AgenticObjects

WHITE PAPER

The Science Analytics Never Shipped

An AI agent can generate findings continuously, at a cost approaching zero. The person receiving them has the same working memory they had in 1984 — the last year a finding from cognitive science made it into an analytics product. This paper is about the research that piled up in between.

Ali Yıldırım

Founder and CEO, D-CAT Group

Twenty-five years in enterprise business intelligence

Computer Engineering, Boğaziçi University · MBA, Istanbul Bilgi University

Version 1.0 · September 2026

Citations verified by two independent passes and one adversarial search for refuting evidence; see §10.

Why I wrote this

I have spent twenty-five years inside business intelligence projects, and the most useful thing I have learned is that building the object correctly is rarely the first problem.

Reports, what-if models, dashboards, balanced scorecards. Organisations become very good at producing these. What they do not reliably produce is anyone consuming them. We have had clients running three and a half thousand reports. The interesting question was never whether those reports were correct. It was whether anyone opened them, and whether anything changed when they did. Our consumption audits at that kind of scale kept coming back low. Low enough that the report count told you almost nothing about how much of the business was actually being read.

For most of those years I filed this under culture. Collaboration tools nobody collaborates in. Dashboards nobody is held to. Culture is real, and I no longer think it is the whole story, because you cannot will a person into consuming three and a half thousand reports. Some of what we have been calling a culture problem is a design problem wearing a culture problem's clothes.

This paper is about that part. The cost of producing analytical output has just collapsed, and the mechanism that consumes it has not changed at all. There is a body of research on that mechanism — how much a person can hold, what makes them accept a machine's claim, what it costs them to check it. It has been published for decades, and almost none of it has reached a product. So I went and read it, and this is what it says.

1. The gap that is not ignorance

Business intelligence has a scientific foundation. This needs saying clearly, because the argument here is not that the industry ignored science. It did not. It imported science at one layer thoroughly, and at the layers above it hardly at all — and the reason turns out not to be the one I assumed for most of my career.

In 1984, William Cleveland and Robert McGill proposed a ranking of the elementary perceptual tasks people perform when reading quantitative information off a graph — judging position, length, angle, area, curvature, shading. The ranking itself was a hypothesis, assembled from prior evidence; two experiments then tested parts of it. Position along a common scale came out most accurate: more accurate than length by factors between 1.4 and 2.5, and 1.96 times as accurate as angle. Area sits low in the ordering but was not measured in either experiment, and length and angle share a rank there, a tie the authors flag as unresolved.

That is a narrower result than the folklore it produced, and the folklore is worth noting too. "Prefer bars to pies" descends from this paper — Cleveland and McGill do write that a pie chart can always be replaced by a bar chart, swapping angle judgments for position judgments. But their actual conclusion goes further than the

advice the industry kept: neither form should be used, they argue, because better methods exist, and they offer dot charts instead. The industry absorbed half a recommendation and built a generation of defaults on it. Practitioner writers, Tufte and Few among them, carried the same tradition into working advice. Ask a BI vendor why their defaults look the way they do and the honest answer traces back here.

The usual next line is that this was the last question the industry asked cognitive science. I believed that version for most of twenty-five years. It is wrong, and it is worth being precise about why, because the true version is a harder problem.

The research did not stop. It moved into fields the analytics industry treats as adjacent rather than as its own. In 1993 Russell, Stefik, Pirolli and Card published a paper titled "The cost structure of sensemaking". In 1996 Shneiderman gave interface design an explicit rule for staging information — overview first, zoom and filter, then details on demand. In 1999 Pirolli and Card modelled information seeking itself as an economic calculation, with "information scent" as the cue by which a person estimates, in advance, whether pursuing something is worth the effort. In 2010 Heer and Bostock re-ran Cleveland and McGill's experiments on crowdsourced participants. In 2024 Burnay, Lega and Bouraga applied cognitive load theory directly to dashboard adoption. And some of this work happens inside vendors rather than outside them: Tableau maintains a research group whose members list perception, cognition and graphical perception among their stated areas.

So the gap is not ignorance. One answer crossed into the product and the rest did not.

Cleveland and McGill's ordering is in the defaults of every BI tool on the market. Shneiderman's staging rule is a design pattern — taught, cited, widely admired, and not a constraint any product is held to. No procurement checklist I have seen asks what it costs a reader to trace a number to its source. No vendor reports a working-memory budget. The perceptual layer got a science that became a default. The layers above it — attention, memory, trust — got a science that stayed in the literature, and defaults that came from vendor convention and dashboard fashion instead.

Knowledge that exists and does not propagate is a question about incentives rather than about curiosity. That is where §9 ends up, and it is the reason this paper is addressed to buyers rather than to designers.

For forty years the failure to propagate did not cost much, because the layers above perception were never the constraint. A human analyst produced a handful of findings a week, and a handful is consumable by default. The gap in the science was real, but nothing was flowing through it.

That condition has ended.

2. The capacity that does not scale

The cost of producing analytical output has collapsed. Systems built on large language models generate findings, comparisons, anomaly reports and

recommendations continuously, at a marginal cost approaching zero. Whatever one thinks of the quality, the volume is not in dispute. The production side has been rebuilt.

The consumption side has not changed, because it cannot. It runs on human working memory, and that capacity is a research finding rather than a product parameter.

Miller's 1956 paper on the magical number seven framed the question for half a century, though it is worth knowing how Miller's own successor reads it: Cowan notes the number "was meant more as a rough estimate and a rhetorical device than as a real capacity limit." Cowan's 2001 review gathered the evidence and put the figure lower. His unit is the chunk: a group of items the mind holds as one thing, so that a familiar phone number costs less than the same digits in random order. Under conditions that stop a reader bundling items into larger chunks, and stop repetition doing the work of memory — the conditions under which a capacity limit can be observed at all — a single central limit averaging about four chunks is implicated. Estimates in the literature range from three to five, and what counts as one chunk changes the answer, so the qualifier matters and stays. But no serious estimate lands anywhere near the volume an agent can produce before a coffee refill.

Cognitive load theory names the mechanism underneath. Sweller's 1988 work showed that when an activity consumes a large share of available processing capacity, that capacity is unavailable for anything else happening at the same time — in his case, the schema-building that constitutes learning. He was studying problem solving, not dashboards, and he offers no threshold and no capacity figure. The transferable part is the shape of the constraint: processing capacity spent on one thing is not spent on another, and it does not expand under load.

Herbert Simon saw the structural version of this in 1971, before either BI or AI existed in their current form. He wrote that a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of sources competing to consume it. My own extension of his argument, and it is an extension: a system designed under information scarcity will misallocate attention under information abundance, because its defaults encode an assumption that no longer holds.

Put these together and the arithmetic of the current moment gets uncomfortable. An agent that surfaces forty findings has not produced ten times the value of one that surfaces four. If capacity is around four and nothing in the delivery mechanism respects that, the value of the forty is not ten times smaller. My own conclusion, and it is an argument rather than a measured result, is that it approaches zero: the reader either abandons the list or samples it arbitrarily, and neither behaviour is a reading strategy. Cowan gives the capacity. What follows from it here is mine.

The obvious objection is that a screen is external memory. Nobody holds forty findings in mind; they are read one at a time and the display does the holding. That is correct, and it is why the capacity limit does not bite during reading. It bites at the moment of deciding, when several findings have to be held against one another,

weighed, and resolved into one choice. Reading is cheap. Comparison is what meets the limit, and comparison is what a decision is made of.

The bottleneck of enterprise analytics has moved. It no longer sits in the warehouse, the model, or the query engine. It sits between the screen and the reader, in the one component that does not ship updates.

This is the layer the industry never built a science for. The rest of this paper is about what that science says, because it exists, it has existed for decades in cognitive psychology and human-computer interaction, and it is remarkably specific about what works and what it costs.

3. The obvious fix, and what the data did to it

Confronted with the trust problem, the industry's instinct has been consistent: if people should not blindly accept a machine's answer, show them the reasoning. Attach the explanation. Let the user verify. It sounds like basic intellectual hygiene — show your working, let the reader judge it — and it has become a default requirement in enterprise AI procurement.

The most direct experimental test of that instinct I have found came from Bansal and colleagues at CHI 2021. They measured whether AI explanations improve the performance of the human-AI team — not the AI alone, but the combined system whose output is the human's final decision. Their result, from the abstract:

> "*While we observed complementary improvements from AI augmentation, they were not increased by explanations. Rather, explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness.*"

Explanations did not make people better at catching the AI's mistakes. They made people more likely to accept whatever the AI said, including when it was wrong. The explanation functioned as a persuasion device rather than a verification aid. A related detail from the same study: displaying a simple confidence score performed well enough that the authors observed no significant improvement from adding explanations on top of it. That is a null result, not a demonstration of equivalence, but it is an expensive null result for anyone building explanation infrastructure.

The boundary of the finding matters. Bansal did not show that explanations are harmful, and I will not claim it. What the study showed is narrower and, for anyone designing analytical systems, more unsettling: the presence of an explanation does not by itself produce scrutiny — the reader stopping to ask whether the claim is actually right. Whatever verification is supposed to happen between the machine's claim and the human's decision does not happen merely because the reasoning is visible.

For enterprise analytics the implication is direct. An agent that delivers every finding wrapped in a fluent rationale is not thereby a transparent system. It is a system that has raised its own acceptance rate.

4. What actually works, and what it costs

If showing reasoning does not produce scrutiny, what does?

Buçinca, Malaya and Gajos tested a family of interventions they call cognitive forcing functions: design elements that compel engagement before acceptance. Three of them, from the paper. *On demand:* the AI's suggestion is hidden until the reader asks for it. *Update:* the reader commits to their own answer first, then sees the AI's and may revise. *Wait:* the recommendation is withheld for thirty seconds while the interface says the AI is still working.

In an experiment with 199 participants, the result cut both ways:

> "*[C]ognitive forcing significantly reduced overreliance compared to the simple explainable AI approaches. However, there was a trade-off: people assigned the least favorable subjective ratings to the designs that reduced the overreliance the most.*"

Stated plainly: the designs that most reduced overreliance — acceptance of an AI suggestion that was wrong — were the designs users rated worst. This is not a failure of the two measures to correlate. In this study they moved in opposite directions.

One precision is worth keeping, because it is easy to lose and I have lost it myself in earlier drafts. What Buçinca measured is overreliance, not decision quality in the round. Overreliance is one component of a good decision and a particularly diagnostic one, but the study does not report that the cognitive forcing conditions produced the best decisions overall. The finding is sharp enough without the upgrade.

Consider what even the narrow version does to the standard playbook of product development. Analytics vendors optimise for what users report back: adoption, NPS, delight. I have not met one that measures anything else about how its output is received. In the specific domain of AI-assisted decisions, a vendor optimising for what users prefer is, in expectation, optimising for acceptance. The market's selection pressure points at the wrong target, and nobody has to act in bad faith for it to happen. The feedback loop does it unassisted.

A further detail resists any one-size-fits-all reading. The authors audited their own work for intervention-generated inequality and found that, on average, the interventions benefited participants higher in Need for Cognition — the disposition to enjoy effortful thinking. The benefit of forcing scrutiny is not distributed equally across readers, which is a different and more precise statement than saying its cost falls unevenly: what the study measured was unequal benefit, and it reports no significant difference in the subjective ratings given by low Need for Cognition

participants. That will matter in §9, because it moves the problem out of interface design and into organisational decisions about who reviews what.

5. The variable underneath

At this point the literature looks like it contradicts itself. Bansal found explanations did not improve team performance and did increase acceptance. A later CSCW study by Vasconcelos and colleagues found explanations can reduce overreliance. Both are careful experiments. The reconciliation is the most useful idea in this body of research.

Vasconcelos and colleagues model the reader as making a strategic choice, formalised in a cost-benefit framework: the costs and benefits of engaging with the task, weighed against the costs and benefits of relying on the AI. Across five studies they moved overreliance by moving the terms. Making the task harder to check raised reliance. Making the explanation harder to follow raised it. Paying participants more for a correct answer lowered it — and that last one is a benefit, not a cost, which is the part of their framework most easily lost in summary.

So the reconciliation is not that explanations work when they are cheap to check. It is that the reader is running a calculation, and an explanation is unlikely to change behaviour unless it changes a term in that calculation. The authors put their own version of this as a suggestion rather than a law, and I will keep it at that strength. Where earlier studies found nothing, the likely reason — and the authors put it hedged, as "could be due in part" — is that those explanations did not sufficiently reduce the cost of verifying the prediction. Adding prose to a claim is not the same as making the claim checkable.

The presence of an explanation was never the variable. Verification cost is the most actionable term in it, and the one nobody measures.

That reframing moves the problem from philosophy to engineering. "Do users trust this system appropriately" is not measurable at design time. "How many seconds and how many steps does it take to trace this number to its source" is.

This idea has a lineage, and pretending otherwise would be both dishonest and unwise. Pirolli and Card modelled information seeking as an economic calculation in 1999: a person pursues a piece of information when the expected value justifies the cost of getting it, and "information scent" is how that estimate is made before the cost is paid. Russell, Stefik, Pirolli and Card had written about the cost structure of sensemaking six years earlier. Treating the reader as an economic agent is neither my idea nor Vasconcelos's.

What is different here is narrower, and as far as I can find it is still unclaimed. Foraging is about the cost of *finding*. This is about the cost of *checking* something already handed to you — a different transaction, and the one that matters when a machine returns an answer rather than a set of search results. And the step I have not

found taken anywhere is the one after the theory: turning that cost into a number that a product reports, budgets against, and is held to.

Making verification cost a number

If verification cost is the term worth managing, someone has to own it as a number. Query latency became manageable the moment it stopped being a complaint and became a percentile on a dashboard. The same move is available here, and four things can be measured before a single reader sees the system:

	What it counts	Where the number comes from
Steps to source	How many actions a reader takes to get from the claim on screen to the record behind it	Count the clicks. This can be done on the interaction design, before anything is built
Trips off the surface	How many times the reader has to leave — into the warehouse, the ERP, a spreadsheet, someone's inbox	Navigation logs. Leaving is the expensive one, because coming back costs attention too
Time to first evidence	Seconds from deciding to check until the first underlying record is visible	The same timing instrumentation that already measures query latency
Rebuild burden	Whether the calculation is shown, reachable in one step, or has to be reconstructed by the reader	Assigned per claim type at design time; three levels are enough

Two things about the measurement matter more than the four components.

Report the spread, not the average. Verification cost fails the way latency fails: the average is reassuring, and the damage is done by the small number of worst cases the average hides. A system whose typical claim traces in four seconds, and whose worst claims need the source system reopened, does not have low verification cost. It has low verification cost on the claims nobody needed to check.

Set the ceiling per claim, not per system. A routine awareness item and a claim that will move a budget do not deserve the same verification budget. Tie the ceiling to a materiality threshold — the instrument financial controls have used for decades — and the design question becomes answerable: above this threshold, a claim ships only if it can be traced in this many steps and this many seconds. Below it, the stream runs.

There is a second number, and it is the one that tells you whether the first one is working. Of the claims a reader acted on, what share did they trace to source first? Call it the verification rate. Cost is the lever; rate is the reading on the dial. A system that drives cost down and sees no movement in rate has learned something worth knowing — that cost was not the binding constraint for those readers, and the engagement has to be bought some other way. This is Vasconcelos's point from the other side. Their fourth study moved behaviour by changing what verification was worth, not what it cost. Price is one term in the calculation, not the whole of it.

I have not validated this scheme, and I want to be exact about its status: it is an engineering proposal, not a research finding. It did not come from nowhere, though.

Four failures recur in the report estates I have worked in — the accumulated portfolios of reports an organisation ends up running — and each one is a verification cost that nobody ever counted.

Someone asks in a meeting where a number came from, and no one in the room can answer, because the answer lives with the analyst who built the report and that analyst has left. The definition of an operational term drifts somewhere inside a transformation layer and goes unnoticed for months, because noticing it would mean leaving the dashboard. A drill-through exists, technically, and deposits the reader on a forty-column table they cannot read. And the business unit stops trusting the figure altogether and rebuilds it in a spreadsheet of its own, having worked out that constructing a parallel version of the truth is cheaper than checking the official one.

The last of those is worth sitting with, because it is also how a three-and-a-half-thousand-report estate comes into existence in the first place. In every case a competent person declined to check something, and declined for a defensible reason: they had made a rough estimate of what checking would cost them and decided against it. We have built a good deal of the reduction ourselves. In the system we ship now every number carries an address, and every recommendation opens the finding underneath it in a single step. What we have not built — and what I have not seen anyone else report either — is the number. Lowering verification cost and measuring it are different disciplines, and only one of them has been attempted.

These are the four things I would instrument first, and I expect anyone with better data to improve them. A paper arguing that claims should be cheap to check owes its readers proposals that are cheap to falsify.

A claim checkable in five seconds lives in a different trust regime from one that requires reopening the source system, and no amount of explanation prose moves a claim from the second regime into the first.

6. Where attention goes

Once the volume of available findings exceeds what a reader can hold, something decides what gets seen first. There is no neutral option. A chronological list is a ranking policy. An alphabetical one is a ranking policy. Whatever surfaces first will disproportionately shape what the organisation believes about itself, so the question of what deserves attention is worth asking scientifically. The research offers three answers and attaches a warning to each.

Surprise. Itti and Baldi gave surprise a formal definition: the divergence between what an observer believed before seeing the data and after — technically, the Kullback-Leibler distance from prior to posterior. Only observations that substantially shift belief count as surprising, however statistically rare they are. They then tested it, and found Bayesian surprise a strong attractor of attention: 72% of gaze shifts went to locations more surprising than average, rising to 84% for regions all observers selected.

Two caveats travel with that number, and I would rather state them than lean on the finding. The surprise they computed was a low-level sensory quantity, modelled on simulated early visual neurons, and what they measured was eye movement over video. Applying it to the ranking of semantic business findings is an extension — one the authors explicitly license, since they argue the theory holds across scales, modalities and levels of abstraction, but an extension nonetheless, and unvalidated in this domain. The second caveat is a design warning: a ranking that optimises for surprise alone rediscovers the logic of the social media feed, where attention capture is the goal rather than the means. Surprise signals information value only when the expectation it violates was well founded in the first place.

Bad news. The second pull on attention is valence — whether a piece of information is good or bad. Baumeister and colleagues surveyed a wide body of work under the title "Bad is Stronger than Good" and concluded that bad information is processed more thoroughly than good, and that bad feedback has more impact than good. The asymmetry survives the obvious alternative explanations: it is not simply that bad news is more eye-catching, or that it carries more information. It is a review rather than an experiment, and the strength of it is the breadth. Left unconstrained, any attention-ranked stream of business findings will drift negative — that inference is mine, not theirs — and a reader whose entire morning consists of risks and declines is not better informed. They hold a systematically distorted model of the business. Balancing the stream is bias correction, not cosmetic reassurance.

The information gap. Loewenstein reinterpreted curiosity as a form of cognitively induced deprivation arising from the perception of a gap in knowledge. The gap itself he defines by two quantities: what one knows and what one wants to know. Curiosity is not the gap; it is the felt deprivation the gap produces once attention lands on it. Separately, Gruber and Ranganath's PACE framework reviews accumulating evidence that curiosity states are related to modulations in the dopaminergic circuit, and that these modulations affect memory encoding and consolidation — for the target of curiosity and for incidental material encountered alongside it. PACE is a prediction-and-appraisal account rather than a continuation of Loewenstein's information-gap theory; the line I am drawing between them is mine.

The design reading I draw from this is a hypothesis rather than a finding, and I want to label it as one: that a well-posed open question may hold attention in a way a closed statement does not. Neither Loewenstein nor Gruber and Ranganath compares open questions with closed statements, and I have found no study that tests it in an analytical setting. There is a second reading, less obvious. Analytical systems have information gaps too. Enterprise data records what happened with precision and records why it happened almost never. The causes of a traffic decline, a stalled category, a churn spike frequently live outside the data, in what Polanyi called tacit knowledge. His formulation is sharper than the way the term is usually borrowed: we can know more than we can tell. The point is not that the knowledge went undocumented, but that a person often cannot fully articulate it on request. A system that can represent its own ignorance explicitly, as a question addressed to the person most likely to hold the answer, does something no dashboard has done: it treats the

organisation's people as a data source of record, and turns the act of consuming intelligence into an act of collecting it. Whether such questions get asked is a design choice the industry has not made, and whether they would be attended to is an empirical question nobody has answered.

7. The single-format assumption

For forty years the industry has assumed one surface can serve every act of information consumption. The dashboard is expected to support the executive glancing for reassurance, the analyst hunting for a cause, and the manager who needs to know what changed since yesterday, all at once.

We built one of these. It is a hospital efficiency dashboard, twenty-five separate reports on a single surface: financial indicators, case-level clinical indicators, anomaly flags, and comparisons against the same period last year. Every report on it is correct. The data model behind it is sound. I would defend the engineering.

What I would no longer defend is the premise. We designed it for a hospital administrator who would sit down, take in twenty-five reports' worth of indicators, notice which anomalies mattered, hold the historical comparison alongside the current figure, and form a judgement. That administrator does not exist. Nobody's working memory holds that, and nobody's calendar contains the hour it would need. We did not get the product wrong so much as we got the reader wrong, and we got the reader wrong because we never asked what a reader is.

That is not a confession of incompetence, and I am not offering it as one. It is a description of the era. Every dashboard I have built or seen built in twenty-five years was designed under the same unexamined assumption: that the surface's job is to present everything relevant, and the reader's job is to cope. The assumption held for as long as an analyst produced a handful of findings a week, because a handful is consumable whatever the interface does with it. It stops holding the moment the volume of output passes what a person can take in. The first place I had to notice that was in our own work.

I want to be careful about what the literature does and does not underwrite here. The three-way split — monitoring, directed awareness, investigation — is my framework, not a taxonomy I am importing. What the research supports is narrower and still useful. The inference I draw — mine, not Cowan's — is that presenting items one at a time, in a deliberate order, is more likely to keep momentary load inside the limits he describes than a wall of forty tiles is. Cowan measured capacity under controlled conditions; he did not compare presentation formats, and neither has anyone else that I have found. Investigation is question-driven and benefits from the curiosity dynamics above. And if attention is the scarce resource Simon described, then its allocation should have a defined end: a format that can be finished respects the budget, while an infinite one spends it silently.

One caution belongs in the same section, because the sequential format is where an earlier finding bites hardest. The smoother the consumption, the lower the friction of acceptance, and Buçinca's result indicates that low-friction acceptance is where overreliance grows. Her task was image classification rather than an analytics stream, so the transfer is inference rather than evidence — but the direction is clear enough to design against. A guided stream is the right format for awareness and the wrong one for consequential decisions, unless scrutiny is deliberately reintroduced at the items that matter. Formats are not interchangeable, and neither are the failure modes they carry.

The system we build now enforces a ceiling on what any single reader receives from any single run, and holds the overflow rather than discarding it. That is a crude instrument, and I no longer think it is the answer. A ceiling throws away what does not fit — but a business does not stop having more than six things happening because a person can only hold four.

The failure I actually want to design against is a system swarming its reader with individually correct records. Every one of them true, addressed, checkable; the reader still drowning. Correctness at the unit does not compose into usability at the surface.

What follows is intent rather than a report of something built. The direction I think is right is to stop treating the individual finding as the unit of consumption. A run can produce a hundred findings without harm, provided no reader ever meets a hundred findings. Between the findings and the person there should be a second layer — composed records that stand for many — and above those, a small number of views. The view is what a person consumes. The finding is what holds the evidence, one step underneath, unchanged.

Cowan has a word for what this is trying to do. A chunk is a group of items the mind holds as one, and every reader facing a wall of tiles is already chunking — unaided, inconsistently, and without leaving a record of how. The proposal is to move that work into the system, where it can be governed, audited and argued with. It also puts the compression where §8 says narrative belongs: a layer over claims that stay individually inspectable, never the layer where truth is stored.

I am deliberately not publishing our numbers. I do not think anyone knows what they should be, ours will change as we learn, and a figure printed here would be obsolete before it was useful. What I will commit to is the direction of the measurement: a system of this kind should be judged on how few things a person has to hold, not on how many the machine found. Generating more was never the hard part. Deciding what a particular person should receive is, and that decision is an organisational one before it is a design one.

8. The cost of telling a story

"Tell a story with your data" may be the most repeated advice in modern analytics. It is offered as a communication virtue, and it rests on something real: narrative transmits information unusually well. The question the industry has not asked is what narrative does to the reader's judgment while it transmits.

Social psychology has an answer. Green and Brock studied narrative transportation — absorption into a story, involving imagery, affect and attentional focus — and built a scale for it. Transportation increased story-consistent beliefs and favourable evaluations of protagonists. In their second experiment they asked readers to circle "false notes," parts of the story that did not ring true, and highly transported readers circled markedly fewer than less-transported ones. Their reading of it: transported participants appeared more accepting, consistent with the idea that they are less likely to doubt, to question, or to engage in disbelieving processing.

It is worth being precise about what was and was not measured, because the looser version of this finding circulates widely. Green and Brock did not measure counterarguing. They say so: conventional thought-listing procedures that sort responses into favourable, unfavourable and neutral were not appropriate for narrative material, and they abandoned them. The false-notes count is a measure of critical evaluation, compared between high and low transportation groups. It is evidence of an association between absorption and reduced critical evaluation. It is not a dose-response curve, and the transportation comparison in that experiment was correlational rather than manipulated, so it does not on its own establish that absorption causes the reduction. I will not present it as more than it is.

Related work points the same way with its own hedges. Bruner suggested that narrative, rather than referring to reality, may create or constitute it — he puts it as a possibility, and so do I.

Place these next to the earlier results and an uncomfortable picture forms. Explanations increase acceptance regardless of correctness. Adding friction reduced overreliance where explanations alone did not. Absorption is associated with reduced critical evaluation. Reading these three literatures together, I arrive at a conclusion none of them states and none of them tested: that the qualities which make analytical output pleasant to consume — fluency, coherence, the shape of a story — are the same qualities that disarm the reader's verification instincts. Each study holds inside its own domain. The convergence is my reading of them, and it is the part of this paper I would most like to see someone test.

An AI agent is, among other things, an extremely cheap narrative generator. Pointing it at business data and instructing it to tell the story is not a neutral formatting choice. It is a decision to optimise for belief transfer in a context where the beliefs may be wrong.

The conclusion is not that narrative is forbidden. Narrative belongs in a specific place: as a presentation layer over claims that remain individually inspectable, never

as the layer where truth is stored. A story can frame the findings. Once the story becomes the finding, the organisation has substituted the most persuasive format for the most verifiable one, and the literature says readers will not notice the substitution happening to them.

9. The management implication

Everything above reduces to one structural problem: in AI-assisted analytics, the signals a vendor optimises and the outcomes a buyer needs have come apart. Economists have a name for this shape — a principal-agent problem, where the party doing the work is rewarded for something other than what the party paying for it actually needs. It has a companion in Goodhart's Law: once a measure becomes a target, it stops being a good measure.

Both apply here with unusual precision. Vendors optimise what users report: satisfaction, adoption, ease. Buçinca's trade-off indicates that these signals select against scrutiny, because the designs that most reduced overreliance were rated worst by the people using them. A market left to satisfaction metrics will converge — rationally, without bad intent — on fluent, frictionless, story-shaped systems that maximise acceptance. Every finding in this paper predicts that outcome, and none of it will appear in an NPS score.

This is why the central question of this paper is a management question. Whether to accept lower user satisfaction in exchange for fewer accepted errors cannot be settled by a product team through A/B testing, because the test's own success metric is the thing in dispute. The exchange has to be stated at that precision, because that is what the evidence supports: in Buçinca's study there was no significant difference in overall performance between the cognitive forcing conditions and the simple explanation conditions. The gain appeared where the AI was wrong, and even then the human-AI teams continued to perform worse than the AI model alone. What friction buys is not a better decision on average. It is a smaller share of decisions in which a wrong machine was believed. It is a governance decision, of the same family as segregation of duties, four-eyes approval, and materiality thresholds: friction imposed deliberately, where the cost of frictionless error exceeds the cost of the friction.

The research also indicates where the friction budget should go. Scrutiny-forcing works in the narrow sense established above, and it works unevenly: its benefits ran, on average, higher among readers already inclined toward effortful thinking. It also costs satisfaction where it is applied, though the study found no significant difference in the subjective ratings of low Need for Cognition participants, so the goodwill cost is not uniform either. So it should be positioned rather than blanketed: reserved for consequential, material claims, while routine awareness flows freely.

For the executive evaluating an analytical system, the literature compresses into four questions that rarely appear on a procurement checklist.

#	The question	What the science supports
1	What does it cost, in seconds and steps, to trace any displayed number to its source — and is that cost measured at all?	Readers weigh the cost and the benefit of checking against the cost of trusting; explanations change behaviour only when they change a term in that calculation. Unmeasured cost means unmanaged trust. *(Vasconcelos et al. 2023)*
2	Where stakes are material, does the system distinguish output that was *seen* from output that was *accepted*?	Adding engagement friction reduced overreliance where explanations alone did not; scrutiny has to be reintroduced deliberately. *(Buçinca et al. 2021)*
3	Is the volume reaching each reader governed by an explicit budget, or only by the system's capacity to generate?	A central capacity limit averaging about four chunks, under conditions where a limit is observable. *(Cowan 2001.* The conclusion that volume beyond capacity yields noise rather than insight is mine, not Cowan's.*)*
4	When a reader wants more depth, does their attention govern how far the analysis goes, or does the system decide unilaterally what everyone gets?	A wealth of information creates a poverty of attention and a need to allocate it efficiently. *(Simon 1971)*

A system that answers these four well may demo worse than one that answers them badly. That, in the end, is the argument: the demo measures delight, and delight is not reliably on the side of the truth.

10. A note on method

This paper was produced with AI assistance, and given its subject the reader deserves to know exactly where.

AI tools were used to scan and cross-reference a body of literature broader than what is cited here, across cognitive psychology, human-computer interaction and organisational science. That scan was not a systematic review: no pre-registered search protocol, no formal inclusion criteria. It was directed exploration, and I describe it as such.

Every finding carrying argumentative weight was then traced to its primary source. All twenty sources were retrieved and read; author lists, years, titles, venues and, where given, volume and page details were checked field by field against publisher records or retrieved copies. Both verbatim quotations were verified character by character against the published abstracts. The participant count in Buçinca et al. was confirmed as reported.

Two independent adversarial verification passes were then run against the retrieved sources, the second by a different reviewer working from a copy of this section withheld, so that it would form its own judgement rather than audit the first pass. Between them they found no fabricated source and no fabricated citation, and a

substantial number of places where an earlier draft stated more than its source supported. All are corrected here.

The first pass found seven: a cognitive-load claim about judgment and failure thresholds that Sweller does not make; a counterarguing result attributed to Green and Brock, who explicitly declined to measure it; a monocausal reading of Vasconcelos that their own fourth study contradicts; an exclusivity claim that made two rows of one table inconsistent with each other; a conclusion about noise attributed to Cowan and Sweller; a citation of Sweller for a claim about presentation format; and the substitution of "decision quality" for the "overreliance" that Buçinca actually measured.

The second pass found more, including two errors the first revision had itself introduced, and one substantive finding the first pass had missed entirely. That finding is worth stating plainly, because it cuts against the argument this paper makes: Buçinca et al. report no significant difference in overall performance between their cognitive forcing conditions and the simple explanation conditions, and note that even with forcing, the human-AI teams continued to perform worse than the AI model alone. The gain was concentrated in the cases where the AI was wrong. §9 has been rewritten to state the exchange at that precision. A paper recommending friction owes its readers the boundary of what friction was shown to buy.

A third pass went looking for evidence that would refute this paper rather than support it, and it found some. An earlier draft claimed that the industry asked cognitive science one question in 1984 and never asked another. That claim was false, and it was in the title's vicinity, the opening and the closing — the three places a reader looks first. Work on the cost of sensemaking, on staging information for attention, on the economics of information seeking, and on cognitive load in dashboards specifically has continued without interruption, some of it inside analytics vendors. §1 has been rewritten around what the evidence actually supports, which is a claim about propagation rather than about existence: the science exists and has not reached the defaults. That is a weaker statement about the industry's curiosity and a stronger one about its incentives, and it makes §9 the destination the paper was always heading for. Five sources were added in the same pass, including the ones that did the refuting.

One claim in this paper rests on my own professional experience rather than on published evidence: the observation in the opening that consumption runs low in large report estates. That comes from client engagements and internal audits at D-CAT, and I report it without a figure because I am not in a position to publish one. It is the observation that prompted the research, not evidence for anything concluded from it.

Cowan's "approximately four" is reported with its qualifier because the figure remains debated. The claim that agent production capacity has grown is qualitative and deliberate: no multiple is cited because none was verified. The hospital dashboard described in §7 is our own work, and it is included deliberately: the argument in this paper is that the failure is structural rather than a matter of

competence, and a structural claim I exempted myself from would not be worth much. Where an inference of my own carries weight — the ranking of business findings by surprise, the drift of unconstrained streams toward the negative, the three-part format taxonomy and the composed-record proposal in §7, the verification-cost instrumentation in §5 — it is marked in the text as mine rather than the cited authors'.

A paper arguing that trust should be managed through verification cost owes its readers a low verification cost of its own. Page ranges and DOIs are given below for that reason.

11. Closing

Forty years ago the industry asked cognitive science how a chart should be drawn, got a rigorous answer, and kept half of it. The science did not stop. It has precise things to say about how many findings a mind can hold, what explanations do to acceptance, what stories do to scrutiny, what surprise does to attention, and what it costs — in the only currency that matters — to check whether a machine is right. The findings themselves are not speculative. They are published, replicated in places, and in several cases produced by people working for analytics vendors. What has not happened is the crossing: none of it is in a default, a procurement question, or a metric anyone reports. What I have built on top of them is a different matter, and I have tried to mark every joint: the verification-cost instrumentation in §5, the extension of surprise to semantic findings in §6, the format taxonomy in §7, and the convergence argument in §8 are mine, and none of them has been tested.

The gap between agent output and human intake will not close from the production side. Models will keep getting faster; the reader will keep being a person whose pure short-term capacity, under the conditions where such a limit can be observed at all, averages about four chunks. The closing has to happen in the layer between the machine's claims and the human's decisions, and that layer needs what the pixel got in 1984: a science, taken seriously, built into the defaults.

This is also where I would revise my own diagnosis of the last twenty-five years. When a client with three and a half thousand reports has almost no consumption, the honest reading is not simply that the organisation lacks a data culture. It is that nobody ever asked what a human being can actually take in, and the system was built as though the answer were unlimited. Culture cannot fix an interface that ignores the reader's capacity. Design can.

The discipline that does this work does not have a settled name. It sits somewhere between evidence engineering, attention economics and trust calibration. The organisations that get there first will not look better in a demo. They will decide better, quarter after quarter, for reasons their dashboards were never built to show.

References

1. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.S. (2021). "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-16. DOI: 10.1145/3411764.3445717
2. Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D. (2001). "Bad is Stronger than Good." *Review of General Psychology*, 5(4), 323-370. DOI: 10.1037/1089-2680.5.4.323
3. Bruner, J. (1991). "The Narrative Construction of Reality." *Critical Inquiry*, 18(1), 1-21. DOI: 10.1086/448619
4. Buçinca, Z., Malaya, M.B., Gajos, K.Z. (2021). "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making." *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188, 21 pages. DOI: 10.1145/3449287
5. Burnay, C., Lega, M., Bouraga, S. (2024). "Business intelligence and cognitive loads: Proposition of a dashboard adoption model." *Data & Knowledge Engineering*, 152, 102310. DOI: 10.1016/j.datak.2024.102310
6. Cleveland, W.S., McGill, R. (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association*, 79(387), 531-554. DOI: 10.1080/01621459.1984.10478080
7. Cowan, N. (2001). "The magical number 4 in short-term memory: A reconsideration of mental storage capacity." *Behavioral and Brain Sciences*, 24(1), 87-114. DOI: 10.1017/S0140525X01003922
8. Green, M.C., Brock, T.C. (2000). "The Role of Transportation in the Persuasiveness of Public Narratives." *Journal of Personality and Social Psychology*, 79(5), 701-721. DOI: 10.1037/0022-3514.79.5.701
9. Gruber, M.J., Ranganath, C. (2019). "How Curiosity Enhances Hippocampus-Dependent Memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) Framework." *Trends in Cognitive Sciences*, 23(12), 1014-1025. DOI: 10.1016/j.tics.2019.10.003
10. Heer, J., Bostock, M. (2010). "Crowdsourcing graphical perception." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203-212. DOI: 10.1145/1753326.1753357
11. Itti, L., Baldi, P. (2009). "Bayesian surprise attracts human attention." *Vision Research*, 49(10), 1295-1306. DOI: 10.1016/j.visres.2008.09.007
12. Loewenstein, G. (1994). "The Psychology of Curiosity: A Review and Reinterpretation." *Psychological Bulletin*, 116(1), 75-98. DOI: 10.1037/0033-2909.116.1.75

13. Miller, G.A. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information." *Psychological Review*, 63(2), 81-97. DOI: 10.1037/h0043158

14. Pirolli, P., Card, S. (1999). "Information foraging." *Psychological Review*, 106(4), 643-675. DOI: 10.1037/0033-295X.106.4.643

15. Polanyi, M. (1966). *The Tacit Dimension.* Garden City, NY: Doubleday.

16. Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K. (1993). "The cost structure of sensemaking." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 269-276. DOI: 10.1145/169059.169209

17. Shneiderman, B. (1996). "The eyes have it: a task by data type taxonomy for information visualizations." *Proceedings 1996 IEEE Symposium on Visual Languages*, 336-343. DOI: 10.1109/VL.1996.545307

18. Simon, H.A. (1971). "Designing Organizations for an Information-Rich World." In M. Greenberger (Ed.), *Computers, Communications, and the Public Interest*, 38-52. Baltimore: The Johns Hopkins Press.

19. Sweller, J. (1988). "Cognitive Load During Problem Solving: Effects on Learning." *Cognitive Science*, 12(2), 257-285. DOI: 10.1207/s15516709cog1202_4

20. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R. (2023). "Explanations Can Reduce Overreliance on AI Systems During Decision-Making." *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Article 129, 38 pages. DOI: 10.1145/3579605

Verification status. All twenty sources were retrieved and read, and their citation details were checked field by field against publisher records or retrieved copies. Fifteen of them went through two independent adversarial passes; the five added in this revision (entries 5, 10, 14, 16, 17) were verified against publisher or Crossref records in a third pass. Three corrections stand from earlier drafts: entry 18 names the volume's editor, and entries 4 and 20 carry ACM article numbers and page counts rather than the page ranges a bibliographic database had normalised them into. Every entry matches its source.

Two authors are named in §1 without being cited: Tufte and Few appear as context for how the 1984 result reached practitioners, and no claim in this paper rests on either. The Tableau research group referred to in §1 was read from the company's own public research page. An earlier draft cited Taleb's *The Black Swan* for the narrative fallacy; that citation was removed, because it was the only work in the list that could not be checked against its primary text and the only one that was not peer reviewed. The argument it supported is carried by Green and Brock and by Bruner. Simon's page range varies across secondary sources; 38-52 is the range printed on the copy consulted.