

How Would You Know?

Two people review the same document and reach the same conclusion. One of them thought about it. The other agreed with it. No inspection anybody currently runs can tell them apart.



WHERE THIS SITS

There is more language in the world than anyone can attend to. The obvious fix, an AI that reads it all for you and tells you what is true, does not work: the judgement that would take requires having been wrong at your own expense, and nothing a model does costs it anything. So the judging stays with people, and almost nothing we currently inspect can tell whether a person actually judged or simply agreed.

01 Why Does It All Feel Like Too Much?

02 Why Can't You Just Be Told?

03 What Holds a Claim Up?

04 **How Would You Know?**

05 Who Pays If It's Wrong?

06 What Would You Build Instead?

Six short pieces, each drawn from a working paper. Each one stands on its own; this is the fourth. They are ordered so that every piece ends by raising the question the next one is called.

Two people, one document

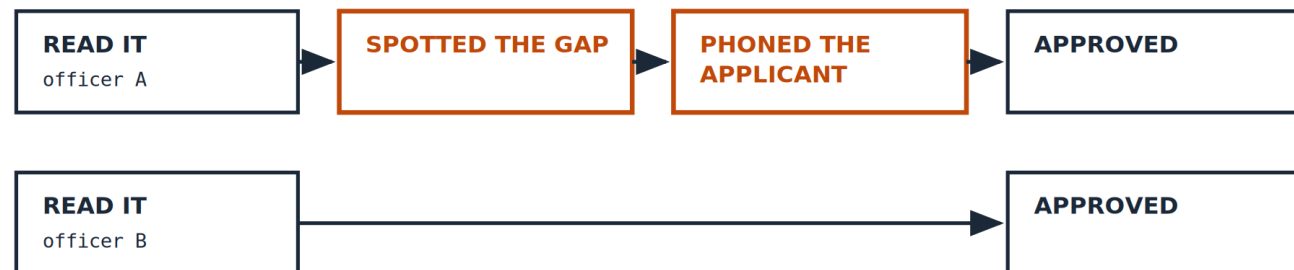
Two loan officers, the same application, the same recommendation. One phones the applicant about a gap in the employment history. One does not.

WHAT THE RECORD SHOWS



identical for both officers, in every field an auditor can reach

WHAT ACTUALLY HAPPENED



the two boxes in the middle are the only difference, and nothing records them

What the audit can reach

Imagine inspecting those two decisions with every power a regulator currently has. You can read the file. You can test the model for bias. You can check the applicant was given reasons. You can review the approval.

All of it is identical. The thing that differed happened in somebody's head and on a telephone. Not because anybody hid it. Because there is no field for it.

The question is not whether, it is how

Most rules ask whether AI was used, and require it to be disclosed. That is the wrong question, because both officers used it.

AS A SCAFFOLD

The output is where thinking starts. You read it looking for what it missed, what it weighted oddly, what you would have asked that it did not.

AS AN AMPLIFIER

The output is where thinking stops. It is fluent, it is plausible, it agrees with itself, and agreeing with it is quicker than not.

Same tool, same document, opposite use. That is the variable that matters, and it is invisible in everything we currently inspect.

1993, and a factor of ten

A final-year engineering project. The analysis package reported that the structure would stand, with the confident internal consistency such packages have.

10 × OUT

How far the stress figures were wrong, because loads had been entered in kilograms where the program expected newtons.

ONE CHECK

A back-of-envelope calculation, done by hand, found it.

The software was not broken. It answered exactly the question it was given and had no way to know the question was wrong. Everything downstream of that answer was fluent, consistent and false.

What you did not check

There is a name for the gap between what you signed your name to and what you would have signed had you done the work yourself. Call it the debt.

INVISIBLE WHEN TAKEN ON

Nothing happens. The document goes out. It is probably fine, and most of the time it is.

IT COMPOUNDS

Each acceptance makes the next one easier, and moves your own standard slightly.

IT ONLY SHOWS AT SCALE

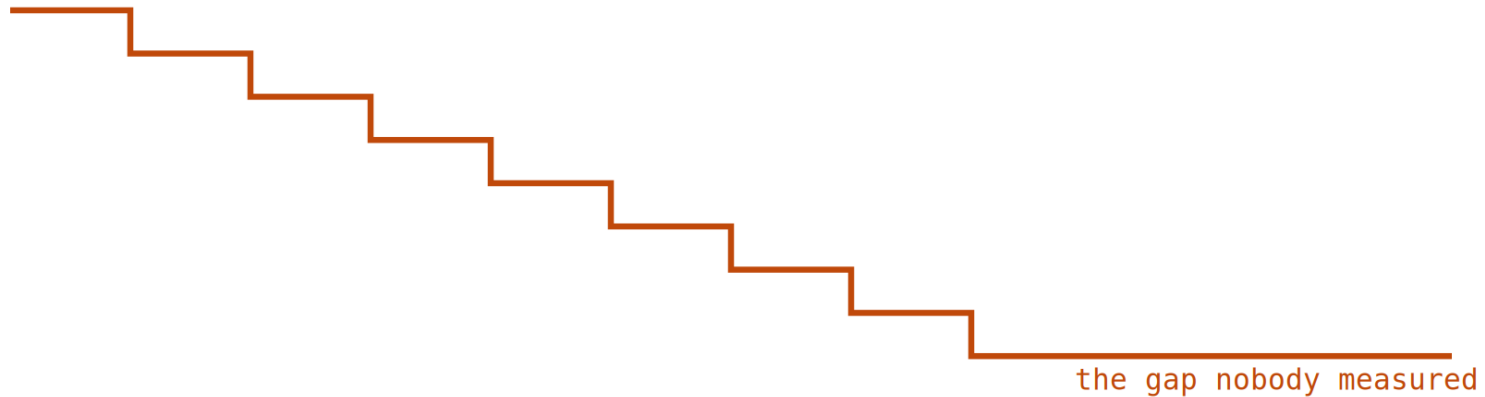
Never in one decision. In a cohort, over years, as a capacity that used to be there.

None of them is an event

Each decision is the same small size. The accumulation is the only thing worth seeing.

WHAT ANYBODY INSPECTS

every document above this line looks fine, and most of them are



one decision, then another, none of them an event

Twelve a year, a hundred and twenty a decade

Two researchers review the same paper, both starting from an AI-drafted review. The first adds a question about a competing method the draft ignored, and deletes a question that misread the notation. The second improves the tone and submits.

Somewhere in that hundred and twenty, the second reviewer stopped being able to do what the first one did, and no particular review was the one where it happened.

Why good prose gets through

None of this is a failure of character. It is a property of fluent writing. Text that is well organised, confident and internally consistent reads as competent, and reading as competent is enough to stop most people looking harder.

Checking is slow and disagreeable. Agreeing is fast and feels like work. The incentive runs the wrong way every single time.

Editing the tone is not reviewing

From the outside, the two look the same. The second reviewer did change the document. There is a tracked edit. There is evidence of engagement.

Any requirement to review AI output before use is satisfied completely by changing three adjectives.

This is the trap waiting for every framework that asks for engagement without asking what the engagement produced.

What we inspect

Take lending, among the most heavily regulated decisions anybody makes with a model.

REASONS

The applicant must be told why they were declined.

FAIRNESS

Outcomes must not discriminate across protected groups.

THE MODEL

It must be validated, documented and monitored.

APPEAL

There are rights to contest the decision after the fact.

These are real protections and they work on what they are aimed at. Every one of them is aimed at the artifact: the model, the decision, the paperwork, the outcome.

Risk below the line anybody looks at

There is a structural precedent and it needs stating carefully, because the substance is not the same.

Before 2008, a great deal of risk accumulated where the inspection regime of the day did not reach. The substance there was different: corrupted incentives, loans written with no documentation, ratings that did not describe what they rated. None of that is what is being alleged here.

What is shared is only the shape: something real building up underneath the layer everybody is watching.

The alert nobody reads

Hospitals have had decision support for decades, and with it a long record of what happens to a safeguard people stop attending to.

over 90%

The rate at which clinicians override decision support alerts, reported repeatedly across the literature on these systems.

Reported across the clinical decision support literature rather than from a single study

A warning overridden nine times in ten has stopped being a safeguard and become a formality. Nothing in any record marks the day it changed, and the compliance report still says the alert is in place.

What that looks like on a Tuesday

If the variable is whether the person thought, rules have to act on the conditions for thinking. Three domains, and in each one the same idea arrives looking completely different.

IN A HOSPITAL

Record what the physician would have ordered before the recommendation is shown. Put a floor under time per decision. Watch how often clinician and system agree, and from which direction.

IN LENDING

Require the officer to name one factor the model did not weight. Cap files decided per day. Track how divergence between officer and model is distributed across a team.

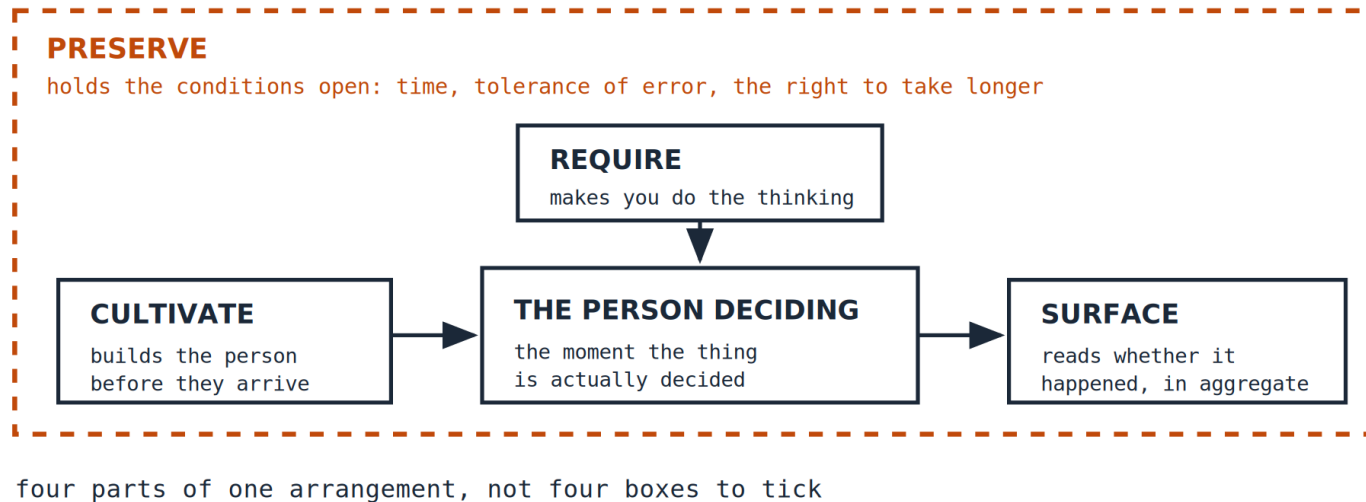
IN PEER REVIEW

Keep junior reviewers on unassisted manuscripts alongside a mentor. Require one issue the draft missed. Cap reviews per year, with a floor on time per review.

None of these inspects the output. All of them change what finishing the work requires.

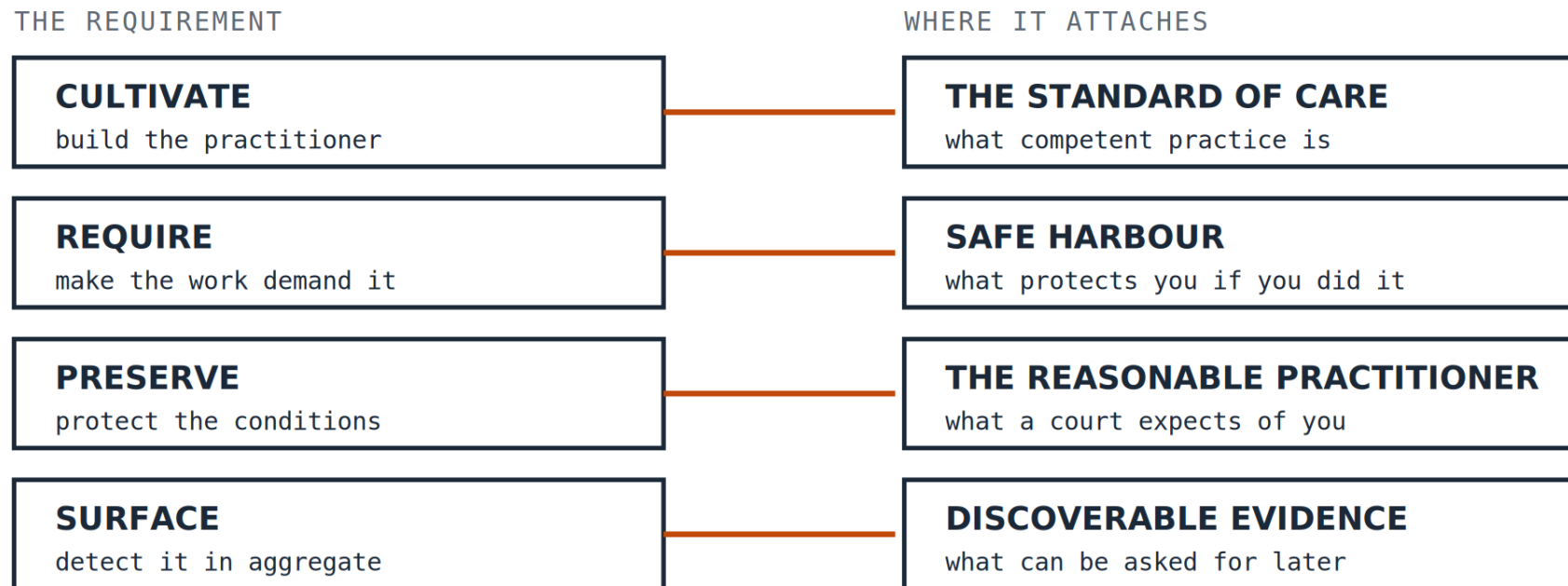
Four things, and not four more audits

Those three look like three different interventions. They are the same four things each time.



It bolts onto law that already exists

Not a new regulator. Each of the four attaches to doctrine that already exists.



How it would be gamed

Any requirement becomes a target. The paper names the three failures in advance rather than waiting to be told about them.

DOCUMENTING INSTEAD OF THINKING

Producing a record that proves you engaged, which takes the time you would have spent engaging.

PLAYING THE MEASURE

If divergence from the model is counted, diverge decoratively. The measure improves and nothing else does.

CEREMONY

The form is completed, everybody knows it means nothing, and it survives because removing it would look like lowering standards.

Gaming any of these still leaves the debt exactly where it was. That is the one consolation, and it is a thin one.

The problem with no answer

Everything in part three rests on somebody senior doing the cultivating: setting the standard, showing a junior what interrogation looks like, naming the number that smells wrong.

The people best placed to do that are the people who have worked alongside these tools the longest, which means they are carrying more of the debt than anyone they would be teaching.

The paper concedes this and does not solve it. It is the hardest problem in the framework and it is the author's own: he has been working this way for years and has no way to audit what it has cost him that does not rely on the very judgement in question.

What this does not claim

Short versions tend to overstate. Here is what this one is not saying.

✗ ~~**Here is the legislation.**~~

There isn't any. No law, no regulator, no institutional structure is proposed.

✗ ~~**Don't use AI for serious decisions.**~~

Not the argument. It is about the difference between two ways of using it, one of which is fine.

✗ ~~**Do these four and you are safe.**~~

Necessary, not sufficient. None of them can supply professional culture or institutional will.

✗ ~~**Surface tells you who is at fault.**~~

It flags, it does not convict. It cannot tell hollowing-out from genuine excellence with the tool.

Try it on the rule you already have

Most organisations have now issued something close to this: disclose when you have used AI, and review the output before you rely on it. It sounds reasonable. Hold it against the three questions.

WHAT DOES IT INSPECT?

Whether you declared it, and whether you looked. Both are properties of the document. Neither is a property of your thinking.

WOULD BOTH OFFICERS PASS?

Yes. Both used the model. Both would disclose it. Both reviewed the recommendation before approving. The rule cannot separate them, and separating them was the only thing worth doing.

WHAT DOES COMPLYING COST?

A tick in a box and a few minutes reading. Nothing that requires you to have thought, which means nothing that stops you not having.

That is not a bad rule. It is a rule aimed at the artifact, in a problem that lives somewhere else.

Where this comes from

This is a short version of a working paper called AI Governance and the Architecture of Practice, which carries the cases, the sources and the objections, including the ones that have not been answered.

It is part of a series on what judgement is, why it has to be carried by people who bear the consequences of being wrong, and what that means for building with machines that bear none.

Working papers are revised when they are shown to be wrong, and what changed is kept on the record.

James Maconochie | Architecture & Attention

HUMILITAS, HUMANITAS, ET CURIOSITAS · SEMPER

Humility, humanity, and curiosity. Always.

