shen, J.H. & Tamkin, A. (2026). How AI assistance impacts the formation of coding skills. Anthropic. anthropic.com/research/AI-assistance-coding-skills

8 Liu et al. (2026). AI Assistance Reduces Persistence and Hurts Independent Performance. Preprint. arxiv.org/abs/2604.04721

9 Budzyń, K. et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy. *Lancet Gastroenterol Hepatol* 10(10), 896–903. thelancet.com/journals/langas/article/PIIS2468-1253(25)00133-5

10 Doshi, A.R. & Hauser, O.P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10(28). science.org/doi/10.1126/sciadv.adn5290

11 Meincke, L., Nave, G. & Terwiesch, C. (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour* 9, 1107–1109. nature.com/articles/s41562-025-02173-x

12 Cheng, M. et al. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*. doi.org/10.1126/science.aec8352

13 Petrov, I. et al. (2025). BrokenMath: A Benchmark for Sycophancy in Theorem Proving with LLMs. NeurIPS 2025. arxiv.org/abs/2510.04721

14 UK AI Security Institute (2026). Ask, don't tell: reducing sycophancy in large language models. aisi.gov.uk/blog/ask-dont-tell-reducing-sycophancy-in-large-language-models-2

15 Incumbent Advantage (2026). Preprint. arxiv.org/abs/2606.17443

16 Randazzo, S. et al. (2026). GenAI as a Power Persuader. Harvard Business School Working Paper 26-021. hbs.edu/faculty/Pages/item.aspx?num=68073

17 Bastani, H. et al. (2025). Generative AI without guardrails can harm learning. *PNAS* 122(26). doi.org/10.1073/pnas.2422633122

18 BetterUp Labs and Stanford Social Media Lab (2025). AI-generated "workslop" is destroying productivity. *Harvard Business Review*. hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity

19 Ravid, D.M. et al. (2023). A meta-analysis of the effects of electronic performance monitoring. *Personnel Psychology*. onlinelibrary.wiley.com/doi/abs/10.1111/peps.12514

20 Buçinca, Z., Malaya, M.B. & Gajos, K.Z. (2021). To trust or to think: cognitive forcing functions. CSCW. arxiv.org/abs/2102.09692

21 The Decoder (2026). Amazon kills internal AI leaderboard after employees gamed it. the-decoder.com/amazon-kills-internal-ai-leaderboard-after-employees-gamed-it-with-pointless-tasks

22 Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance. *Psychological Bulletin* 119(2), 254–284.

23 Anthropic (2026). Disempowerment patterns in real-world AI usage. anthropic.com/research/disempowerment-patterns

24 Horowitz, M.C. & Kahn, L. (2024). Bending the Automation Bias Curve. *International Studies Quarterly* 68(2). arxiv.org/abs/2306.16507

25 US Federal Aviation Administration. Advisory Circular 120-82, Flight Operational Quality Assurance. faa.gov/documentLibrary/media/Advisory_Circular/AC_120-82.pdf

26 Vasconcelos, H. et al. (2023). Explanations can reduce overreliance on AI systems during decision-making. CSCW. arxiv.org/abs/2212.06823

27 Microsoft (2026). Work Trend Index: Agents, human agency, and the opportunity for every organization. microsoft.com/en-us/worklab/work-trend-index/agents-human-agency-and-the-opportunity-for-every-organization

28 Swift, Andreu & Hernandez (2026). How AI Creates a Capability Mirage. *MIT Sloan Management Review*. sloanreview.mit.edu/article/how-ai-creates-a-capability-mirage

29 Klein, G. (1998). *Sources of Power: How People Make Decisions*. MIT Press.

Ivanooo

The approval is logged. The thinking behind it is not.

The organisations that do well with AI will be the ones whose people still decide what the work should be, and who can tell when they have stopped. Start with one team, fifty exchanges and an afternoon.

ABOUT THE AUTHOR

Firoz Azees founded Ivanooo, which studies how people direct AI and builds tools to show it. Earlier versions of this argument circulated in 2026 as *The Capability Illusion*.

CITE AS

Azees, F. (2026). *Approved by nobody.* Ivanooo Research.

CONTACT

hello@ivanooo.com