WHITE PAPER · AI IN THE THINKING PROFESSIONS

What does this conclusion rest on?

An organization can meet every quality requirement and still not be able to show what a recommendation rests on. Whoever can show it also gets to move faster. This white paper is about those two sides.

VERSION

6.3 · August 20, 2026 · English edition

FOR

the thinking professions: research, consulting, publishing, and education, and the clients they serve

BASIS

every claim labeled, justification in the back

CONTENTS

IN BRIEF

Sooner or later a client calls about a recommendation in a report and wants to know what it is based on. That answer exists only if, during the research, someone recorded which findings support the recommendation and what material those findings lean on. No framework for knowledge organizations asks for traceability. ISO 20252, the standard for market, opinion, and social research, which covers policy research, tests whether the study as a whole is documented well enough to be reproduced. Where one specific conclusion came from falls outside that test.

AI raises the stakes, because a fabricated source is indistinguishable from a real one to the naked eye. With a transcript there is a safety net: the original sits right next to it. With a recommendation there is none. It takes shape in the researcher's head, and that is where the trail ends.

A better language model will not fix this. A different process design will. And that design does more than protect: the same agreement that removes the risk unlocks the gain, because work that demonstrably rests on material gets to move faster. Three of the measures in Part III cost nothing and can start this week. And sometimes the outcome is that AI adds nothing in a given spot. We put that on paper too.

INTRODUCTION

The question that comes months later

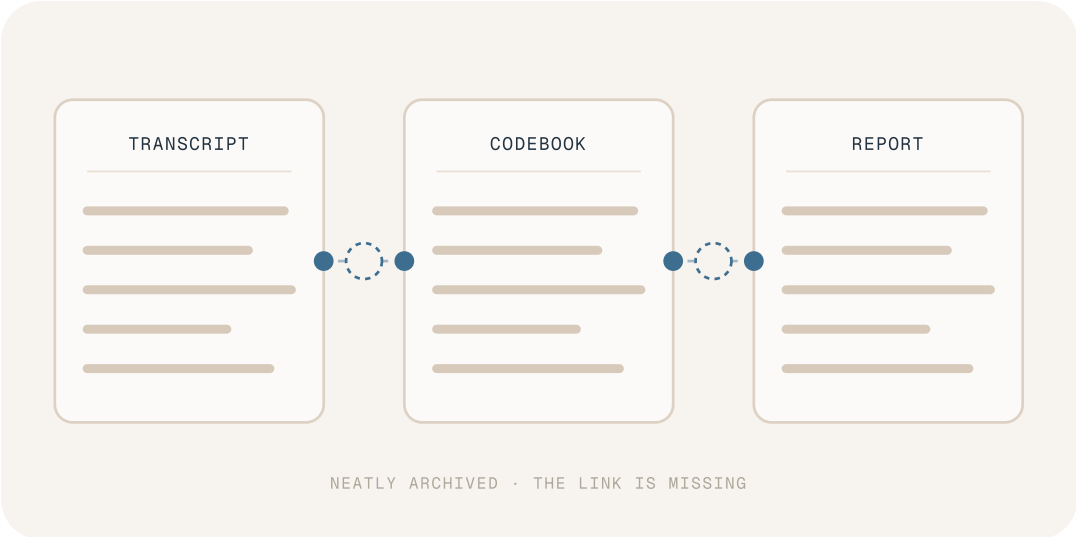
A knowledge organization records a great deal. The method is written up, the transcripts are stored, the codebook and the draft versions sit in the archive. Every link in the chain is there. What the standard way of working does not produce by itself is the connection across that chain: which findings a recommendation is based on, and what material those findings rest on.

As long as nobody asks, nothing seems wrong. The question often comes months after delivery, when the report starts to create friction somewhere: a city council is debating one recommendation, and the client wants to know which interviews and which analysis it rests on. Whether that answer exists at that point does not depend on how careful the work was. It depends on one design choice: was the connection recorded during the research, or does it have to be reconstructed from the archive after the fact?

This white paper is about that missing connection.

One agreement up front. Every factual claim in this document carries a label, VERIFIED, PLAUSIBLE or DISPUTED, and the justification per source sits in the back. That way the reader does not have to take this document on faith. Which is the point.

EVERYTHING SAVED, NOT THE LINK



Each piece is neatly archived. Only the link between the pieces is recorded nowhere.

Traceable is not the same as reproducible

Two terms look alike and get mixed up in practice. Reproducible means a study can be repeated with the same method and lead to comparable results. Traceable means every conclusion in the report can be followed back to the material it is based on. Quality systems secure the first. Clients ask about the second.

It happens to the firms that should know better

How this goes wrong, the field itself shows.

- **Deloitte Australia** refunded part of a government contract after fabricated sources and a fabricated quote from a court ruling were found in the report. VERIFIED
- **KPMG** pulled an international advisory report on AI offline after researchers at GPTZero checked its 45 source references: five held up, the rest turned out to be partly fabricated, distorted, or too vague to verify. VERIFIED
- **A Dutch lawyer** cited case law in court that did not exist, was called on it by the judge, and showed up at the same court in a later case with nonexistent rulings again. PLAUSIBLE

Since this spring the set is complete: EY Canada (May 2026) and PwC Middle East (late July 2026, four reports) also had to withdraw or correct work after fabricated sources were found. With that, each of the four big firms had its own case within a single year. PLAUSIBLE

None of these documents stood out until someone went and checked the sources. That is the core of it: this was work that looked like finished work, made by people who know their trade.

What the standard tests and what the client asks

Traceability is not a new word in this field. It appears in the quality handbooks and in the stated quality principles of nearly every knowledge organization: conclusions must be traceable, the researcher remains ultimately responsible, AI supports the expertise and does not replace it. That is meant seriously and taken seriously.

The misunderstanding sits in what gets tested next. ISO 20252 asks whether method, execution, and analysis are documented well enough for the study to be repeatable. That is a test of the whole. The client tests the part: they single out one conclusion and want to know where it came from, and whether a language model had a hand in it.

So an organization can be certified, work with integrity, and still owe an answer on that one question. That gap existed before AI existed. AI has made it visible, and bigger.

How often this happens, nobody has measured, and a percentage here would therefore be fabricated. In its place stands a test, and it takes fifteen minutes: take the last report you delivered, pick one recommendation, and try to show which findings it rests on and what material those findings rest on. If that works within fifteen minutes, this white paper confirms what you already have in place. If it does not, what follows shows where the trail breaks and what to do about it.

The standard tests the whole; the client points at one sentence and asks what it rests on.

How the rest of this document is built

Part I describes the state of play: where AI has already entered the research process, which rules apply, and what the existing quality frameworks do and do not cover.

Part II shows where the risk really sits, using SEAL: the four phases knowledge work moves through, in each of which the traceability of a claim can be lost. The risk is not in the transcription, and not in the finding that can be laid next to your own material, but in the step from finding to recommendation.

Part III is practical: three measures that require no budget, and TRACES, the six layers to test a system or a vendor against. Sometimes the outcome of that test is that AI adds nothing in a given spot. That, too, is an outcome.

I

PART I

Where things stand

CHAPTER 1

AI is already used in the analysis itself, mostly out of sight

Four moments from daily practice in a research department.

- A **researcher** pastes an interview transcript into a private ChatGPT account to have it cleaned up.
- A **project lead** has a client's confidential policy document summarized because the meeting starts at two.
- A **junior** hands the literature scan to a language model and copies the hits without checking them.
- A **senior editor** has the closing chapter rewritten, the evening before delivery.

None of these four actions is registered anywhere. They show up in no project file and no quality report, and precisely because of that, the scale of the use is so hard to measure.

Try to measure it with a survey and you will come in low. An employee working from a private account is not quick to report that in a questionnaire from their own employer. Behavioral data gets closer to reality.

Two measurements, independent of each other, land in the same place.

- **LayerX**, telemetry on what actually happens in browsers: about 45 percent of employees use generative AI. Most of it runs through unmanaged private accounts, and a sizable share of the uploaded files contains sensitive or personal data. PLAUSIBLE
- **Verizon**, Data Breach Investigations Report 2026: based on more than 850,000 logged events, the same percentage, three times as much as a year earlier. PLAUSIBLE

WHAT THIS NUMBER DOES AND DOES NOT SAY

The LayerX measurement covers employees in general, not knowledge organizations. A measurement for this sector does not exist, and that is a finding in itself: nobody has counted how many Dutch knowledge organizations use AI in the substantive analysis. The figure gives a direction, not a picture of the sector.

CHECK WHAT A PERCENTAGE MEASURES

Around AI adoption, percentages circulate that measure something other than what they seem to say. Take the 77 percent from the same LayerX study: that is not an adoption figure but the share of users pasting company data into an AI prompt. Anyone who repeats that number as the share of employees using AI is citing the source wrong. The control question is always the same: what was measured, among whom, and by whom. PLAUSIBLE

For a knowledge organization, something weighs in here that counts for less in many other sectors: the material itself. Interview data and confidential policy documents almost always contain personal data. A transcript in a private account is then not only a quality risk but a GDPR matter. The organization remains the data controller, even when a single researcher acted on their own initiative.

For whoever carries that responsibility, the question of whether AI is used has become too small. The usable question is where and how it happens, and whether anyone can show that afterward.

CHAPTER 2

Which rules apply and who oversees them?

AI legislation, meanwhile, is getting stricter. The rollout comes in stages, with no single hard deadline. And which rules apply depends heavily on the type of use and the risk class. With so much incorrect information going around, a clear overview pays off right now.

Two obligations from the European AI Act (Regulation (EU) 2024/1689) already touch every knowledge organization. The literacy obligation has been running since February 2025: every organization must be able to demonstrate that employees who work with AI know what they are doing. The transparency obligation applies since August 2, 2026, together with the enforcement powers of the national supervisory authorities. The panel below gives the full calendar. VERIFIED

The Digital Omnibus

The Digital Omnibus, the package the Council of the European Union and the European Parliament formally adopted on June 29, 2026, postpones the heaviest obligations for high-risk AI: from August 2026 to late 2027 and 2028.

But Article 4 and Article 50 stand. Anyone who reads this as “it has been postponed, I have time” is reading it wrong: precisely the obligations that touch most organizations, literacy and transparency, stay on schedule. VERIFIED

AI ACT · WHAT APPLIES WHEN

FEB 2, 2025 · IN FORCE

Article 4 (AI literacy) & Article 5 (prohibited practices)

Already running. An organization must be demonstrably AI-literate.

AUG 2, 2026 · IN FORCE

Article 50 (transparency) + enforcement powers

From this point, national supervisory authorities can enforce, including on Article 4.

DEC 2, 2026

End of the transition period for watermarking existing systems

DEC 2, 2027 · POSTPONED

High-risk AI (stand-alone, Annex III)

AUG 2, 2028 · POSTPONED

High-risk AI (embedded, Annex I)

Source: Regulation (EU) 2024/1689 and the Digital Omnibus (Council of the EU, June 29, 2026).

The first enforcement moment came from a professional body

The first real Dutch AI incident did not come from a regulator but from the legal profession. In February 2026, the deans of the Dutch bar issued formal warnings to three lawyers for improper use of generative AI in court filings; two of them were required to complete an AI course. The trigger: judges flagging references to case law that does not exist. PLAUSIBLE

Note the precise nature of this. It stayed at these warnings, the mildest form of oversight; it did not come to formal rulings by a Disciplinary Board. PLAUSIBLE

Even in that mild form, the lesson stands: your own professional body and your client will test you sooner than the legislator will.

CHAPTER 3

No framework for knowledge organizations requires traceability

Knowledge organizations in policy research and evaluation operate inside a compact web of sector-specific quality agreements. Yet none of those frameworks requires an organization to record, per conclusion or recommendation, where it came from.

Take the four frameworks that apply everywhere.

- **ISO 20252** is a quality standard for research organizations, built around documenting method, execution, and analysis so the study remains repeatable.
- The Dutch **Code of Conduct for Research and Statistics**, the industry code of MOA, VBO, and VSO, is about handling data with integrity, privacy, and data protection, and does not mention AI at all.
- The **Periodic Evaluation Regulation (RPE)** determines when government policy must be evaluated and requires methodological quality and independent review.
- The **Policy Evaluation Toolbox** from the Dutch Ministry of Finance requires nothing itself: it is an aid that walks clients and evaluators through five steps, from context and scope to reporting and communication, and so helps meet the RPE's quality requirements.

All these frameworks revolve around the process: is the study reproducible, with the same method and the same data? The client asks a different question: they point at a recommendation and want to know what it is based on.

> Reproducing a study and tracing a finding are two different things.

That is how a gap opens up between what the quality principles say and what the organization can demonstrate. Traceability is a principle there, not a mechanism. Nothing is set up to enforce the principle, and after the fact nobody can reconstruct the origin.

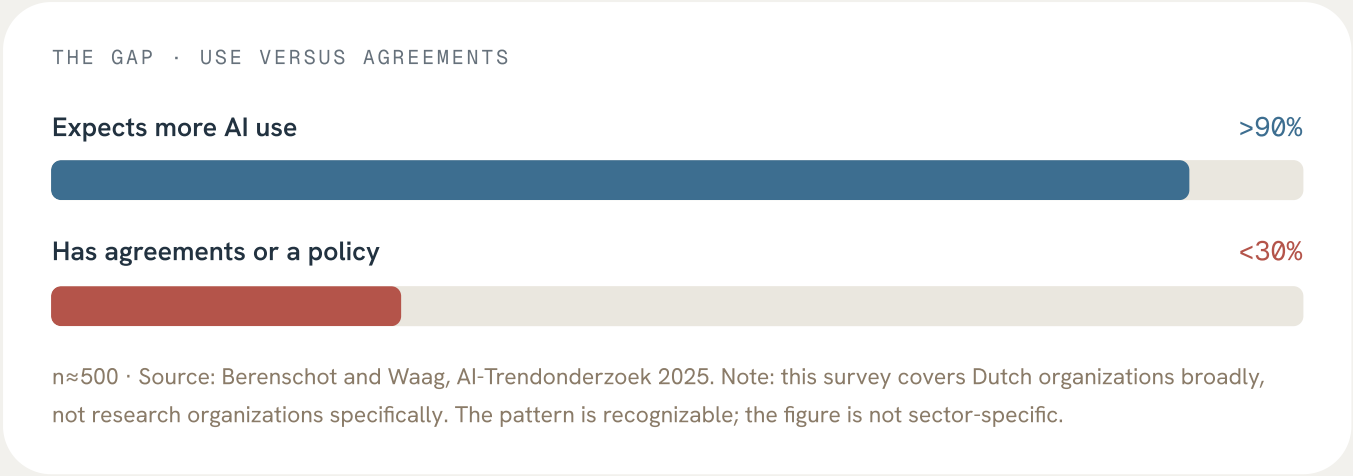
Academia does have one framework with a real traceability requirement, the Netherlands Code of Conduct for Research Integrity, with a revision around AI in the works and expected in 2026. That code applies to scientific research at knowledge institutions, not to an independent knowledge organization. So the sector will have to regulate itself.

CHAPTER 4

Nearly everyone uses AI, hardly anyone has it organized

The rules and the pressure are running ahead of where most organizations currently stand. That gap is the real problem, not the use itself.

A survey by Berenschot and Waag Futurelab among roughly 500 respondents shows that over 90 percent expect more AI use inside their own organization in the near future, while fewer than 30 percent have agreements or a policy. Let alone a system. VERIFIED



Put those numbers side by side: nearly everyone expects more AI use; fewer than a third handles it deliberately and with intent. That is not a sign of carelessness but of organizations that have not yet been handed a workable starting point.

And where policy does exist, it has a clear shape. Within that group, the guidelines are mostly about:

- privacy, data protection, and legal compliance (53 percent);
- general rules of use (51 percent).

Agreements on procurement, collaboration with AI vendors, and employee representation are scarce. And working control mechanisms are missing at most organizations, according to the researchers. PLAUSIBLE

The policy that exists, then, mostly governs what may go into an AI system. Who checks what comes out before it lands in a report is rarely recorded, even among the front-runners. The rest of this document is about that second half.

A second gap runs deeper and gets named less often. In qualitative evaluation research, the researcher is often already under pressure from the client.

As far back as 2011, sociologist Vasco Lub called it an open secret, on the Dutch platform Sociale Vraagstukken, that researchers are regularly pressed to phrase critical passages less negatively, often because a client insists that everyone stay constructive. PLAUSIBLE

That pressure is not new. What is new is how silently it can be given in to: a model softens a sharp phrasing into an acceptable one in seconds, where that used to cost a conversation and a conscience. Of that softening, nothing stays visible.

II

PART II

The real risk

CHAPTER 5

A fabricated source looks like a real one

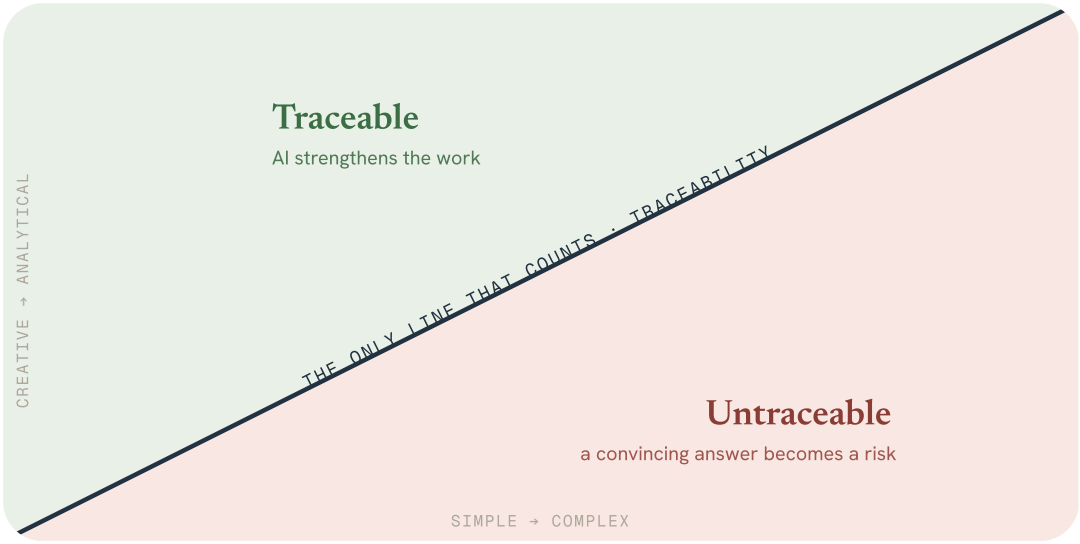
The most dangerous property of AI in research is not that it makes errors. Every tool makes errors. The dangerous part is that these errors look excellent.

A study of 758 consultants (Dell'Acqua et al., *Organization Science*, 2026) made both sides visible. VERIFIED

- **Where AI helped:** creative and writing work for a fictional shoe brand, from product ideas and market segmentation to persuasive copy and memos. Quality rose by roughly 40 percent and the work went a quarter faster.
- **Where AI hurt:** a task that resembles evaluation work, advice to leadership based on spreadsheet figures and interviews that contradicted each other. AI users scored 19 percentage points below colleagues without AI, and the odds of a wrong answer being adopted actually went up.

The dividing line between AI that adds value and AI that does damage therefore does not run between creative and analytical work, and not between simple and complex either. The dividing line runs between traceable and untraceable.

THE DIVIDING LINE THAT MATTERS



The usual axes do not matter. The only line that counts cuts straight across them.

The question is whether someone checking the report can find out, for every finding, conclusion, and recommendation, where it came from. Where that is possible, AI is a powerful aid. Where it is not, a convincing answer is more dangerous than a weak one, because it slips past the checks with ease.

CHAPTER 6

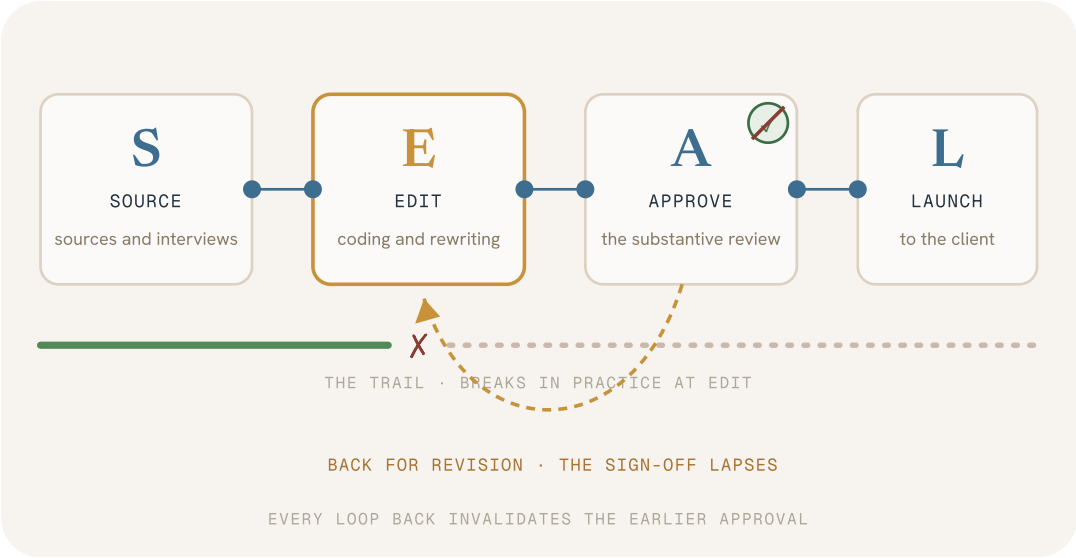
SEAL: the four phases where traceability gets lost

To point at where in the process a recommendation comes loose from its source, we use a fixed model, SEAL.

> **SEAL** describes the four phases knowledge production moves through: **Source, Edit, Approve, Launch**. In each phase, the traceability of a source can be lost, and the weakest phase determines what the work is worth.

For an evaluation report, Source is collecting sources and interviews; Edit is everything that happens to the material after that, up to the review, so coding, analyzing, and summarizing included; Approve is the substantive review; and Launch is the moment the report goes to the client. The work does not pass through the four phases once: after a review, a draft goes back for revision, and every time that happens, the earlier approval becomes invalid.

SEAL · FOUR PHASES, ONE LOOP BACK



The work does not run left to right once: after every review, a draft goes back for revision, and with that, the sign-off that was already in place lapses.

Things can go wrong in every phase, but not every phase carries the same risk. Where the trail breaks in practice is the subject of the next chapter.

CHAPTER 7

The unguarded step sits between finding and recommendation

The risk does not spread evenly across the four phases.

- **The transcript** can be checked: lay it next to the original and walk through the deviations.
- **The finding** can too: it points to passages and documents that sit in the archive.
- **The recommendation** cannot. It arises from interpretation, and no source text sits next to it for comparison.

That is exactly where a language model, under deadline pressure, smooths a weakly supported conclusion into fluent advice that leads back to nothing. And it is the recommendation that the client, the advisory committee, or the regulator will press on.

That is where the question runs aground. The finding is not fabricated; the connection between finding and advice was simply never recorded. In SEAL terms: **the unguarded step sits at Edit.** In editing, a sentence gets summarized, softened, or merged with another, and the reference to the passage it rests on gets lost along the way. In the final report, there is no way to see this happened.

The next phase should catch that, and right there sits a blind spot. In practice, Approve is a colleague or project lead reading through the draft: does the structure hold, is anything off, can we defend this to the client. That is useful and essential, but **it is reading, not checking.**

Whether a statement truly rests on the interviews and files cannot be pulled from the text itself. For that, the material has to come up next to every conclusion, and a read-through leaves no time for it. The sign-off, moreover, usually covers the report as a whole: the read-through tests the whole, while the client's question will later be about one specific conclusion or recommendation.

That is how the break from Edit survives the review. An approval that can handle that question looks different: the reviewer gets the material next to each conclusion, and somewhere it is recorded who approved which version.

CHAPTER 8

Experience is no protection against a fluent error

In the spring of 2026, a former editor-in-chief of NRC and De Standaard was suspended as a fellow at Mediahuis. Fifteen of 53 blog posts contained quotes the people involved had never said; seven of those quoted denied the statements. VERIFIED

He confirmed himself how it had gone: the quotes arose while summarizing reports with AI. It did not happen to him out of ignorance, but out of routine: summaries checked just short of well enough.

This case is no outlier. In January 2026, GPTZero, a company that detects AI text and fabricated sources, went through 4,841 accepted papers at NeurIPS, the largest AI conference in the world, and found 100 fabricated references across more than 50 papers. VERIFIED

- Part were **entirely fabricated**: nonexistent authors, a made-up title, a link that leads nowhere.
- Part were **more refined**: a decent paper with fabricated authors, or a fabricated position inside a sometimes fabricated organization.

Each of those papers had been reviewed by three to five peers, and the conference had a policy banning fabricated citations. It slipped through anyway, among people who build the technology themselves and were already trained and selected to the maximum. So better training or stricter selection is not the answer.

Through both cases runs the same thread: **work that looks finished no longer gets checked.** The check verifies that a reference looks right, not that it exists. Whether a source is real only shows when someone opens it, and with convincing work, that does not happen by itself. The process has to force it.

CHAPTER 9

Vigilance wears out, process design does not

The obvious measures are paying closer attention and training people. That helps, but it does not fix the problem structurally. And there is evidence for that.

In a randomized controlled trial (NEJM AI, 2025), 44 physicians first received 20 hours of AI literacy training. Even so, the physicians shown flawed AI advice reached only 73.3 percent diagnostic accuracy, against 84.9 percent in the group with correct advice: a difference of 11.6 percentage points, rising to 14.0 percentage points after adjusting for covariates. VERIFIED

Trained, alert, and they still followed the bad advice.

MEASUREMENT CONTEXT

The errors in this study were deliberately injected into the AI advice. It therefore measures automation bias, the tendency to follow a machine's advice, not the model's day-to-day error rate.

The mechanism is human. Reading an expertly reasoned piece, we take fluency and logic as competence, and doubt subsides. Training covers the legal literacy obligation, and that is useful, but it does not neutralize this reflex.

The distance from those physicians to research practice is small. An experienced researcher handed a draft finding from a language model is in the same position: the AI literacy course completed, the legal obligation covered, and still their own judgment weakens the moment the proposal is wrong but fluent.

Chapter 5 showed the same pattern: on a task where the sources contradicted each other, consultants with AI performed below colleagues without it. In evaluation research, that is not the exception but the core of the trade: interviews contradict each other, files contradict the interviews, and exactly there is where expertise makes the difference.

Organizations that want to use AI responsibly and well often underestimate this. The training itself is good and useful: it covers the legal obligation and teaches people what the systems can and cannot do. But whoever stops there is taking a risk.

Without deliberate AI use, systemic or set in policy, that keeps findings and sources linked, AI pulls the level down. Including among those who took the course seriously. The time savings are real. So is the price: conclusions worse than what the same team delivered without AI.

> Attention wears out. A fixed rule does not.

The difference between attention and a recorded rule starts to count as volume grows. Whoever reviews hundreds of reports a year will sooner or later miss one. Not from incompetence, but because nobody stays sharp a thousand times in a row. At those volumes, a miss is inevitable.

CHAPTER 10

The first incident in policy research is a matter of time

The KPMG case from the introduction is the exact mirror of this work: a consulting firm that put its own name under fabricated sources. A comparable case at a Dutch knowledge organization in policy research is not yet known. That does not weaken the argument; it strengthens it.

In the Netherlands, quality standards almost always arrive after an incident. The Stapel affair led, years later, to the research integrity code. Open data became the norm once research funder NWO made a data management plan a condition for grants. The pattern repeats, and the first public case in this corner is a matter of time.

Two consultants at Necker, Lucyenne van de Coevering and Marieke Oprel, laid out what is at stake in May 2026 in the trade journal Binnenlands Bestuur. Democracy runs on accountability: for every decision, you must be able to trace what it rests on. Policy research supplies precisely the reports and analyses on which administrators and politicians base their decisions. So it does not sit at the edge of that process, but at its core. VERIFIED

The moment AI co-writes those documents without it being traceable what a conclusion rests on, the chain of accountability breaks at exactly that point. And a language model cannot take over that accountability: it is not elected, it cannot be called to account, and it carries no political responsibility. That stays with the human.

So the responsibility stays where it always was. Only the proof of that responsibility has become harder.

THE ANSWER WITHIN ONE MINUTE

The client calls and asks what a particular recommendation rests on.

The employee opens not the archive but the report itself, in the digital environment it was made in. Next to the recommendation sit exactly the findings it is based on. Next to each finding sit the passages it rests on: fragments from interviews and documents, each with a reference to the full source. Where AI was used, it can be pulled up what text went in, what came out, and what the researcher kept or changed. And on the whole it says who reviewed and approved which version, with a date.

The answer to the client is then not a reconstruction but a printout: this recommendation rests on these three findings, and those rest on this material. Time between question and answer: under a minute. No digging through the archive, no calling the colleague who left, and no need to know a single sentence by heart.

III PART III The approach

CHAPTER 11

Sort first, then choose

Most organizations start with the wrong question. They want to know which AI tool is safe. The usable question is: **which documents may be processed with AI at all?**

Because the moment AI enters the picture, information splits into two kinds that call for opposite treatment: documents with personal data, and knowledge without names. Almost all the worry belongs to the first. Almost all the safe gain sits in the second.

THE SORT, IN TWO GROUPS

Per document: does it contain a name, or anything that could identify a person?

1 · PERSONAL DATA

Files, interview data, client documents, personal data from policy documents.

Stays out of AI for now, until the legal basis, the assessment, or the anonymization is in order.

2 · KNOWLEDGE WITHOUT NAMES

Methodology, standards frameworks, published policy, protocols, your own reports without personal data.

This is where the safe AI gain sits. Start here.

When in doubt about a document, it is always group 1. The doubt is the answer.

Treat these two as one whole and you are forced to apply the strictest measures across the board. The outcome is usually that practically nothing is allowed anymore, or that a system arrives that is too big and complex for anyone to maintain.

Sort first, and you set up the part without personal data properly and postpone the sensitive part until the legal basis is in order. That sorting takes half an hour and determines the entire approach from there.

CHAPTER 12

Three measures that cost no money

Before the question of which system to build or buy, there are three things that can happen today, cost nothing, and require no purchase. Together they remove more real risk than any tool choice.

01

Organization accounts

Instead of private accounts. A private account runs on consumer terms: the text entered can often be used for training, and nobody sees what is being processed. An organization account turns that around: business terms, a data processing agreement where needed, and visibility into the use.

02

Two-factor authentication

On every AI account. Such an account stores the full conversation history, so every transcript and policy document that ever went in. One leaked password opens that entire archive; this is the cheapest way to keep that door shut.

03

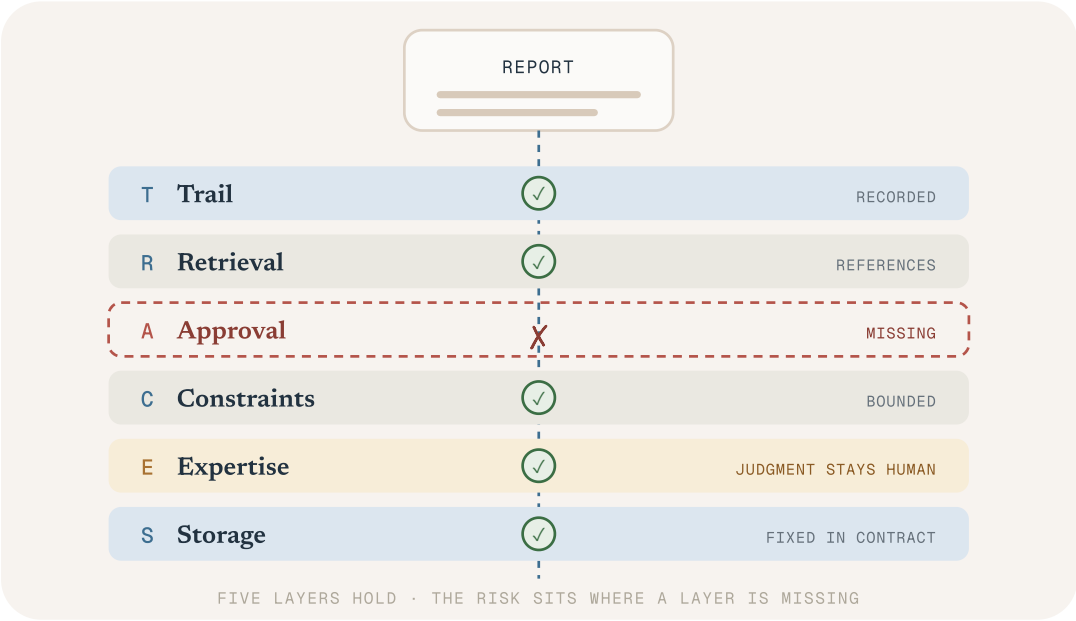
Agreements on one page

Not a policy memo, but a single page with three answers: which data never goes into an AI service, who signs off before a report goes to the client, and who an employee reports an error or a doubt to.

A WARNING THAT RUNS AGAINST INTUITION

When hidden AI use is suspected, attention quickly goes to the juniors. But it sits just as much with the senior researchers and the leadership, and often more so, because that is where the freedom to ignore agreements lives. So look at every layer.

TRACES · THE CROSS-SECTION



Test a system layer by layer. Nearly every widely discussed AI incident traces back to the layer that was not there.

CHAPTER 13

TRACES: the six layers to test a system against

SEAL shows which phase leaks. TRACES is how you test whether a system can plug that leak.

So suppose an organization has something built, or buys something. How do you establish that it is sound?

A system that holds on to origins is more than a model with a text box around it. It consists of six layers, and an organization tests with the vendor whether each layer is there. If a layer is missing, a risk sits there, and nearly every widely discussed AI incident traces back to one or more missing layers.

> **TRACES** are the six layers on which you test an AI system: Trail, Retrieval, Approval, Constraints, Expertise, Storage. If a layer is missing, a risk sits there.

THE VENDOR TEST · SIX LAYERS ON ONE CARD

THE LAYER	THE QUESTION YOU ASK	DISQUALIFYING ANSWER
T **Trail**	Produce the log of a random finding from last month.	Someone had to fill something in for it. Then in three months it stops happening.
R **Retrieval**	Show a claim with its source reference, and what happens on a question without a legitimate source.	Something fluent comes out anyway.
A **Approval**	What do the approval steps look like, and on what share does the reviewer still change something?	That number sits at zero: things get checked off, not checked.
C **Constraints**	Which actions does the system carry out on its own, and which sit technically outside the model?	The model gauges that itself.
E **Expertise**	Which tasks stay emphatically with people, and how does the design keep that role from hollowing out?	The vendor does not understand the question.
S **Storage**	The data processing agreement, the server location, the retention periods, and in black and white: no training on your data.	A reference to the general terms and conditions.

Six layers, six questions, and per question the answer that ends the exercise. The details per layer follow below.

T Trail

When something goes wrong, and somewhere something always goes wrong, the first question is always the same: what exactly happened?

The Trail layer records, for every finding, which source was used, what the conclusion was, who approved it, and when. That is not there to police employees, but to account for the work.

The pitfall is a logbook someone has to keep by hand. Under time pressure, every extra action is the first thing to go, so such a logbook stops being filled in after two weeks. A good Trail layer therefore asks for nothing extra: it writes along and records as a byproduct of the work, and every action in the system leaves a trace by itself.

Without this layer, every serious incident turns from a repairable error into a loss of trust. A client or regulator then cannot see whether it is one case or a pattern, and whether the rest of the work holds up.

THE QUESTION FOR THE VENDOR

Produce the log of a random finding from last month. And: did someone have to fill something in for it? If so, in three months it stops happening.

R Retrieval

A language model always gives an answer, even when it lacks the necessary knowledge. It then invents something that fits in form, logic, and tone, and is hard to catch on errors for exactly that reason. The Retrieval layer turns that around:

- the system may only answer from a bounded source collection managed by the organization;
- every answer points to where it came from; and when the source is not there, the system says so instead of inventing something.

There is a second property, less known, that matters in research. Search always yields something: it returns the best-fitting hits, even when nothing truly relevant exists. What it cannot do is demonstrate that something is missing. Which statement has no source, which recommendation leads back to nothing: those are negative questions, and a system that only searches never answers them. That requires recording which finding hangs on which source, as a stored connection and not as a search query.

THE QUESTION FOR THE VENDOR

Show a claim with its source reference, and what it shows on a question without a legitimate source. Does it say honestly that the answer is not there, or does something fluent come out anyway? The second is disqualifying, however good the rest is. And after that: does the system show which statements in this report have no source?

A Approval

Everything that goes to the client should first be confirmed by a person. That sounds self-evident, but the Approval layer is subtler than a click on a confirm button.

An approval that consists of one click waters down into box-ticking within weeks. The way out is not more eyes but fewer checkpoints. Adding a source is simple and waits on nothing. Whoever changes a statement gets an automatic check: does the new phrasing contradict something elsewhere in the report? Whoever changes a conclusion or recommendation cannot proceed without the reviewer's sign-off. The editor-in-chief with the fabricated quotes had no ill will. The check had become a routine.

THE QUESTION FOR THE VENDOR

What do the approval steps look like, what does the user see at that moment, and which changes get approved automatically? And one number from the system: on what share of the approvals does the reviewer still change something? Whoever truly reviews finds something now and then. If that number sits at zero, there is no checking going on, only checking off.

C Constraints

A language model can do one thing: produce text. It can write a sentence that sounds as if calculation happened, without any calculation happening, or report that all names have been removed from a document, without every name being gone. From the sentence itself, there is no telling. So these actions, calculating, anonymizing, publishing, should not hang on the model's own estimate of itself.

The Constraints layer places sensitive actions outside the model, in fixed rules. A model that decides on its own to anonymize a quote, or recalculates a sample percentage without a calculation module, or writes a draft to a shared folder, does something nobody can verify afterward. The difference from checking afterward is essential. A checkpoint can be ignored. A boundary that is technically fixed cannot. That the system cannot do something gives more certainty than any after-the-fact check.

THE QUESTION FOR THE VENDOR

Which actions does the system carry out on its own, and which sit technically outside the model?

E Expertise

This layer is not technical but a design choice, and perhaps the most important of the six. A system can be set up so that it takes over the routine work and leaves the judgment with the user, and sharpens that judgment or not. It can also be set up so that it quietly takes over the

critical thinking.

What that means differs per role.

- A **junior** who never builds a first analysis alone anymore learns the trade differently, and not necessarily better.
- A **project lead** who leaves the synthesis to a model stops building their own judgment from that point on.
- A **researcher** who no longer reads critically loses touch with their own expertise.

This layer, moreover, can never be fully automated. A system can establish that a source covers a statement. A system cannot establish that a source technically covers the statement but puts the reader on the wrong track, because the quoting was selective or the coverage stretched. That takes professional knowledge. The work does get considerably more manageable: the system shows, next to every statement, the passage it rests on, so the reviewer lays statement and passage side by side instead of working through the whole source.

THE QUESTION FOR THE VENDOR

Which tasks does the system take over, and which tasks stay emphatically with people? And how does the configuration or architecture prevent the human role from slowly hollowing out? A vendor who does not understand this question has not thought about it, and that is an answer in itself.

S Storage

This layer is about where information goes and what happens to it. Where does the data sit, in which country, with which party? Does the vendor, or the vendor's vendor, train on what users enter? What happens to the data when the contract ends? And are the answers to those questions fixed in the contract, or do they rest on a promise made in a sales call?

For an organization working with interview data, for instance, this is not a matter of preference: without a data processing agreement, the processing is not legally valid, however well everything else is arranged.

THE QUESTION FOR THE VENDOR

The data processing agreement itself, the location of the physical servers, the retention periods, and in black and white that there is no training on the organization's data. Whoever stays vague here, or points to general terms that do not properly cover this, is out.

CHAPTER 14

Consistency is measurable

For a field where reproducibility is a quality requirement, the strongest piece of equipment sits here. A system is not stable because the vendor says so, but because a measurement shows it. Ask the same question ten times and count how often the same answer appears with the same source reference. Nine or ten out of ten is stable; six out of ten means the answer partly depends on chance. That number belongs in the methods section, next to the response rate: a quality figure that tells the reader how reliable the tool was that the research was made with.

A system that gives three different recommendations to the same question does not serve as substantiation, however good each separate answer looks. A system that points to the same passage ten times out of ten does.

THE TEST, STEP BY STEP

1. Pick one question the system must handle in the real work, for example: what does this recommendation rest on? Keep the exact wording.

2. Ask precisely that question ten times, each in a fresh conversation, so the system does not build on the previous answer.

3. Note two things per round: the answer in one sentence, and the source the system points to.

4. Count. The same answer with the same source reference counts as a hit. A different answer, or the same answer with a different source, does not.

5. Write the result as a fraction: eight out of ten. That number goes into the methods section. Half an hour of work.

Counting is not judgment: whoever repeats the test lands on the same number. The vendor is not needed for it: anyone can run the test today. And the test measures exactly what the client silently assumes: that the research gives the same answer tomorrow as it does today.

THE MEASUREMENT METHOD

Five real research questions. Each question ten times, in a clean conversation. Note how often the finding and the source both match.

Below eight out of ten, we report a warning. That is not an existing standard but our own choice, based on the idea that one in five diverging answers in a report cannot be explained away. A different threshold is fine, as long as it is fixed in advance and not after the fact.

CHAPTER 15

Measure first, then choose

Before the question of buying or building anything can be asked, an uncomfortable fact has to be faced. And it shapes how most organizations decide whether AI works or not: by feel.

In a controlled study of experienced developers (METR, 2025), the participants expected AI to make them about a quarter faster. Afterward, they reported that it had indeed felt faster. The measurement, with the clock running, showed the opposite: they were almost 20 percent slower with AI. And even after the fact, they kept believing it went faster. The gap between assumed and measured productivity came to about 40 percentage points. VERIFIED

MEASUREMENT CONTEXT

This involved 16 experienced developers in codebases they had written themselves, with the AI tools of early 2025. The researchers themselves warn against broad generalization, but the mechanism is not bound to the trade.

In research, the odds of this effect are greater rather than smaller. With AI, something is there right away: a draft text, a summary, an analysis. What nobody registers is the time that leaks away afterward in checking, rewriting, and verifying sources, and precisely that follow-up work is the biggest part.

A researcher rewriting an AI draft three times, a project lead walking through every finding, a senior editor checking the sources: they all experience the speed of the first version and forget the hours after it.

That makes self-reporting an unreliable measuring instrument. “Our team experiences AI as a big time-saver” is not evidence. Whoever really wants to know whether AI helps measures before the rollout and measures the same thing again after. A week of tallying beats any survey.

CHAPTER 16

Buy, build, or neither

Does a knowledge organization buy a ready-made tool, or have something built?

The honest answer: buying usually wins. For routine work that is the same everywhere, transcription and translation, a good existing tool is almost always cheaper and faster than a custom build.

But the moment source requirements, confidential material, and the demand for traceable work come together, the standard tools fall short. And in policy research, that is not the exception: every report that goes to the client has to be right and has to be defensible. That is the work.

There, a different form wins: the organization buys in the AI capacity and has the layer around it made to measure, the search across its own sources, the source referencing, the approval flow. That layer is where it counts, and exactly what a generic tool does not deliver.

And then the third outcome, every bit as serious as the first two: sometimes the answer is neither. A simple script without AI, a better process, or nothing at all. If a thorough analysis shows that AI adds nothing, that is an answer as valuable as building.

What all three routes share is the order: **first the analysis, then the choice.** Not the other way around. Buying or building something and then looking for what it is for is exactly where most failed AI projects begin.

CHAPTER 17

The question your client is going to ask

Everything in this document runs toward one moment: the client, the committee, or the regulator asks what a recommendation rests on. Whoever withstands that moment is selling not accuracy but something more valuable: a research lead who stands behind the work in a room where nobody has recomputed the report themselves.

What an organization must show at that moment is precisely what the six layers deliver: the source next to the finding, the bridge to the recommendation, and the log of who approved what, and when. That setup comes before the question, because after the fact, nobody reconstructs the origin.

> The client is not paying for accuracy, but for the certainty that every conclusion can be followed back without having to check it themselves.

That is what traceability buys an organization: not another quality mark, but the difference between a report that survives a room of critical professionals and a report that cannot withstand one.

And there is a second side, one that rarely gets named in these discussions. The same setup that can handle the client's question makes room in the work itself: what demonstrably rests on material does not need to be checked three times out of caution and gets to move faster. The risk and the gain come from the same agreement.

CLOSING

In closing

CLOSING

What this white paper does not do

This document does not give legal advice. Whether a particular data flow is permitted, and whether a data protection impact assessment (DPIA) is required, is a question for the data protection officer or a lawyer. What this document does is supply the questions that make that conversation worthwhile.

Nor does it promise that the risk disappears. It becomes manageable, not zero. Even with everything in order, risks remain, and that calls for an agreement, rather than the illusion that it has all been sealed shut.

And it sells nothing: anyone who wants to take this further on their own has everything they need. The line we draw out loud: systems that assess, select, or admit people fall under the heaviest regime and belong with specialist parties, not with us.

VERGELABS

What we do

This document sells nothing: anyone who wants to take it further alone has everything they need. And for those who would rather take it on together, VergeLabs is here.

We work in one order: first look at where an organization is really losing time or money, and only then build or advise. Sometimes the outcome is that a smaller intervention is enough, or that AI adds nothing. We say that too, because a good answer is worth more than a sold project.

The same agreement that removes the risk unlocks the gain. After such a project, you can show what every judgment rests on, and what gets to move faster because of it.

01 · Advice and analysis

We map out what AI may be used for and where it pays off, and which documents are better kept out of an AI service.

02 · Training and AI literacy

We teach employees how the systems work and what safe use means. That also covers the statutory literacy obligation.

03 · Building systems that hold on to the source

We build systems that show what a finding rests on and leave the decision with the researcher. The six TRACES layers are the hard condition.

04 · Setting it up on the work floor

We arrange the agreements, the business accounts, and the checks, so AI use becomes governed and visible instead of a shadow practice.

The first step is usually just a good conversation.

One line back is enough: hallo@vergelabs.nl.
Or book an introduction at vergelabs.nl.

SCAN · VERGELABS.NL

Justification, claim by claim

Every factual claim carries a label: VERIFIED PLAUSIBLE DISPUTED. Where a figure appears, the measurement context appears with it. The labels are assigned conservatively: in case of doubt, the lower label. Where this document makes a choice that is not a standard, it says so, among others at the consistency threshold

in chapter 14. Five claims were re-checked against the primary source as a sample on August 17, 2026; those carry a verification date below.

AI use among employees, ~45 percent · LayerX, Enterprise AI & SaaS Data Security Report 2025 · browser telemetry, not a survey; the 77 percent is the share of users pasting company data into a prompt, not an adoption figure · vendor of browser security PLAUSIBLE

Same order of magnitude, independent measurement · Verizon, Data Breach Investigations Report 2026 · more than 850,000 logged events, three times as much as a year earlier PLAUSIBLE

>90 percent expect more AI use, <30 percent have agreements; policy mostly privacy (53 percent) and rules of use (51 percent) · Berenschot and Waag Futurelab, AI-Trendonderzoek 2025 · n≈500, Dutch organizations broadly, not sector-specific · re-checked 8-17-2026; the 2026 edition appears mid-September VERIFIED

Client pressure on phrasings · Vasco Lub, Sociale Vraagstukken, 2011 · essay by a peer, not a measurement PLAUSIBLE

758 consultants: ±40 percent higher quality and a quarter faster inside the boundary, 19 percentage points lower outside it · Dell'Acqua et al., *Organization Science*, 2026 (vol. 37, no. 2; earlier Harvard Business School working paper, 2023) · randomized field experiment · re-checked 8-17-2026 VERIFIED

44 physicians, 20 hours of training: 84.9 to 73.3 percent with flawed advice, difference 14.0 percentage points after adjustment · NEJM AI, 2025 · randomized trial; the errors were deliberately injected, the study measures automation bias VERIFIED

Developers ~19 percent slower with AI, while believing they were faster · METR, 2025 · randomized, 16 experienced developers in their own codebases; the researchers themselves warn against broad generalization VERIFIED

NeurIPS: 100 confirmed fabricated references in 51 to 53 papers · GPTZero, report January 21, 2026 · 4,841 of 5,290 accepted papers checked; each paper had three to five reviewers · primary source is public, re-checked 8-17-2026 VERIFIED

Deloitte Australia refunded part of a government contract · fall 2025, report for the department of employment; fabricated academic sources and a fabricated quote from a court ruling · international press coverage, incl. Financial Times and AFR VERIFIED

KPMG withdrew AI advisory report; 5 of 45 source references held up · GPTZero analysis, June 2026, of the report "Redefining excellence in the age of agentic AI" (October 2025) · Financial Times; also The Register, TechCrunch, and accountant.nl · re-checked 8-17-2026 VERIFIED

EY Canada (May 2026) and PwC Middle East (four reports, late July 2026) · press coverage, PwC via GPTZero in the Financial Times · not yet laid next to the primary articles PLAUSIBLE

Three lawyers received a formal warning from the deans; two were required to complete an AI course · NOS, February 22, 2026 · it stayed at the mildest form of oversight, no disciplinary rulings PLAUSIBLE

Former editor-in-chief of NRC and De Standaard suspended: fabricated quotes in 15 of 53 blog posts, seven of those quoted deny them · NRC, March 19, 2026; NOS, March 20, 2026 · he himself confirmed the quotes arose while summarizing reports with AI · re-checked 8-17-2026 VERIFIED

Accountability as the core of policy research · Van de Coevering and Oprel (Necker), Binnenlands Bestuur, May 2026 · interpretation by peers, not a measurement VERIFIED

The legal calendar · Regulation (EU) 2024/1689 and the Digital Omnibus (adopted June 29, 2026) · literacy since February 2, 2025; transparency and enforcement since August 2, 2026; watermarking transition until December 2, 2026; high-risk postponed to December 2, 2027 and August 2, 2028 VERIFIED

Frameworks consulted without figure claims · ISO 20252 · Code of Conduct for Research and Statistics (MOA, VBO, and VSO) · Periodic Evaluation Regulation · Policy Evaluation Toolbox · Netherlands Code of Conduct for Research Integrity, revision around AI expected in 2026.

Appendix · what you can do today

Without us, without budget, and without a tool. Whoever carries out only this appendix stands further than most organizations.

1 · Sort into two groups: half an hour

Go through your material: does it contain a name or an identifiable person? If yes, or in doubt: group 1, out of AI for now. If no: group 2, this is where your safe gain sits.

2 · Three free measures: an afternoon

Organization accounts instead of private ones, two-factor authentication on every AI account, and one page of agreements: what never goes into AI, who signs off, where you report an error.

3 · Ask the six TRACES questions

Take them to the next vendor meeting: Trail, Retrieval, Approval, Constraints, Expertise, Storage. If the vendor works with a demo, ask them there too.

4 · Measure a week: before forming a view

Tally the time of a task with and without AI for a week, before anyone forms an opinion. Feel is not evidence; a measured week is.