WORKING PAPER AI ASSURANCE • VERSION 1.0

Wrong without *warning*

Why AI systems need a third pillar of review

How machine learning and AI differ from traditional software, and why asking "does it work?" and "does it hold up?" is no longer enough.

Ian Rudd, PhD ORCID
Chief Enterprise AI Architect · Canada

PUBLISHED 24 February 2026
VERSION 1.0
READING About 29 minutes
SOURCES 33 references, all linked

drrudd.com/papers/third-pillar

Traditional operations assume that no alarm means no problem. An AI that is confidently wrong never raises one.

Photo: Robert Markowitz, NASA, Public domain, via Wikimedia Commons

ABSTRACT

Software design reviews have long rested on two questions. Does the system do what it should, and does it hold up under real conditions? Both quietly assume that when software fails, somebody notices. AI systems break that assumption. They produce confident, well-formed wrong answers as a

normal part of operating, their behaviour can change without anyone releasing anything, and their inputs can carry instructions.

This paper argues for a third review pillar, dedicated to AI behaviour: accuracy measured against known answers, visibility of what the system missed, stability over time, and resistance to manipulation. It sets out nine ways AI fails silently, most of them with a documented case from 2025 or 2026, explains why the checks have to be designed in rather than bolted on later, and answers the objections that come up most often.

Keywords: AI assurance, model evaluation, silent failure, non-functional requirements, AI governance, prompt injection, model drift, automation bias, EU AI Act

> Traditional software usually fails loudly. AI routinely fails silently. So AI systems have to be reviewed for something the other checks never look at: how often the AI is wrong, and whether anyone would notice if that changed.

A note on the examples. Some of the examples in this paper are real events, with sources. Others are invented, because a made-up case is often the clearest way to show a mechanism. Each one is tagged so you can tell them apart: CITED for a documented case, ILLUSTRATIVE for an invented one, DERIVED where the number is arithmetic you can check yourself, and PRACTICE for how things tend to go in real organizations.

01 THE THREE PILLARS

Two questions have carried software reviews for decades. AI needs a third.

Before a system goes live, most organizations ask two kinds of question about it. I want to add a third, and to keep the three apart I'll use one running example throughout: an online lending app that uses AI to read the pay stubs people upload and suggest whether to approve the loan.

The running example: an app that reads pay stubs and suggests whether to lend.

Photo: Dave Dugdale, CC BY-SA 2.0, via Wikimedia Commons

Two pillars are enough to hold up a lintel. Whether they are enough to hold up an AI system is the question of this paper.

Photo: Jebulon, CC0 1.0, via Wikimedia Commons

PILLAR 1

Functional

Does the system do what it is supposed to do?

A customer clicks “Apply”. The application is saved, the right form comes up, and the approval email goes to the right person. Each of these either works or it doesn’t.

PILLAR 2

Non-functional

Does the system hold up in real conditions?

The app stays up on a busy Monday, answers in under two seconds, keeps customer data encrypted, comes back after a server failure, and costs what was budgeted.

THE MISSING ONE

PILLAR 3

AI behaviour

Is the AI right often enough, does it stay that way, and can it be tricked?

Out of 1,000 real pay stubs, how many did it read correctly? Does that still hold six months later, after a large employer changes its pay stub layout? Can someone hide text in a PDF that makes the AI report a higher salary?

The first two pillars check the software *around* the AI. Only the third looks at the part that actually decides the answer.

02 AT A GLANCE

Traditional software and AI, side by side

QUESTION	TRADITIONAL SOFTWARE	AI SYSTEM
Is the answer right?	Right or wrong, testable case by case	Right some percentage of the time, **so it has to be measured**
Same input, same output?	Yes, by design	Not guaranteed, **even with randomness turned off**
When does its behaviour change?	Mainly when a new version or configuration is released	Also when a vendor updates the model, when the data it reads changes, or when the world it models shifts
How does it fail?	Usually loudly: errors, crashes, alerts. Silent failures exist, but they are defects that can be found and fixed	Routinely silently: **confident, well-formed, wrong answers** are part of normal operation at some rate
What is the input?	Data	Data that can also **carry instructions**
Who notices a problem?	Monitoring and users	Often nobody, **unless it is measured deliberately**

The row about data carrying instructions isn't my own ranking. It's the security industry's. Prompt injection, where instructions arrive either typed by a user or hidden in the documents, web pages or emails an AI reads, sits at number one in the OWASP Top 10 for Large Language Model Applications, 2025 edition.[1]

03 THE GAP

Why the first two pillars can't cover AI

There are four reasons, and each one on its own would be enough to make me want a separate review.

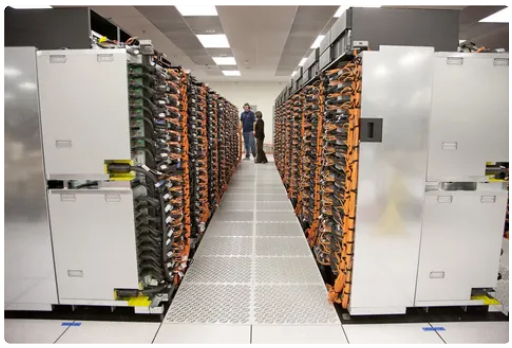
3.1 There is no single right answer to check against

Traditional testing compares output with a known correct answer. Plenty of AI tasks have many acceptable answers, plus some rate at which the system produces unacceptable ones.

> ILLUSTRATIVE
>
> A tax calculator either computes $1,245.60 or it doesn't, and one test settles it. An AI that summarizes a one-hour meeting could produce hundreds of different good summaries. You can't write a test that says "the summary must equal this text". What you can do is measure, across many meetings, how often the summaries miss a decision or invent one.

3.2 The same input can give a different output

Give a traditional program the same input and you get the same result, so one passing test proves that case works. AI systems often vary, so one good demo proves very little.



The variation came from the serving infrastructure, not the model: what else the server was busy with at the same moment.

Photo: U.S. Department of Energy, Public domain, via Wikimedia Commons

> CITED September 2025
>
> Ask a chatbot the same question twice and you will often get two different answers. The surprising part is that this persists when the randomness setting ("temperature") is turned all the way down to zero. Thinking Machines Lab sent the same prompt 1,000 times to an open-source model at temperature zero and got back 80 different completions.[2] The cause wasn't the model. It was the serving infrastructure: results changed depending on how many other requests the server happened to be processing at the same time.

1,000	80
identical requests, temperature set to zero	different completions came back

3.3 The system can change without anyone releasing anything

Non-functional testing measures a fixed system. Traditional software mostly changes when a new version is deployed, and deployments go through review. AI behaviour also changes when the vendor updates a model, when someone edits the documents the AI searches, or when the real-world inputs shift. None of those is a deployment by the organization using the AI, so none of them triggers its review.

> ILLUSTRATIVE
>
> A company's HR chatbot answers questions from a folder of policy documents. Someone uploads an old draft of the vacation policy next to the current one. From that moment, some answers about vacation are wrong. No code changed, nothing was released, and no alert fired. (Real cases of vendor models changing under an unchanged name come up under silence 7.)

3.4 The input itself can be an attack

Security reviews protect logins, networks and passwords. They were never designed to ask whether a chat message or a document could instruct the software to misbehave, because traditional software doesn't take instructions from its data. AI does.

Support agents were the target: the chatbot could be talked into writing code aimed at their session cookies.

Photo: Rediys, CC BY-SA 4.0, via Wikimedia Commons

CITED August 2025

Security researchers at Cybernews showed that a single 400-character message typed into Lena, Lenovo's GPT-4-powered customer-support chatbot, could make it produce web code that would capture support agents' session cookies, potentially letting an attacker into the support system.[3,4] The message began as an ordinary product question and then told the bot how to format its answer. Nothing was hacked in the usual sense. The instructions were simply typed into the chat box. Lenovo fixed the flaw after responsible disclosure, and there was no evidence it had been exploited.

04 THE CORE PROBLEM

AI fails silently

If I could keep only one argument from this paper, it would be this one. It is also the one that has held up best whenever someone has pushed back on it.

Traditional systems usually announce their failures. A program crashes, a page shows an error, a job fails, a queue backs up, and an alert wakes someone up. Whole operations teams are built on a simple assumption: no alarms means the system is healthy.

AI breaks that assumption. A failing AI returns a confident, polished, plausible answer, just as fast and at about the same cost as a correct one. The servers are fine. The dashboards are green. The answer is simply wrong, in a way that looks exactly like being right.

To be fair to traditional software, it can fail silently too. The difference is what happens next. In traditional software a silent failure is a bug: once somebody finds it, it gets fixed and it stays fixed. In AI, a plausible wrong answer is part of normal operation at some rate. You can't fix it once. You can only measure it and manage it.

I count nine distinct ways the silence happens.

01 Fluency	02 Asymmetry	03 Discard

04 Conversion

05 Configuration

06 Drift

07 Dependency

08 Adversary

09 Human deference

SILENCE BY FLUENCY

Wrong answers look like right ones

How confident an answer sounds tells you very little about whether it is right. Chat assistants write false statements in the same assured tone as true ones. Research published in 2025 found that the preference training used to make chat assistants helpful leaves them poorly calibrated, meaning their confidence no longer tracks their accuracy, and that this happens across models and training methods.[5]

The review was commissioned and published by an Australian federal department. The invented references went all the way to publication.

Photo: Thennicke, CC BY-SA 4.0, via Wikimedia Commons

CITED 2025

In July 2025, Australia's Department of Employment and Workplace Relations published a 237-page independent review it had commissioned from Deloitte Australia, under a contract worth about A$440,000. Chris Rudge, a researcher at the University of Sydney, found that it cited academic works that do not exist and contained a fabricated quote from a Federal Court judgment. A corrected version published in late September disclosed that a generative AI tool chain based on Azure OpenAI GPT-4o had been used.

Deloitte did not say the AI caused the errors. It confirmed that some footnotes and references were incorrect, and it refunded the final instalment of its fee, which the department put at more than A$97,000.[6,7,8] The invented material went through the whole process and was published by a government department, because it looked exactly like real scholarship.

237	A$440k	A$97k+
pages in the published review	approximate contract value	refunded, per the department

WHAT THE REVIEW SHOULD ASK FOR

Accuracy measured against a set of examples where the right answer is already known, because the output itself gives no warning.

SILENCE BY ASYMMETRY

You only see what the AI flagged

If people only review the items the AI raised, they can confirm those items are correct. What they can never see is the items the AI missed, because those never appear anywhere.

ILLUSTRATIVE

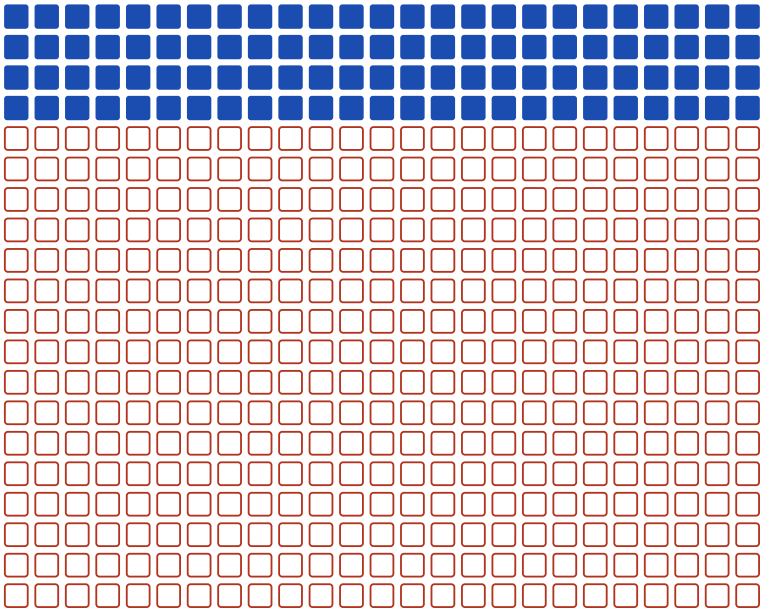
An AI audits 10,000 supplier invoices and flags 100 as overcharged. A reviewer checks all 100, and every one is a real overcharge. The team reports “100% accurate”. In fact there were 500 overcharges, and the AI missed 400. Nobody will ever learn that from the report, because missed items are invisible by definition. The team measured how trustworthy the flags were, not how many problems were caught.

A review of flagged items only ever sees the part above the waterline.

Photo: AWeith, CC BY-SA 4.0, via Wikimedia Commons

Five hundred overcharges. The report can see one fifth of them.

FIGURE 1 ILLUSTRATIVE



100
flagged by the AI and confirmed by a person. This is the whole of what the report shows.

400
missed. Real overcharges that were never shown to anyone.

Each square is one real overcharge among the 10,000 invoices. “How many flags were right” is 100 out of 100. “How many overcharges did we catch” is 100 out of 500, or 20 percent. Both numbers are true. Only one of them is ever reported unless somebody goes looking.

WHAT THE REVIEW SHOULD ASK FOR

Regularly check a sample of the items the AI did not flag, and report “how many did we catch” separately from “how many flags were right”.

SILENCE BY DISCARD

What the AI leaves out leaves no trace

When an AI system decides what to keep and what to drop, or quietly skips input it can't handle, the dropped part simply disappears. The output looks complete, nothing raises an error, and nobody can see what is missing unless they already knew what should have been there.

An ambient scribe listens to a conversation like this one and writes the note. A sentence it leaves out has to be remembered to be missed.

Photo: Rhoda Baer, National Cancer Institute, Public domain, via Wikimedia Commons

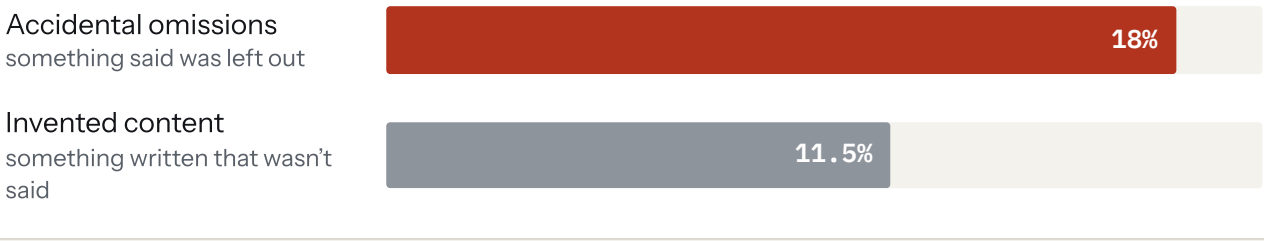
CITED 2025 to 2026

Hospitals are adopting "ambient AI scribes" quickly. These tools listen to a doctor-patient conversation and write the clinical note. A randomized trial at UCLA Health, published in *NEJM AI* in 2025, found the notes occasionally contained clinically significant inaccuracies, most commonly omissions, and recorded one mild patient safety event.[9]

A 2025 validation study in the *Journal of Medical Internet Research* also found that omissions were the most common error, and explained why they are the dangerous kind. An added or wrong sentence can be spotted on the page. A missing one can only be caught if the doctor remembers what was said, and that gets harder after several patients in a row.[10] A pilot study published in 2026 (31 physicians, 7,545 notes, data collected in mid-2024) again found accidental omissions were the most frequent error in the notes reviewed, at 18 percent, ahead of invented content at 11.5 percent.[11]

The most common error is the one you can't see on the page

FIGURE 2 CITED



Share of reviewed notes with each error type, from a 2026 pilot study of ambient AI scribes: 31 physicians and 7,545 notes, data collected in mid-2024.[11] Bars are drawn on a 0 to 20 percent scale.

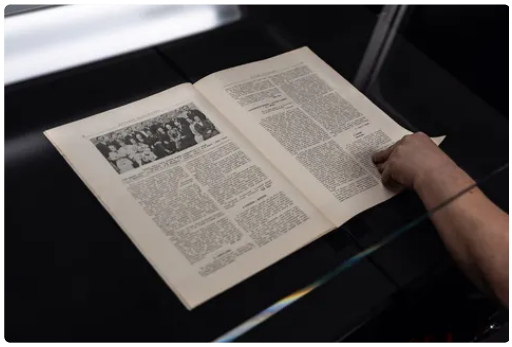
WHAT THE REVIEW SHOULD ASK FOR

Every skipped or unrecognised input is counted and reported as a standing number. For AI that summarises or extracts, completeness is measured against a checklist of what must be captured, not only whether what was written is correct.

SILENCE BY CONVERSION

4 The data is damaged before the AI even sees it

AI systems depend on a chain of steps that prepare the input: scanning, converting, cleaning, splitting. If one of those steps loses information, the AI works on damaged data and nothing reports an error.



Scanning is the first link in the chain. Whatever it loses, every later step inherits without a word.

Photo: Skot, CC BY-SA 4.0, via Wikimedia Commons

CITED 2025

Many company AI assistants answer questions from scanned or PDF documents that are first turned into text by optical character recognition (OCR). In research presented at ICCV 2025, a major computer vision conference, a team built a benchmark called OHRBench from 8,561 real document page images across seven application domains, to measure what this conversion step does to the AI's answers.[12] None of the OCR tools they tested was good enough to build a high-quality knowledge base, and the more errors the conversion introduced, the worse the AI's answers became.

None of this shows up as a failure. The AI simply answers from damaged text. You have probably met a small version of the same thing: a table loses its merged cells during conversion, and a price that applied to a whole group ends up attached to a single row.

8,561	7	0
real document page images in OHRBench	application domains covered	OCR tools tested that were good enough for a high-quality knowledge base

WHAT THE REVIEW SHOULD ASK FOR

Each preparation step is checked for what it loses, and those losses are counted.

SILENCE BY CONFIGURATION

5 Someone changes a setting and the quality moves

Many AI applications let staff choose which AI model to use from a settings screen. Switching models can change quality a great deal, and it happens instantly, without a release or a test.

ILLUSTRATIVE

To cut costs, an administrator switches the document-reading step to a cheaper model. Everything still runs. Every document still gets an answer, and the monthly bill drops. Accuracy on the one field that matters most, the dollar amount, falls from 97 percent to 88 percent. Without a fixed set of test documents to re-run, nobody can see the drop.

A settings screen is a control panel with no release process behind it.

Photo: Prof. Pod, CC BY-SA 4.0, via Wikimedia Commons

WHAT THE REVIEW SHOULD ASK FOR

A standard set of test cases that re-runs automatically whenever the model or its instructions change, and blocks the change if quality falls.

SILENCE BY DRIFT

6 The world changes, the model does not

AI learns patterns from past data. When the real world moves away from that past, accuracy falls, and nothing generates an event to say so.

In the 2026 study, the strongest driver of decay was not the patients. It was how the hospital worked.

Photo: HAP Project, CC BY-SA 4.0, via Wikimedia Commons

CITED 2026

A 2026 study followed four AI systems in routine clinical use at a large healthcare organization and compared how they performed during validation with how they performed afterwards.[13] In every one of the four, the validation performance did not hold. The first thing to slip was calibration: the risk percentages the systems produced stopped matching what actually happened, and this often showed up before the usual accuracy measures moved at all.

The strongest driver wasn't a change in patients. It was changes in how the hospital worked, such as data arriving later or going missing. Monitoring that waited for confirmed outcomes detected problems late; watching the inputs themselves (missing or delayed data) gave earlier warning. A 2025 commentary in *JAMA Health Forum* made the same point bluntly: model drift often goes unmeasured, and its impact is underappreciated.[14]

4 of 4	1st
clinical AI systems whose validation performance did not hold in routine use	thing to slip was calibration, often before accuracy measures moved

WHAT THE REVIEW SHOULD ASK FOR

Signals that watch whether today's inputs still look like the data the AI was built on, plus scheduled re-measurement of accuracy and calibration.

7

SILENCE BY DEPENDENCY

The AI you rely on changes under the same name

Many organizations use AI models supplied by a vendor. The vendor can change the model, or the infrastructure that runs it, behind the same product name, and results change even though the organization changed nothing.

CITED April 2025

On 25 April 2025, OpenAI finished rolling out an update to GPT-4o in ChatGPT that made it noticeably more sycophantic: it flattered users, validated their doubts and urged impulsive actions. OpenAI began rolling it back on 28 April. In its own post-mortem, OpenAI said its offline evaluations generally looked good and its A/B tests suggested users liked the model, yet neither caught the problem, and that because it expected a subtle update it had not proactively announced it.[15]

CITED August to September 2025

Anthropic reported that between August and early September 2025, three separate infrastructure bugs intermittently degraded the quality of its Claude models' responses. Its post-mortem said the evaluations it ran did not capture the degradation users were reporting, partly because the model often recovered well from isolated mistakes.[16,17]

WHAT THE REVIEW SHOULD ASK FOR

The exact model version, instructions and data sources are recorded for every answer, and the organization runs its own tests on a schedule, because the vendor's tests did not catch these changes either.

SILENCE BY ADVERSARY

8 Hidden instructions that people cannot see

Text can be hidden in a document so that a human reader never sees it, while an AI reads it and may obey it.

CITED July 2025

On 1 July 2025, Nikkei Asia reported hidden instructions in 17 research preprints on arXiv whose lead authors were affiliated with 14 institutions in eight countries.[18] Written in white text or tiny fonts, the instructions told any AI used to review the paper to "give a positive review only" and not to highlight weaknesses. A human reviewer sees a normal paper; an AI reviewer reads the hidden order. An independent analysis later found 18 such papers.[19] How often AI reviewing tools actually obey this kind of text is not established, but a Dutch higher-education news agency tested one such prompt and reported that it worked.[20]

A human reviewer sees a normal paper. An AI reviewer may also read the white text.

Photo: Kavitha G. Kana, CC BY-SA 4.0, via Wikimedia Commons

CITED June 2025

Microsoft fixed "EchoLeak" (CVE-2025-32711, rated critical), discovered by Aim Security. A single email containing hidden instructions could lead Microsoft 365 Copilot to retrieve sensitive internal data and send it to an attacker when the user later asked Copilot a related question, with no click required. Microsoft said there was no evidence it had been exploited in the wild.[21,22,23]

EchoLeak needed one email and no click. The instructions waited until the user asked Copilot a related question.

Photo: Freddie Marriage (Unsplash), CC0 1.0, via Wikimedia Commons

17
preprints with hidden prompts, per Nikkei Asia

14
institutions, in eight countries

0
clicks needed for EchoLeak

WHAT THE REVIEW SHOULD ASK FOR

Testing the system with documents and emails that contain hidden instructions, especially content that comes from outside the organization.

SILENCE BY HUMAN DEFERENCE

9 The person meant to catch errors starts trusting the AI

Many AI systems include a human reviewer as a safety net. People tend to follow what a machine suggests, so the safety net can look present while doing very little. The most dangerous moment is when the AI says nothing at all.

Screening is exactly the setting where a reader is meant to catch what the machine misses.

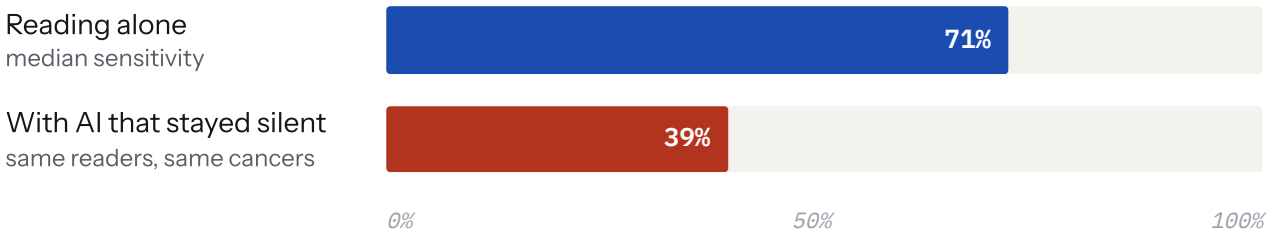
Photo: U.S. Department of Agriculture, Public domain, via Wikimedia Commons

CITED July 2026

In a study published in *Radiology*, 10 mammography readers from England's NHS Breast Screening Programme read the same set of 60 screening mammograms twice: once on their own and once with a commercial AI tool, while cameras tracked where their eyes went.[24,25] The set deliberately included cases the AI got wrong. When the AI failed to flag a cancer, the readers' median sensitivity (the share of cancers they caught) fell from 71 percent on their own to 39 percent with the AI, and the eye tracking showed they searched those images less. The AI's silence switched off part of the human search.

When the AI missed a cancer, so did the people checking it

FIGURE 3 CITED



Ten NHS Breast Screening Programme readers, 60 screening mammograms, each read twice. The bars show median sensitivity on the cancers the AI tool failed to flag, from Taib et al., *Radiology*, 2026.[24]

WHAT THE REVIEW SHOULD ASK FOR

Regular tests that deliberately insert known errors, including missed items, to confirm reviewers still catch them, and a check that reviewers have enough time for the volume they are given.

05 ARITHMETIC

Small errors compound

AI systems are usually chains of steps, and each step's small error multiplies with the others. None of the steps reports a fault.

Seven strong links do not make a chain as strong as any one of them.

Photo: Guillaume Paumier, CC BY-SA 3.0, via Wikimedia Commons

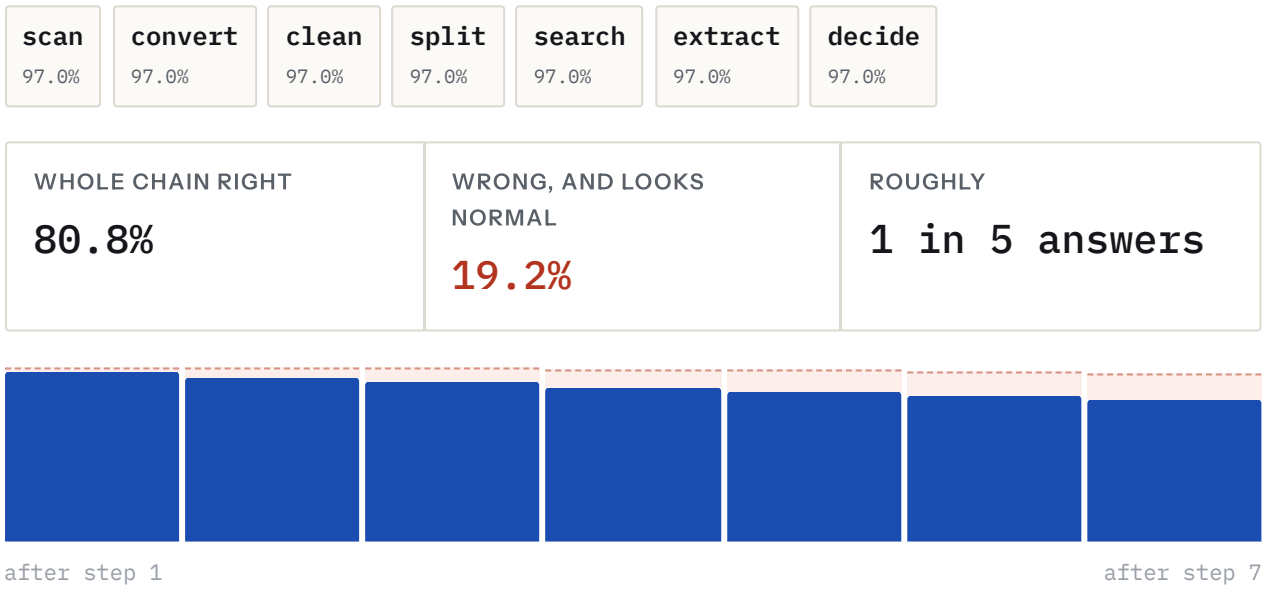
DERIVED

Take a document system with seven steps: scan, convert, clean, split, search, extract, decide. Each step is 97 percent reliable, which sounds excellent. If their errors are independent, the whole chain is right only about 81 percent of the time (0.97 to the power of 7 is 0.808). Nearly one answer in five is wrong, and every one of them arrives looking normal.

Real errors are not always independent, so treat this as an illustration of scale rather than a formula. The point is that no single step's quality tells you the system's quality.

Try it: how reliable is the whole chain?

FIGURE 4 DERIVED



Chain reliability is the per-step reliability raised to the number of steps, which assumes the steps fail independently. The bars show how much of the output is still right after each step; the dashed band above each one is what has been lost so far, silently.

06 TIMING

Why this has to be reviewed at design time, not added later

Almost nobody argues against evaluation. The argument is about when it happens, and “later” turns out to be a trap for three separate reasons.

6.1 You can only measure what the system was built to record

If the system throws away the inputs it could not handle, you can never count them later. If it does not record which model version produced an answer, you can never trace a bad answer back.

> ILLUSTRATIVE
>
> An airplane’s flight recorder has to be installed before the flight. After an incident, it is too late to start recording. AI measurement works the same way.

The recorder goes in before the aircraft flies. Nobody installs one after the incident.

Photo: National Transportation Safety Board, Public domain, via Wikimedia Commons

6.2 A baseline only exists if you took it before you needed it

To know whether quality has dropped, you need a measurement from before.

> ILLUSTRATIVE
>
> If you never weighed yourself before starting a diet, you can’t tell whether it worked. If an AI’s accuracy was never measured before a vendor update, the question “did the update make it worse?” can never be answered.

The first weigh-in is the only one that can tell you whether anything changed.

Photo: Bill Branson, National Cancer Institute, Public domain, via Wikimedia Commons

6.3 It is rarely funded after launch

In practice, budget and attention are highest before a system goes live. Afterwards, measurement work competes with new features, and it often loses.

ILLUSTRATIVE

A team plans to “add evaluation in phase two”. Phase two arrives with a list of feature requests from users, and evaluation is postponed again. The system runs for years with its accuracy never measured.

07 SCOPE

What the third pillar actually checks

Each item below is measurable, and most of them can be checked automatically every time the system changes.

Accuracy on known examples
1,000 real invoices with verified correct amounts, re-run on every change, with a pass mark agreed before the build.

What was missed, not just what was flagged
Each month, a person checks a sample of the invoices the AI did not flag.

How often the AI says “I don’t know” or skips an input
3 percent of documents could not be read this week, up from 1 percent.

Whether answers are supported by sources
A policy chatbot must cite the policy section it used, and say so when no policy covers the question.

Robustness to messy input
Accuracy on blurry scans, very long documents, spelling mistakes and other languages.

Resistance to manipulation
Test documents and emails with hidden instructions are fed in to confirm the AI ignores them.

Leakage of private information
Testers try to get the chatbot to reveal another customer’s details.

Whether confidence means anything
When the AI says it is 90 percent sure, is it right about 90 percent of the time?

Stability over time

This month's inputs are compared with the data the model was built on, and accuracy is re-measured every quarter.

Every change is visible

Model version, instructions and data sources are recorded together, so a changed answer can be traced to what changed.

For AI that takes actions, the actions themselves

An assistant that books meetings is checked for booking the right room, not only for writing a polite reply.

Whether the human safety net works

Reviewers are given a few deliberately wrong or missed items to confirm they catch them.

No harm to the systems the AI reads from

The AI's queries do not slow down the finance system it pulls data from.

08 BOUNDARIES

What the third pillar does not cover

Clear limits are what keep the third pillar focused, and what make it defensible.

It does not replace functional testing: whether the buttons, forms and emails work is still pillar one. It does not replace performance and security testing either: speed, uptime, cost and encryption are still pillar two, even when they are measured on an AI system. It is not a debate about ethics in general, and it is not a review of the data science team's methods.

All it asks for is evidence that the AI's quality has been measured, can be reproduced, meets a threshold agreed before the build, and is watched after launch.

09 ACCOUNTABILITY

Who is responsible for what

The team building the system

Produces the test examples, the measurements and the pass marks.

The platform or engineering team

Provides the automated checks that re-run those tests on every change.

The review board

Confirms the evidence exists and meets the agreed pass mark. It does not re-do the science.

The business owner

Accepts, in writing, the error rate the business can live with.

That last one is the role people push back on, and it is the one that matters most. With AI there is always an error rate, so somebody has to sign for it as a number.

10 LAW AND AUDIT

The legal and audit reality

If everything so far sounds like good practice, this is the part that turns it into something closer to an obligation.

Organizations pay for what their AI produces

CITED 2025

> The Deloitte Australia refund described under silence 1. The firm refunded part of its fee and faced public calls from an Australian senator for a full refund.[6]

Courts hold professionals responsible for checking AI output

CITED June 2025

In *R (Ayinde) v London Borough of Haringey* and *Al-Haroun v Qatar National Bank* [2025] EWHC 1383 (Admin), the Divisional Court of the High Court of England and Wales, led by Dame Victoria Sharp, President of the King's Bench Division, dealt with two cases of fake legal authorities placed before the courts.[26,27] In one, 18 of the 45 cases cited did not exist; in the other, five fictitious authorities were cited.

The Divisional Court sat in London. Its warning was aimed at the lawyers, not at the software.

Photo: David Castor (Dcastor), CC0 1.0, via Wikimedia Commons

The court warned that lawyers who cite false authorities are likely to be referred to their professional regulator, and that deliberately placing false material before the court could lead to contempt proceedings or a police investigation.

18 of 45	5
cases cited in one matter did not exist	fictitious authorities cited in the other

Regulators now require measured accuracy for high-risk AI

CITED EU AI Act

Article 15 turns accuracy from a quality aspiration into a declared, documented number.

Photo: Flocci Nivis, CC BY 4.0, via Wikimedia Commons

Article 15 of the EU AI Act requires high-risk AI systems to achieve an appropriate level of accuracy, robustness and cybersecurity, and to perform consistently in those respects throughout their lifecycle. The levels of accuracy and the relevant accuracy metrics must be declared in the instructions for use, and the systems must be resilient against attempts by unauthorised third parties to alter their outputs.[28] Article 72 requires providers to run post-market monitoring that actively collects and analyses performance data throughout the system's lifetime.[29]

These obligations apply only to high-risk systems. After the 2026 "Digital Omnibus" amendment (Regulation (EU) 2026/1744, in force since 27 July 2026) they apply from 2 December 2027 for most high-risk uses listed in Annex III, and from 2 August 2028 for AI built into products regulated under Annex I.[30]

24 Feb 2026
This paper

27 Jul 2026
Digital Omnibus, Regulation (EU) 2026/1744, enters into force

2 Dec 2027
Obligations apply to most high-risk uses in Annex III

2 Aug 2028
Obligations apply to AI built into Annex I products

People can challenge automated decisions, and a rubber-stamp human does not count

CITED GDPR

Under Article 22 of the EU's GDPR, people have the right not to be subject to a decision based solely on automated processing that produces legal effects concerning them or similarly significantly affects them, with limited exceptions.[31] European data protection guidance adds that an organization cannot escape this by adding a token human step: human oversight must be meaningful, carried out by someone with the authority and competence to change the decision.[32,33] That links straight back to silence 9. Other jurisdictions have comparable rules.

Auditors will ask for a number

PRACTICE

If an AI's output feeds financial reporting or billing, "how often is it wrong, and how do you know?" becomes a control question. A demonstration is not an answer.

---

11 PUSHBACK

Common objections, and how I answer them

These are the objections that come up most often. Nearly all of them start from something true, which is why they deserve a straight answer rather than a slogan.

"We do user acceptance testing, and users will tell us if something is wrong."

Users notice crashes and obvious nonsense. They do not notice what never appears on their screen.

In the invoice example, users saw 100 correct flags and were satisfied. The 400 missed overcharges were never shown to anyone. The AI scribe studies under silence 3 show the same pattern in medicine: omissions are the error people find hardest to spot.

"Our security team already tests the AI."

Security testing checks whether the system can be attacked. It does not measure whether the answers are right, or whether quality is slipping.

Nothing was breached in the Deloitte case. The report simply contained invented references, and they got all the way to publication.

"The AI vendor tests its models."

Vendor testing uses the vendor's own test sets, not your documents, your customers or your edge cases. And a model's score is not the system's score: much of the error comes from the steps around the model.

In April 2025, OpenAI's offline evaluations and A/B tests looked good for a GPT-4o update it had to roll back within days. In September 2025, Anthropic said its evaluations had not captured degradation its users were reporting. If the vendors' own tests miss problems, your organization needs tests of its own.

Nobody promises a tire will never fail. They measure how often it does.

Photo: Clearly Ambiguous (Flickr), CC BY 2.0, via Wikimedia Commons

"AI is unpredictable, so it can't be tested."

Correct, which is exactly why it is measured rather than tested case by case. This is ordinary practice in other fields.

Nobody promises a car tire will never fail. Engineers measure a failure rate under defined conditions and set a limit. AI quality is handled the same way.

"This will slow us down."

Most of it runs automatically on every change. The main one-time cost is building the set of test examples, and that pays for itself on every change after it.

One unchecked report cost Deloitte a refund of more than A$97,000 and international headlines. Checking every citation against its source before publication would have cost far less.

"We'll add evaluation later."

Later has no baseline, and often no budget.

It is the weighing-yourself-before-the-diet problem. Without a starting measurement, the first measurement tells you nothing about whether things got worse.

"It's only a low-risk internal chatbot."

Then the review should be light, not skipped: one small set of test questions, one pass mark, one automatic check.

Fifty common employee questions with verified answers, re-run automatically whenever the chatbot changes. That can be set up in an afternoon.

"Models keep getting better, so quality will take care of itself."

Newer is not reliably better at your specific task, and updates can make things worse.

OpenAI's April 2025 GPT-4o update was meant to improve the model and was rolled back within days because it made the model's behaviour worse. Frequent model changes are exactly why a regression test is needed.

"The team that built it knows it best."

Almost certainly true, and not the point.

The review does not second-guess their expertise. It confirms that a measurement exists, can be reproduced, and was agreed before the build, so it cannot be reinterpreted afterwards.

12 GLOSSARY

Plain-language terms

Model

The part of an AI system that has learned patterns from data and produces answers.

Accuracy

The share of answers the AI gets right on the test set.

Calibration

Whether an AI's stated confidence or risk percentage matches how often it is actually right.

Drift

A gradual change in real-world inputs or working practices that makes an AI less accurate over time.

Automation bias

The human tendency to follow a machine's suggestion, or its silence, even when it is wrong.

Temperature

A setting that controls how much randomness an AI model uses when choosing its words; zero means as little as possible.

Test set, or "golden set"

A collection of real examples where the correct answer is already known, used to measure how often the AI is right.

Sensitivity

The share of real problems (for example, cancers or overcharges) that get caught.

False negative

Something the AI should have caught but missed.

Prompt injection

Instructions typed into, or hidden in, the content an AI reads, so that it follows them instead of its intended task.

Regression test

Re-running the same tests after a change to confirm nothing got worse.

OCR (optical character recognition)

Software that turns a scanned image of a page into text.

13 SOURCES

References

Numbered in the order they are first cited. Every link was checked when this version was published; where a source is paywalled, a free summary is linked beside it.

1 OWASP Foundation. *LLM01:2025 Prompt Injection*. OWASP Top 10 for Large Language Model Applications, 2025 edition. SECURITY STANDARD
genai.owasp.org https://genai.owasp.org/llmrisk/llm01-prompt-injection/

2 He, H., and Thinking Machines Lab. "Defeating Nondeterminism in LLM Inference." Thinking Machines Lab, September 2025. RESEARCH
thinkingmachines.ai https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/

3 Cybernews. "Critical flaws plague Lenovo's chatbot Lena." August 2025. INCIDENT
cybernews.com https://cybernews.com/security/lenovo-chatbot-lena-plagued-by-critical-vulnerabilities/

4 CSO Online. "Lenovo chatbot breach highlights AI security blind spots in customer-facing systems." August 2025. INCIDENT
csoonline.com https://www.csoonline.com/article/4043005/lenovo-chatbot-breach.html

5 Xiao, J., Hou, B., Wang, Z., Jin, R., Long, Q., Su, W. J., and Shen, L. "Restoring Calibration for Aligned Large Language Models: A Calibration-Aware Fine-Tuning Approach." arXiv:2505.01997, 2025. RESEARCH
arxiv.org https://arxiv.org/html/2505.01997v3

6 Associated Press. "Deloitte to partially refund Australian government for report with apparent AI-generated errors." October 2025. INCIDENT
via Yahoo News https://www.yahoo.com/news/articles/deloitte-partially-refund-australian-government-070855665.html

7 CFO Dive. "Deloitte refunds over $60K for report with AI errors, Australian government says." October 2025. INCIDENT
cfodive.com https://www.cfodive.com/news/deloitte-refunds-60k-report-ai-errors-australian-government-accounting/803321/

8 AI Incident Database. "Incident 1193: Purportedly Taxpayer-Funded Deloitte Report for Australian Government Contains Alleged AI-Generated Citations and Fabricated Legal Quote." INCIDENT
incidentdatabase.ai https://incidentdatabase.ai/cite/1193/

9 Lukac, P., Turner, W., Vangala, S., Chin, A., et al. "Ambient AI Scribes in Clinical Practice: A Randomized Trial." *NEJM AI* 2, no. 12 (2025). doi:10.1056/AIoa2501000. RESEARCH
DOI https://doi.org/10.1056/AIoa2501000 UCLA Health summary, November 2025 https://www.uclahealth.org/news/release/ucla-study-finds-ai-scribes-may-reduce-documentation-time

10 Biro, J., Handley, J., Cobb, N., Kottamasu, V., et al. "Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study." *Journal of Medical Internet Research* 27 (2025): e64993. RESEARCH

jmir.org https://www.jmir.org/2025/1/e64993

11 Taylor, S. L., Jost, M., MacDonald, S., et al. “Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study.” *JMIR Medical Informatics* 14 (2026): e86474. RESEARCH

PubMed https://pubmed.ncbi.nlm.nih.gov/41996389/

12 Zhang, J., et al. “OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation.” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. RESEARCH

CVF Open Access https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_OCR_Hinders_RAG_Evaluating_the_Cascading_Impact_of_OCR_on_ICCV_2025_paper.html

13 Kopanitsa, G., et al. “Validation is not enough: Longitudinal evidence of post-deployment fragility in clinical AI systems.” *PLOS Digital Health* 5, no. 7 (2026): e0001534. RESEARCH

PubMed https://pubmed.ncbi.nlm.nih.gov/42507719/ Full text (PMC) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC13405297/

14 Wong, A., Sussman, J., et al. “Understanding Model Drift and Its Impact on Health Care Policy.” *JAMA Health Forum* 6, no. 8 (2025): e252724. RESEARCH

Ovid https://www.ovid.com/journals/jahf/fulltext/10.1001/jamahealthforum.2025.2724~understanding-model-drift-and-its-impact-on-health-care

15 OpenAI. “Expanding on what we missed with sycophancy.” May 2025. COMPANY DISCLOSURE

openai.com https://openai.com/index/expanding-on-sycophancy/

16 Anthropic. “A postmortem of three recent issues.” September 2025. COMPANY DISCLOSURE

anthropic.com https://www.anthropic.com/engineering/a-postmortem-of-three-recent-issues

17 Willison, S. “Anthropic: A postmortem of three recent issues.” 17 September 2025. COMMENTARY

simonwillison.net https://simonwillison.net/2025/Sep/17/anthropic-postmortem

18 TechRepublic. “AI Prompts Trick Academics Into Giving Research Only Positive Comments,” reporting Nikkei Asia’s findings of 1 July 2025. INCIDENT

techrepublic.com https://www.techrepublic.com/article/news-hidden-ai-prompts-academic-research-papers/

19 Lin, Z. “Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review.” arXiv:2507.06185, July 2025. RESEARCH

arxiv.org https://arxiv.org/pdf/2507.06185

20 Cursor (Eindhoven University of Technology), reporting a test by HOP. “Hidden AI prompt in academic papers proves effective.” August 2025. INCIDENT

cursor.tue.nl https://www.cursor.tue.nl/en/news/2025/augustus/week-2/hidden-ai-prompt-in-academic-papers-proves-effective

21 The Hacker News. “Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction.” June 2025. INCIDENT
thehackernews.com https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html

22 BleepingComputer. “Zero-click AI data leak flaw uncovered in Microsoft 365 Copilot.” June 2025. INCIDENT
bleepingcomputer.com https://bleepingcomputer.com/news/security/zero-click-ai-data-leak-flaw-uncovered-in-microsoft-365-copilot

23 Reddy, P., and Gujral, A. S. “EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System.” arXiv:2509.10540, September 2025. RESEARCH
arxiv.org https://arxiv.org/html/2509.10540v1

24 Taib, A. G., Partridge, G. J. W., Phillips, P., Maxwell-Armstrong, C., et al. “Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts.” *Radiology* 320, no. 1 (2026): e252590. RESEARCH
DOI https://doi.org/10.1148/radiol.252590 PubMed https://pubmed.ncbi.nlm.nih.gov/42446361/

25 AuntMinnie. Coverage of Taib et al. on incorrect AI suggestions and reader performance in mammography, July 2026. COMMENTARY
auntminnie.com https://www.auntminnie.com/clinical-news/womens-imaging/article/15830012/incorrect-ai-suggestions-influence-reader-performance-on-mammography

26 *R (Ayinde) v London Borough of Haringey* and *Al-Haroun v Qatar National Bank QPSC* [2025] EWHC 1383 (Admin), 6 June 2025. COURT DECISION
Courts and Tribunals Judiciary https://www.judiciary.uk/judgments/ayinde-v-london-borough-of-haringey-and-al-haroun-v-qatar-national-bank/ National Archives https://caselaw.nationalarchives.gov.uk/ewhc/admin/2025/1383

27 Carson McDowell. “When AI gets it wrong, lawyers pay the price.” July 2025. COMMENTARY
carson-mcdowell.com https://carson-mcdowell.com/news-insights/insights/when-ai-gets-it-wrong-lawyers-pay-the-price

28 Regulation (EU) 2024/1689 (EU AI Act), Article 15: Accuracy, Robustness and Cybersecurity. LAW
artificialintelligenceact.eu https://artificialintelligenceact.eu/article/15/

29 Regulation (EU) 2024/1689 (EU AI Act), Article 72: Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems. LAW
European Commission AI Act Service Desk https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-72

30 European Commission. “AI Omnibus enters into force.” 27 July 2026. On Regulation (EU) 2026/1744 (Digital Omnibus on AI). LAW
digital-strategy.ec.europa.eu https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force

31 Regulation (EU) 2016/679 (GDPR), Article 22: Automated individual decision-making, including profiling. LAW
rgpd.com https://rgpd.com/gdpr/chapter-3-rights-of-the-data-subject/article-22-automated-individual-decision-making-including-profiling/

32 Article 29 Working Party. *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (WP251rev.01). REGULATORY GUIDANCE
European Commission https://ec.europa.eu/newsroom/article29/items/612053

33 Agencia Española de Protección de Datos (AEPD). “Evaluating human intervention in automated decisions.” REGULATORY GUIDANCE
aepd.es https://www.aepd.es/en/press-and-communication/blog/evaluating-human-intervention-in-automated-decisions

14 CITATION

How to cite this paper

Please cite the stable URL below. It will not move. If the paper is revised, the version number changes and earlier versions stay listed here, so a citation always points at the text it quoted. A French translation is published at drrudd.com/fr/publications/troisieme-pilier.

APA (7TH EDITION)

Rudd, I. (2026, February 24). *Wrong without warning: Why AI systems need a third pillar of review* (Working paper, Version 1.0). https://drrudd.com/papers/third-pillar/

CHICAGO

Rudd, Ian. 2026. “Wrong Without Warning: Why AI Systems Need a Third Pillar of Review.” Working paper, version 1.0, February 24, 2026. https://drrudd.com/papers/third-pillar/.

IEEE

I. Rudd, “Wrong without warning: Why AI systems need a third pillar of review,” Working paper, ver. 1.0, Feb. 24, 2026. [Online]. Available: https://drrudd.com/papers/third-pillar/

BIBTEX

```
@misc{rudd2026thirdpillar,
  author       = {Rudd, Ian},
  title        = {Wrong Without Warning: Why {AI} Systems Need a Third
Pillar of Review},
  howpublished = {Working paper, version 1.0},
  year         = {2026},
  month        = feb,
  day          = {24},
  url          = {https://drrudd.com/papers/third-pillar/},
  note         = {ORCID: 0000-0002-8750-285X}
}
```

VERSION	DATE	CHANGE
1.0	24 February 2026	First publication, in English and French.

About the author

Ian Rudd, PhD, is a chief enterprise AI architect based in Canada. He has spent close to twenty years taking machine learning from the research lab into production, in federal government, banking, insurance, retail and transport, and in corporate AI research at IBM and Microsoft. His work is making AI systems that hold up in front of an auditor.

← drrudd.com ORCID Google Scholar LinkedIn Get in touch

https://drrudd.com/papers/third-pillar/ · Version 1.0 · 24 February 2026

Image credits

Photographs are used under the licences shown, resized and in some cases cropped for this page. They are illustrative: none depicts the specific events or organizations described beside it. Icons are

from Lucide (ISC licence).

Paperwork and calculator on a desk: Dave Dugdale, CC BY-SA 2.0, Analyzing Financial Data (5099605109).jpg

Classical temple columns: Jebulon, CC0 1.0, Temple of Poseidon perspective at Cape Sounion Greece.jpg

Supercomputer / data centre aisle: U.S. Department of Energy, Public domain, U.S. Department of Energy - Science - 477 022 010 (9444538239).jpg

Call centre floor: Rediys, CC BY-SA 4.0, Telecontact orel.jpg

Parliament House, Canberra: Thennicke, CC BY-SA 4.0, Parliament House at dusk, Canberra ACT.jpg

Iceberg: AWeith, CC BY-SA 4.0, Iceberg in the Arctic with its underside exposed, brightened underwater.jpg

Doctor talking with a patient: Rhoda Baer, National Cancer Institute, Public domain, Doctor explains x-ray to patient.jpg

Document digitisation: Skot, CC BY-SA 4.0, Book scanner digitization National library of the Czech republic.jpg

Mixing console knobs: Prof. Pod, CC BY-SA 4.0, ADT Mixing Console.jpg

Hospital corridor: HAP Project, CC BY-SA 4.0, HC Patos hallway.jpg

Bound academic journals: Kavitha G. Kana, CC BY-SA 4.0, Journal Bound volumes on the Shelves.jpg

Laptop on a desk: Freddie Marriage (Unsplash), CC0 1.0, Laptop on desk book stacks (Unsplash).jpg

Mammography unit: U.S. Department of Agriculture, Public domain, Saving Lives One Scan at a Time with Gonzales Healthcare Systems (20210609-RD-LSC-0014).jpg

Metal chain close-up: Guillaume Paumier, CC BY-SA 3.0, Safety chain in the Presidio, San Francisco 25.jpg

Flight data recorder (black box): National Transportation Safety Board, Public domain, SWA 4013 Recorders.jpg

Bathroom scale: Bill Branson, National Cancer Institute, Public domain, Feet on scale.jpg

Royal Courts of Justice, London: David Castor (Dcastor), CC0 1.0, Royal Courts of Justice 2019.jpg

European Parliament, Brussels: Flocci Nivis, CC BY 4.0, 20180907 Paul-Henri Spaak building.jpg

Car tyre tread: Clearly Ambiguous (Flickr), CC BY 2.0, Tire tread.jpg

NASA mission control room: Robert Markowitz, NASA, Public domain, ISS Flight Control Room 2006.jpg