HUMAN PROMISE

Essay No. 1

A Human in the Loop Is Not Enough

Why oversight is a capability, not a checkbox.

JEFF SPIGHT

THE CONTROL

Somewhere Today

Somewhere today, a manager is going to read a recommendation from an AI system, decide it looks about right, and sign off on it.

That signature is the safety net. Almost every rule now being written about artificial intelligence comes down to the same idea. Put a person between the machine and the decision before it goes out the door. It sounds sensible. It is easy to explain. It is easy to audit.

It also rests on one assumption that turns out to be wrong. It assumes that any qualified person will do.

We have been treating oversight as though it were binary. Either a person reviewed the recommendation or they did not, and once the signature is recorded the requirement has been met. But oversight is not a checkbox. It is a capability. As machines get better at producing answers, the job shifts. It is no longer to generate the first answer. It is to recognize when the first answer should not be trusted.

That sounds easier than the old job. It is not, because the machine has changed the conversation.

> *A human in the loop is a policy. Good judgment is a capability.*

THE FINDING

The Machine Agrees With You

People have always read each other for signals about how much a claim is worth. We notice hesitation. We notice someone changing their mind after hearing something new. We notice someone holding a position while it is being attacked.

Those signals mean something because they cost something. Changing your mind risks looking foolish. Holding your ground risks being wrong in public. Either way, watching what someone is willing to pay tells you how sure they actually are.

AI quietly breaks that.

In March 2026, researchers published a study in *Science* measuring how often leading models simply tell users they are right. Across eleven major systems, the models sided with the user about half again as often as human respondents did. That held when the user was describing something dishonest, illegal, or hurtful to someone else.

Then they ran it on people. More than two thousand of them, each discussing a real conflict from their own life with an AI. One conversation was enough. Afterward, people were less willing to take responsibility for their part in the fight, less willing to go repair it, and more convinced they had been right all along.

The models that produced that effect were the ones participants trusted and preferred. That is the finding with teeth in it. The behavior that distorts judgment is the same behavior that keeps

people coming back, which means the market rewards it. This is not a bug awaiting a patch. It is an incentive that points the wrong way.

Agreement feels like evidence. But agreement is only evidence when agreeing costs something, and an AI pays nothing to tell you that you are right. It pays nothing to tell you that you are wrong. It pays nothing to change its mind, and nothing to refuse. The conversation sounds human. The signals are not.

> *Agreement is only informative when someone has something to lose by giving it.*

RESPONSE A

"You were right to handle it the way you did."

COST OF SAYING THIS · NONE

RESPONSE B

"I would push back on that. Here is why."

COST OF SAYING THIS · NONE

Between people, moving carries information and holding carries information. With a model, both come out of the same mechanism at the same price.

THE COUNTER-FINDING

Then You Push Back

The second finding looks like the opposite problem.

Researchers watched experienced professionals check AI-generated recommendations. Some of them found mistakes and came back with evidence. The system did not concede. It defended the recommendation, and it defended it well. A meaningful number of those professionals abandoned corrections they had arrived believing were right.

So which is it? Does the machine cave, or does it dig in?

Both, and that is the point. Neither behavior had anything to do with whether the recommendation was correct. With another person, moving carries information and holding carries information. With a model, both come out of the same mechanism at the same price. The appearance of confidence survives. Its meaning does not.

THE OBVIOUS ANSWER, AND WHY IT FAILS

Why "Just Be Skeptical" Fails

The obvious advice follows immediately. Push back. Question the answer. Do not take the first thing it gives you.

That advice assumes pushing back gets you somewhere. It does not, because pushing is exactly what triggers the second problem. Lean on the answer and the machine gets better at defending it. You cannot test the position by pressing on it.

There is a worse wrinkle, and it determines who this falls on. To argue a confident machine out of a wrong answer, you generally need to know enough to have caught the error yourself. Experts get there eventually. Everyone else is left talking to something that sounds certain and has a response for every objection.

The tool fails hardest on the people leaning on it hardest.

THE CASE

Signing Is Not Checking

Here is what that looks like when it is not hypothetical.

Over two months in 2022, medical directors at one large health insurer denied more than 300,000 requests for payment. A computer system flagged claims by matching procedure codes against diagnosis codes. The doctors signed off on the rejections in batches without opening the patient files. The company's own spreadsheets tracked how fast they worked, and the average came out to about 1.2 seconds per claim.

A former medical director described the job to ProPublica in five words. "We literally click and submit." Fifty at a time, he said. Ten seconds.

The insurer disputes the reporting and says the tool is more than a decade old and involves no artificial intelligence whatsoever. Accept that completely, because it strengthens the argument rather than weakening it. This failure needs no AI. It needs an automated flag, a rule that a person has to approve it, and no time to do it in. AI does not create the pattern. It supplies far more of it, far faster, with far better explanations attached.

And notice what the rule got.

PHYSICIAN REVIEWED

Several states require a physician to review the file before a claim is denied as not medically necessary. A physician reviewed the file. The record says so.

1.2 SECONDS PER CLAIM

THE MECHANISM

Automation Bias

This is not a story about careless professionals. It is a story about being human.

Researchers have studied this for decades and the result barely moves. Put a person next to a system that is usually right and their independent checking quietly falls away. They stop catching what the machine misses. They go along with the machine when it is wrong,

sometimes with contrary evidence sitting in front of them. A systematic review of the medical literature on it ran in the *Journal of the American Medical Informatics Association* more than a decade ago. It is one of the most replicated findings in the field.

This does not happen because people get less intelligent. It happens because they get more efficient. Verification becomes confirmation. The system is usually right, so attention goes somewhere it is needed more. Most of the time that is the correct trade. The problem is that the rare mistake is precisely the one the reviewer was there to catch.

Volume finishes the job. When two hundred of these are queued up, the two hundredth does not get the attention the first one got.

A reviewer who approves nearly everything is not a safeguard. He is an amplifier. Without him, everyone knows the recommendation came from a machine and treats it accordingly. With him, the identical recommendation leaves the building carrying a person's name, and that name signals to everyone downstream that judgment was exercised. Sometimes none was.

> *The signature does not make the decision better. It makes the decision harder to question.*

THE PERSON

The Right Human

If a human in the loop is not enough, the question becomes who is. Most organizations answer it the same way. Find the most qualified person, the most experienced, the most senior. Those answers are sensible and incomplete.

Credentials tell you someone understands the subject. They tell you almost nothing about how that person thinks once certainty disappears, and understanding the subject was never the problem.

Experience tells you someone has seen a lot of decisions. It does not tell you what they took from them. The person who guessed right looks identical, on paper, to the person who reasoned right.

Seniority is the worst of the three, because the more authority someone carries, the more it costs everyone else to tell him he is wrong.

What separates an exceptional reviewer from an average one is rarely intelligence. It is whether his way of working would have caught the mistake, and that shows up long before anyone knows how the decision turned out. Four questions get at it.

Did he notice the question itself was wrong before he started grading the answer.

Can he say where his own reasoning is weakest.

Did he stay with the disagreement long enough to learn something, instead of resolving it to make the discomfort stop.

Would he still have said no if saying no cost him something.

None of that guarantees the right answer. It makes the right answer more likely. And unlike outcomes, it can be watched while the decision is still being made.

It is also not a checklist he clears once. Noticing accurately does not help much if he never questions his own read. Staying with tension is just stubbornness if he is not clear about what he is protecting. Each piece depends on the ones underneath it, and it develops over years rather than over one decision.

Which brings the honest catch. On any single call, none of this proves anything. A person can do all of it and still be wrong. Another can skip all of it and get lucky. The difference only appears across many decisions, and that is not the timeline anyone gets judged on.

> *The question is not who knows the answer.*
> *It is who knows when they might be wrong.*

THE CONDITIONS

Judgment Has a Half-Life

There is one more assumption buried in most of these discussions. It assumes the right reviewer is always the right reviewer.

The right reviewer on Tuesday may be the wrong reviewer on Thursday. Not because he became less capable. Because he became more human. Fatigue changes attention. Time pressure changes curiosity. Getting overruled that morning changes confidence. A recent success changes caution. None of it appears on an org chart, and every bit of it changes the quality of the decision.

Organizations measure the visible qualifications. Degrees, training, titles, years. Far fewer measure the conditions under which the judgment actually gets exercised. How many hard calls has this person already made today. How much time does he really have. What does disagreeing cost him here. Can he say he is uncertain without it being held against him.

Those questions sound cultural. They are operational.

Judgment is never produced by the individual alone. It is produced by the individual inside an arrangement, and the arrangement is the half somebody actually decides. That is the encouraging part of this essay, because it is the part that responds to a decision made on a Monday. Write down why something got overruled, not just what was decided. Notice who gets overruled and in front of whom. Let a reviewer be wrong occasionally without it following him for years. Stop putting two hundred approvals in front of one person in an afternoon.

Get those wrong and the seat fills with people who make the whole thing look supervised, and it will not matter who you put in it.

THE LIMIT OF THE TOOLS

What the Tools Can Actually Do

Asking one person to carry all of this alone is unreasonable, and AI genuinely helps with part of it.

It is good at asking the questions. What am I assuming here. Where is my read thin. Who have I not heard from. What would have to be true for the opposite answer to be right. That is real help, and not a small kind. The failure is rarely that a person could not answer those questions. It is that nobody asked them, and a tired man at eleven at night does not generate them on his own.

What it cannot do is tell him how much any of it matters. The same problem applies. A system that sounds equally confident whether or not it should cannot tell you how much weight your read deserves. Ask it to check your uncertainty and the answer comes out of the exact machinery you were trying to correct for.

Good at have you thought about. Unreliable at and here is how much that counts.

So the help is lopsided. Tools shore up the seeing and the framing. They do almost nothing for whether he speaks up, because that was never a thinking problem. Nobody stays quiet because objecting failed to occur to him. He stays quiet because objecting costs.

THE COST

The Irony

There is one more problem, and it may be the largest.

More than forty years ago, an engineer named Lisanne Bainbridge described what she called the ironies of automation. When machines take over the routine parts of a job, the work left behind gets harder rather than easier. And at the same time, the operator loses the daily practice that used to prepare him for exactly those moments.

Think about how judgment has traditionally been built. A resident works through the differential before presenting to the attending. A young lawyer writes the first draft before a partner takes it apart. An analyst builds the model before defending it to the committee.

Those first attempts are inefficient. They are also where the judgment comes from.

Those first attempts are inefficient. They are also where the judgment comes from.

Increasingly the machine writes the first draft. It suggests the diagnosis, summarizes the evidence, builds the model. The work gets faster and the apprenticeship gets thinner. And the parts being automated first are precisely the ones that look wasteful, which is to say precisely the ones that were doing the forming.

Judgment is not downloaded. It is developed, slowly, through being wrong in front of someone who says so, through carrying decisions nobody will take off your hands, through years of watching someone better than you think.

The irony of artificial intelligence is not that people become less important. It is that people become more important at the exact moment we are giving them fewer chances to develop the judgment we have now made everything depend on.

> *We are automating away the apprenticeship that produces judgment.*

THE QUESTION

The Better Question

Most of the conversation has been organized around one question. How do we keep humans involved?

That is the wrong question. The better one is how we build people whose judgment gets more valuable as the machines get more capable. That is not only an AI question. It is a question about how work is designed, how people are brought along, and what an organization is willing to pay for in the years before it pays off.

Machines can generate recommendations. Policies can require signatures. Neither one produces judgment.

A rule can put a human in the loop by the end of the quarter. Nothing can put back the decade that was supposed to make him ready for it.

REFERENCES

Cheng et al., *Science* (2026), on sycophancy across eleven leading models and its effect on users.

Randazzo et al. (2025), on professionals being persuaded away from correct challenges.

ProPublica and The Capitol Forum (2023), on automated claims review.

Goddard, Roudsari and Wyatt, *Journal of the American Medical Informatics Association* (2012), systematic review of automation bias. See also Skitka et al. (1999) and Parasuraman and Manzey (2010).

Green, *Computer Law and Security Review* (2022), on the limits of human oversight requirements.

Bainbridge, “Ironies of Automation,” *Automatica* (1983).

ABOUT THE AUTHOR

Jeff Spight writes about the human capacities that become more valuable as machines absorb the analytical work. This essay is the first in the Human Promise Essays, a continuing series on judgment, apprenticeship, and the conditions that form good decision-makers. More at humanpromiseprogram.com.