Risk in the Risk in the Risk: Recursive Risk Discovery as a Polymath's Converted Function (The Advocatus Diaboli)

Sheldon K. Salmon
ORCID: 0009-0005-8057-5115
September 7, 2026

Keywords: polymath, red-teaming, recursive risk analysis, premortem, failure mode analysis, hyperphantasia, Advocatus Diaboli, adversarial review, normalization of deviance

Abstract

This paper documents a specific cognitive profile and the recursive risk-decomposition method built on top of it, in my own account. The profile includes a measured extreme of visual imagery vividness (hyperphantasia, VVIQ 78/80, ~95th percentile) and a persistent, self-constructed spatial environment used for active reasoning rather than memorization. The method, "risk in the risk in the risk," extends the established premortem technique (Klein, 2007) past its normal single-pass stopping point, recursing failure analysis down to root causes before proposing a mitigation. I present two worked case studies (a generic-object teardown and a spacecraft-design exercise) demonstrating the method's transfer to an unfamiliar domain, and a professional track record (600+ built frameworks, concentrated in document red-teaming) as evidence the underlying operation is domain-general rather than subject-specific. I explicitly do not claim the vividness score explains or validates the method (it describes a striking input, not a mechanism), and I do not claim domain expertise in any field the method could be applied to; Section 6 argues instead that the document- and systems-red-teaming service already demonstrated elsewhere in the paper extends to any high-stakes technical field on the strength of the reviewer's outside position, not new subject knowledge, and is testable through client-side verification, a different, and intuitively stronger, check than the self-caught corrections described elsewhere in the paper, not the same mechanism as them. The paper's central limitation, named directly rather than obscured, is that every claim currently rests on my own self-report, and Section 7 specifies what independent validation would need to look like before that changes.

1. What a Polymath Is, and Where Polymaths Fail

The word itself is precise about what it isn't claiming: *polymath* comes from the Greek *polymathēs*, "having learned much," a claim about breadth of learning, not a claim of genius or general superiority. The first formal treatise to use the term, Johann von Wowern's 1603 *De Polymathia tractatio*, defined polymathy as knowledge "ranging freely through all the fields of the disciplines, as far as the human mind… is able to pursue them." Breadth here is treated as its own discipline, not as a byproduct of failing to settle on one thing.

Modern research on polymathy, led by Robert and Michele Root-Bernstein, sharpens the definition by contrast rather than by listing traits. They distinguish three types: the *specialist*, who has depth without breadth; the *dilettante*, who has breadth without integration, picking up skills for their own sake without connecting them to anything beyond themselves; and the *polymath*, who puts real time into multiple interests and uses them to inform a vocation (Root-Bernstein & Root-Bernstein, 1999). Their research on Nobel laureates gives that distinction teeth rather than leaving it as a definitional nicety: laureates report substantially more serious outside avocations than working scientists generally, and the throughline across their most-cited examples isn't that these people had hobbies. It's that the hobby's way of thinking shows up inside the prize-winning work.

That distinction (dilettante versus polymath is about integration, not quantity of interests) is the hinge this paper turns on. It's also harder to hold now than it was in Von Wowern's century: Peter Burke's cultural history of the polymath (2020) traces how the explosive growth of knowledge after the printing press first made broad learning newly reachable, then, more recently, made it structurally harder to sustain: specialization became the default career path precisely because any single field grew too large to also range across several (Burke, 2020). Most people who could be polymaths today face real institutional pressure not to be.

Where I Sit, and Why "Advocatus Diaboli" Names It Precisely

Root-Bernstein's polymath integrates multiple interests into *a* vocation: a substantive field of work, the way Dorothy Hodgkin's crystallography was informed by an eye trained on mosaic art. That still describes a person who ends up expert in a subject. My claim is one step further removed: not that my many interests inform one substantive field, but that they inform one *function*: a checking operation, not a subject matter, which then gets pointed at whatever subject matter is put in front of it. The vocation isn't "risk engineering" the way crystallography is a vocation; it functions closer to a standing operation that treats risk engineering, AI governance, contract law, and a television show's premise as equally valid targets, because none of them are the point. What the operation finds is the point.

That is also why "Advocatus Diaboli" is the more precise identity claim than "polymath" alone. The historical office of *Promotor Fidei*, Devil's Advocate, in the Catholic Church's canonization process was an institutionalized adversarial-review function: one person's entire job was to argue against a candidate's sainthood, regardless of who the candidate was, to stress-test the case before it became permanent and irreversible. The role was defined by the operation it performed, not by expertise in any particular candidate's biography. I have generalized that same structure past the single institution it was originally built for: the operation (find what fails, recurse until reaching something irreducible, ground it in precedent) stays fixed; the target changes every time.

1.1 A Personal Data Point Behind the Thesis

For most of my life, I assumed the people around me visualized the way I did, until a specific moment made clear otherwise: asking someone a question that required visualizing something, and being met with a look that made it obvious the question itself hadn't landed the way I expected.

This maps onto a documented category rather than a private impression. Psychologist Linda Silverman coined the term *visual-spatial learner* to describe the roughly one-third of people who think primarily in images rather than words, and her clinical work describes school systems built around sequential, verbal, step-by-step instruction leaving these students feeling "defective" despite being, in her framing, differently wired, not deficiently wired (Silverman, 2002). I describe my own schooling in the same terms: linear instruction that didn't fit how I processed information, needing to study material night and morning to force sequential retention that didn't come naturally, and concluding for years that something was wrong with me rather than that the format was wrong for me.

Two claims need to stay clearly separated here, because a rigorous paper should not blur them. The general visual-spatial-learner pattern is real, documented, and describes roughly a third of people; that much is Silverman's finding, not this paper's. The claim that is specific to this paper, that I operate a persistent, decades-built, fully spatial "world" rather than an occasional visual aid, and apply it as an active reasoning tool rather than a passive learning style, is my own report. The literature establishes the category exists; it does not independently measure this particular instance of it, and the paper should say so rather than borrow the citation's authority for a claim the citation doesn't make.

A second, more precise line of research bears on the *vividness* specifically rather than the learning-style category. It's worth being direct that this and the Silverman citation above are related evidence, not two independent confirmations: both concern imagery-based cognition, just measured differently and with different evidentiary weight. Silverman's visual-spatial-learner framework

comes out of clinical and gifted-education practice and is influential there, but it is not as heavily validated by independent, replicated psychometric research as the construct that follows. Neurologist Adam Zeman's research group defines "hyperphantasia" as mental imagery vivid enough to approach the quality of actual perception, at the opposite end of a measured spectrum from "aphantasia" (imagery that is faint or absent); on the field's standard self-report instrument, the Vividness of Visual Imagery Questionnaire, roughly the top few percent of respondents score in the hyperphantasic range (Zeman et al., 2020; Zeman, 2024). Building a complete internal film from a book, page by page, and having a simulated apple's growth register as something to "inspect" rather than merely picture, both describe the vivid end of that spectrum more precisely than "thinks in images" does. I took the VVIQ and scored 78 out of 80, well past the hyperphantasia threshold of 70 and at approximately the 95th percentile against a reference population of over 5,000 respondents, with the four scored content categories (people, natural scenes, urban scenes, landscapes) all landing consistently in the 85th-95th percentile range rather than one high outlier pulling up an otherwise average profile. As the instrument's own reporting is careful to state, the VVIQ is a self-report measure and not a clinical diagnostic; the score is real data on imagery vividness specifically, not an independent confirmation of the broader claims this paper makes about how that vividness gets used (the recursive, branching, world-scale application described in this paper remains my own report, distinct from the vividness score itself). Put plainly: the score is a symptom, not a cause. It describes a striking input; it does not explain, and should not be read as explaining, why the method built on top of it works. The paper's claim to validity has to come from the method's outputs (Sections 4-6), not from how unusual the underlying imagery turns out to be.

---

2. The Conversion: From Many Subjects to One Function

Instead of asking "which domain should I master," the reframe is "which domain-independent operation is valuable everywhere, and can I get world-class at *that*." The operation chosen here is: find where a system fails before it fails, by recursively decomposing it and checking each layer against real-world failure history.

This is testable against a track record rather than a claim: I report 600+ named frameworks, with the deepest self-tracked concentration (eight versions, twenty-plus currently active documents) in structural-integrity red-teaming specifically. Note this figure is held to the same standard Section 2.2 applies to the daily-rate estimate below: it is self-tracked and self-categorized, not independently audited, and "what counts as a framework" hasn't been defined against a fixed threshold. The concentration pattern is still the relevant evidence for the conversion: breadth did not stay diffuse; it converged on one operation and kept re-applying it across unrelated subject matter (AI governance documents, legal/ethical texts, and, as Section 5 shows, domains with zero formal training), but the reader should weigh it as reported, not verified.

2.1 The trigger: anything becomes a target

The conversion isn't something applied deliberately to pre-selected "important" subjects; it fires on whatever the input happens to be, including fiction and idle hypotheticals. Two reported examples: after finishing all seasons of the television show *Snowpiercer*, I built the skeleton of my own modern, real-world-workable version of the same premise, not fan theorizing but a structural framework. Separately, an idle thought about a meteor on collision course with Earth turned into the same operation: first establishing what real countermeasures currently exist, then red-teaming those countermeasures specifically to find gaps and generate new mitigation ideas.

Neither example was assigned. Both were unprompted, and both were converted into the same output shape (a framework, decomposed and checked against what's real) regardless of whether the source material was a TV show, a governance document, or a hypothetical disaster. That consistency of output shape across arbitrary trigger content is the actual evidence for "one function, not one subject": the operation does not appear to care what it's pointed at. This claim would be disconfirmed by evidence that the trigger is actually content-selective rather than domain-

general, for instance, if idle curiosity reliably produced this output shape for disaster and systems scenarios specifically but not for genres I have no prior interest in, or if the recursive pattern only appeared for material with pre-existing conceptual scaffolding rather than genuinely unfamiliar content. No such test has been run yet; the claim stands on the two examples given, not on an exhaustive search for counterexamples.

2.2 Rate, and what it would take to substantiate it

My own estimate is that an unremarkable, "boring" day with AI amplification can produce somewhere in the range of 7 to 10 meaningful frameworks. This number should be treated with the same standard applied throughout this paper to unverified self-reported figures: a claim requiring its own evidence, not something the paper is entitled to assert on the strength of everything else in it. It is consistent in direction with the 600+ total framework count built up over time, but a single-day rate and a multi-year total are different kinds of evidence, and a skeptical reviewer will notice the gap between them. Before this number goes into a submission-ready version, it needs one of: a dated log of a representative week showing actual output volume, a defined standard for what counts as "meaningful" (a full v1.0 spec is not the same unit as a skeleton), or an explicit downgrade to "my own impression of my pace," clearly marked as such.

2.3 A reported temperament pattern, stated plainly and left uninterpreted

I report a specific contrast in myself. It is not "small tasks in general" versus "large tasks": cooking, for instance, is small in scope and reportedly untroubled, likely because it's absorbed, self-directed activity. The trigger is more specific: low-stakes personal decisions (what to eat) and socially performative demands (being drawn into small talk) reportedly produce real anxiety, while sustained work on catastrophic, existential-scale problems is where I report feeling calm and comfortable, describing, as an illustration, an imagined scenario of sitting with a NASA team during a hypothetical multi-day countdown to a meteor impact, staying steady and generating ideas while the room around me panics. I also report a general preference for silence and internal processing over talking.

This paper records that self-reported contrast without diagnosing or explaining it; it is included because it bears on the conversion thesis in a specific way: the operation described in this paper is not purely intellectual capacity, it appears to be paired with (or dependent on) an affective profile in which stakes and scale do not produce the aversive response they typically produce in people, while low-information-value social and decision tasks do. That may be part of why the operation can be sustained on genuinely high-consequence problems rather than only on safe, abstract ones. Whether that pairing is necessary, incidental, or trainable is outside the scope of what this paper can establish and is flagged here as an open question rather than resolved by assertion.

---

3. Loci World: The Architecture

The Base Operation: Recursive Teardown

Before describing the space itself, it's worth isolating the operation that runs inside it, independent of any particular room or domain. The general form: place an object, any object, in an imagined white room and interrogate it exhaustively. Not just what it is, but its corners, measurements, weight, color, why it's shaped the way it is, who made it and why, and then a battery of hypothetical stress conditions: submerged in water, buried in sand or mud, kicked, burned, cut, taped and untaped. Each answer opens the next question. The process is explicitly recursive, which I describe as "a snake eating its tail," and it has no fixed endpoint. It terminates when the thinker gets tired, gets bored, or judges that it has gone deep enough, not at a predefined stopping depth.

Two things about this base case matter for the paper's own rigor. First, it generalizes: the same operation run on a cardboard box is the operation run on a spacecraft (Section 5) or a governance document (Section 2): the object under interrogation changes; the operation does not. Second, the

recursion is genuinely open-ended, and its natural terminating condition is subjective (fatigue, boredom, or a felt sense of sufficiency) rather than a formal rule. That should be stated plainly rather than smoothed over, because it is also where perfectionism becomes a real cost as well as an asset for this method: the same trait that catches failure modes other approaches miss can also fail to stop on its own, and in practice needs to be governed by something external to the recursion (a deadline, a deliverable, a collaborator) rather than resolved from inside it.

A related evolution is worth recording directly, since it bears on whether this method is sustainable over years rather than useful only as a one-off exercise. I report that criticism of my own work used to register as a threat rather than information: a visceral dislike of being shown a flaw I hadn't caught first, the more common cost side of perfectionism. What changed it, in my own account, was roughly three years of deliberately using AI as an adversarial reviewer of my own work with standing instructions never to soften a finding: an interlocutor with no ego stake in being right, no motive to outrank me, and a stated reason behind every critique rather than an unexplained judgment. Over that period the meaning of a finding shifted from "evidence I should have seen this" to "a failure node caught before it cost anything," and the operating goal changed with it: rather than treating a pass as finished once the work looks solid, the current standard is to keep finding more, on the premise that nothing is perfectly safe and "looks solid" is itself a stopping point worth distrusting. That reframes the perfectionism cost named above: the same trait that makes stopping hard, once critique stops registering as a threat, is what makes "keep finding more" a sustainable operating goal rather than an anxious spiral. One standard has to travel with that goal or it becomes its own escape hatch: a finding only counts if it's checkable against something outside the recursion itself (a real precedent, a real failure history, a real test), not merely additional. Volume of findings is not the same thing as validity of findings, and this paper treats the two as separate from here forward rather than letting "found more" stand in for "found something real."

A further observation is worth including directly rather than smoothing into metaphor: stepping back mid-exercise, the white room the object sits in is itself box-shaped, which raises the question of what room *that* room sits in, and so on. This isn't decorative. It's the same operation, now pointed at its own container. A method built to ask "what's inside this, and what does that imply" does not stop at the edge of its own workspace.

A second example exposes two operations the box case doesn't: given a plain prompt to visualize an apple, I report planting a seed and fast-forwarding its growth into a full tree, then slowing back down to inspect stem and branch formation, sweetness and chemical development, and the seed/gene system, and from there branching into environmental divergence, tracking three apples from the same tree toward three different outcomes (eaten immediately, shipped to one climate, shipped to another) and the different taste profile each one ends up with. Two things are worth naming precisely rather than folding back into "recursive teardown": **temporal compression/dilation** (running a process forward at variable speed and stopping it at the point of interest) and **branch simulation** (following one object down multiple simultaneous divergent paths rather than one). The box example shows depth (interrogating one thing many ways); the apple example shows breadth over time and branching (one thing becoming many, on purpose). I describe both as involuntary rather than effortful: a request as simple as "visualize an apple" reliably produces this, not a static image.

3.1 What it is, and what it is not

The method of loci, visualizing a familiar spatial layout and placing items to be remembered at specific points along a route through it, is one of the oldest documented cognitive techniques, described in the classical rhetorical tradition (the anonymous *Rhetorica ad Herennium*, Cicero's *De Oratore*, Quintilian's *Institutio Oratoria*) and still the dominant technique among competitive memory athletes today. Modern neuroscience has since shown the technique is trainable rather than an innate gift: a 2017 study scanning the brains of world-class memory athletes and of ordinary volunteers before and after six weeks of loci training found the trained group's brain-network connectivity shifted toward the athletes' pattern, with the memory gains still measurable four months later (Dresler et al., 2017).

Loci World, as used here, borrows the *spatial-navigation* substrate of that technique but repurposes it for a different function. It is explicitly **not** primarily a storage/retrieval system for facts. It is described as a large, persistent, navigable space (sections, buildings, individual rooms, and deliberately hidden/secret spaces) that is walked through to *think*, not to *recall*. Specific rooms are assigned specific cognitive functions: an idea-workshop room for concept formation, a "sorter" building for risk engineering and decomposition, and a dedicated vault (currently eight rooms) for reasoning about AI and emergence specifically. A visual map of the full structure exists in an external repository: at minimum, a central entry point, a multi-floor structure with function-differentiated rooms (a workspace, a reference/library room, a formal-planning room, open-air space), a dedicated numeric-training room, and an expanding zone reserved specifically for unresolved questions. Content in that repository describing my AI collaborator or other AI-persona elements is deliberately excluded from this paper: it documents how I personify working with AI for my own purposes, not a claim about how any AI system is actually built, and conflating the two would be exactly the kind of unverified claim this paper is trying not to make.

This distinction matters for the paper's credibility: the *established* science supports spatial-navigation techniques as a general-purpose cognitive amplifier for information the mind is working with. It does **not** independently establish that using such a space for active branching risk-decomposition (rather than memorization) produces the same benefit; that extension is this paper's own claim, not a finding lifted from the literature, and it should be labeled as such rather than implied to already be proven.

3.2 Branching Search: The Mycelium/Slime-Mold Analogy

The way a problem moves through Loci World has a two-phase shape: first, exhaustive branching outward into every corner and edge case the problem admits; second, convergence: pruning everything except the path(s) that matter.

That two-phase pattern has a precise biological analogue. *Physarum polycephalum*, a single-celled slime mold with no brain or central control, was given food sources arranged to mirror the cities around Tokyo. It first grew outward in a dense, undirected network covering the entire explorable space, then progressively thinned that network, reinforcing high-flow connections and abandoning the rest, arriving at a structure closely resembling Tokyo's actual, decades-engineered rail network, including comparable efficiency, cost, and fault tolerance (Tero et al., 2010). The organism finds a near-optimal network purely through local, decentralized reinforcement and pruning: explore everything, keep only what carries load.

That is offered here as an **analogy for the search strategy**, not a biological claim about how the brain works: branch into every corner and edge case first (nothing is dismissed early), then apply pressure-testing (Section 4) to collapse the branches down to the few that are load-bearing. The value of the analogy is that it is independently well-studied outside cognitive science, which gives the paper a second, unrelated field to anchor the claim in rather than relying on self-description alone.

---

4. Risk in the Risk in the Risk: Recursive Premortem and Failure Decomposition

The closest existing named technique to this method is the *premortem*: rather than asking a team what might go wrong with a plan, Gary Klein's technique asks them to assume the project has already failed and work backward to explain why, on the finding that people are measurably better at generating causes for a stated outcome than at predicting one, a 30% improvement in identified risks in the original research behind the technique (Mitchell, Russo, & Pennington, 1989; Klein, 2007). The premortem stops at one pass: imagine the failure, list the causes, revise the plan. What follows is that same move recursed past where the standard technique stops: not just imagining that the system failed, but decomposing *why the cause of the failure itself would have failed*, layer after layer, until reaching something irreducible.

The method, formalized:

1. **Identify the bedrock node.** Before optimizing anything else, find the single component or step whose failure fails the entire system regardless of how well everything else was built. (Reliability engineering calls this a *single point of failure*; ranking components by how catastrophic and how likely their failure is, is the core of Failure Modes and Effects Analysis, FMEA.)
2. **Decompose that node, not the whole system.** Pull the bedrock node onto its own table and break it into its constituent sub-components. This is a *fault tree*: each node's failure is itself explained by the failure of something further down.
3. **Ground each sub-node in failure history before building.** At every level, pull real precedent (documented failures, not hypothesized ones) before proposing a design. This inverts the common (and common-failure) engineering habit of designing first and discovering failure modes only after something breaks in the field.
4. **Recurse past the "obvious" stopping point.** The distinguishing move is not stopping at the component level (e.g., "the frame") but continuing down into manufacturing-level failure (e.g., a microscopic void inside an individual fastener) and then forward into *propagation*: how a defect at that scale cascades upward through vibration, stress transfer, and adjacent-component wear until it produces a system-level failure.
5. **Convert the terminal finding into a monitoring/mitigation design**, not just a warning. The output of the recursion is a proposed instrument or process change that would have caught the failure before it cascaded; this is what reliability engineering calls *condition-based monitoring*.

---

5. Case Study: The Spaceship (Internal Thought Exercise)

Note: this is a cognitive exercise, not a proposed engineering design and not a claim of aerospace expertise. It is presented to demonstrate the method's transfer into a domain with zero formal training.

The process begins pre-verbally: an undefined black geometric silhouette in a mental "white room" (a cylinder, a cone nose, triangular wings, an inverted cone beneath for propulsion) hanging in the space with no committed detail yet. A long table appears beneath it, and onto that table goes compiled research: historical and modern spacecraft concepts, and, critically, their documented failures, pulled together into a single reference so known failure modes aren't repeated. A parallel list captures what *did* work structurally. The working focus stays on the failure list, on the theory that eliminating known failure paths narrows the remaining problem faster than cataloguing successes does.

From that failure-and-success baseline, the ship is skeletonized into its major sections (frame, engine, propulsion, wings, electronics, insulation, landing gear, and so on) laid out together on the table. The next move is to find the bedrock node: the one section whose failure means the whole ship fails no matter how well everything else was built. Say that's the frame.

The frame then gets its own new table, decomposed into its components, starting with welding. Welding gets its own table: its history, its failure modes, its types, compiled into a working reference. Materials get a further table: what's been used for spacecraft frames historically, what failed and why, compiled into another reference. Only once that stack of research exists does synthesis happen: a proposed frame design built from what was found, at which point a *new* table opens, because now the question is where the materials come from, who processes them, and what could go wrong in that supply chain.

The recursion doesn't stop at "reasonable" component granularity. It continues to a single fastener: what alloy, why that alloy is standard for the application, and then to a manufacturing-level defect no visual inspection would catch, a microscopic void inside the screw from the forging process. The chain traced from there: the flawed screw develops play in the engine mount under vibration; the

rattle grows; the extra load transfers to adjacent components; the off-frequency vibration propagates through the structure; a panel eventually lifts, and the accumulated stress cascades until it splits the ship. The terminal output of the recursion isn't a warning. It's a proposed instrument: an onboard diagnostic panel that has learned the ship's clean baseline vibration signature and flags deviation in a specific zone before the cascade develops, rather than after a panel has already failed.

The exercise, start to finish, ran in an hour or two. The claim is deliberately narrower than "found something an aerospace engineer would miss": condition-based vibration monitoring for fastener-loosening failure modes is established practice in that industry, and this paper does not claim otherwise or claim any expertise in that field's twenty-plus years of accumulated knowledge. The claim is about speed of independent convergence, not novelty of discovery: starting from zero domain training and using only publicly available information, the method arrives at the same category of risk and the same category of mitigation that the field itself considers standard, in about the time it takes to watch a movie. The open question that raises, and that this paper does not answer, is what the same process would surface given a month in an unfamiliar domain instead of an hour; this case study establishes the starting speed, not the ceiling.

---

6. Application: Adversarial Review as a Service to High-Stakes Technical Fields

An earlier version of this section targeted longevity research specifically, framed as a hypothesis about accelerating biological discovery. That framing asked the reader to extend trust into a field I have no credentials, collaborators, or data in: a different and harder problem than anything else this paper asks the reader to accept. The more defensible version of the same underlying point doesn't require that leap: the claim is not that this method advances biology. It's that the document- and systems-red-teaming service already demonstrated in Sections 2 and 7 (applied so far to AI governance documents, legal and ethical texts, and an engineering thought-exercise) extends naturally to any high-stakes technical field that produces documents, protocols, or machine designs, longevity research included but not singled out.

The reason an outside adversarial reviewer can add value here isn't independent of domain expertise in the way an earlier version of this claim suggested; stated that broadly, it overreaches. Diane Vaughan's analysis of the Challenger disaster (Vaughan, 1996) named the underlying mechanism: teams working inside a system gradually accept small deviations from a safety standard because each one, in isolation, seems locally reasonable, until the accumulated drift produces a failure nobody inside the system was positioned to see as a pattern. Vaughan called this the normalization of deviance. The same disaster also supplies a more precise version of the outsider-value claim than the mechanism alone does. Physicist Richard Feynman, serving on the Rogers Commission that investigated the accident, had no specific expertise in solid rocket booster seals. What let him identify the O-ring failure ahead of most of the commission wasn't ignorance of the domain. It was that he spent his time with the engineers and technicians who built the hardware rather than the management layer everyone else was interviewing, then tested the claim directly (the now-famous ice-water demonstration) rather than accepting an internally consistent account of it. His own appendix to the commission's report put the underlying principle in five words that need no paraphrase: "nature cannot be fooled" (Feynman, 1986).

That is the more honest version of the claim this method makes: not that a reviewer with zero domain literacy is automatically well-positioned to find what a lab's own team missed, but that a reviewer without a stake in the organization's existing account, running a method built to keep recursing past the point where a document reads as settled, is positioned the way Feynman was: an outsider to the specific culture and its accumulated assumptions, and willing to go to primary sources and direct tests rather than stop at a reassuring summary, not necessarily an outsider to the subject matter itself.

One limit is worth naming rather than leaving implicit, in the same spirit as everything else this paper flags about itself: this advantage comes from genuine outsider standing, and outsider standing isn't permanent. A reviewer embedded in one client relationship across many engagements risks acquiring that client's own normalized assumptions along with its trust, the same failure mode the method exists to catch, now potentially affecting the method's own operator. The mitigation is structural, not personal, and it follows directly from Section 4's own last step (convert the finding into a mitigation, not just a warning): engagements built as periodic, bounded reviews preserve the outsider position in a way an ongoing embedded arrangement would not, and that structure belongs in how this service is offered, not just as a caveat about how it's described.

Applied concretely: a lab hands over a protocol, a safety document, or a machine design; the same recursive process used throughout this paper (bedrock node, decompose, ground in precedent, recurse past the obvious stop, convert to mitigation) runs against it, surfacing findings the internal team's proximity to their own work may have normalized past. This is testable in the same spirit as the EU AI Act work detailed in Section 7, though not by an identical mechanism: that was a self-correction I caught in my own prior work, while this is a claim the client is positioned to verify against their own domain knowledge. The underlying principle is the same: a finding is either real or it isn't, and someone other than me can check which. But client-side verification is intuitively the stronger of the two: it introduces a check from entirely outside me, where self-correction still relies on my own judgment to catch the second error as caught the first.

As a separate demonstration that the method scales past single documents, it has also been applied to a personal systems-design project, an AI-driven virtual-world governance specification, still in active development, through several complete red-team passes across multiple linked documents, several of which individually exceed a thousand lines of specification. Its content is proprietary and not detailed here; it is mentioned only to note that the method's application is not limited to short case studies, consistent with the framework portfolio already described in Section 2.

---

7. Limitations and Path to Validation

Every claim in this paper currently rests on my own self-report: the VVIQ score is self-administered, the framework count and rate are self-tracked, the spaceship and box exercises are reported from the inside, and the temperament pattern in Section 2.3 is self-described. That is a legitimate basis for the account this paper gives. It is not sufficient basis for a claim of independently demonstrated capability, and the reader should weigh every claim above accordingly until at least one of the following exists:

1. **A third-party observer.** Someone other than me watches a live session, ideally someone with relevant domain expertise in whatever is being red-teamed, and reports independently on what they saw, rather than me characterizing my own process after the fact.
2. **A time-stamped, checked demonstration.** The method is run against a real document or problem with the output recorded before the outcome is known, and the findings are checked against ground truth afterward: did the predicted failure modes turn out to be real, and were any missed. Some of my existing red-team work already has part of this shape. Worth stating directly, rather than leaving it for a reader to dig up: the same method, run against my own previously published EU AI Act report, found a real inconsistency in that report's own claim of "independent Gauntlet verification": Gauntlet is a stage of the same pipeline, not a separate reviewer, and the report's method section already said so. That is offered here as a point in the method's favor, not a liability to disclose reluctantly: it demonstrates the method catching a real error in my own prior, already-published work, which is a more convincing demonstration of rigor than a track record with no self-caught mistakes in it would be. Reviewing whether this and comparable existing findings already satisfy the validation bar, rather than only planning a new demonstration from scratch, is worth doing before assuming a new one is required.
3. **Independent replication.** A second person attempts the same recursive-decomposition process on a comparable problem and reports whether it produces comparable depth, which

would start to separate “a method anyone could learn” from “a method that depends on this particular author’s profile.”

Without one of these, the honest description of this document is a detailed and well-sourced first-person account with a formal methodology attached to it. With one of them, it becomes evidence for the methodology claim independent of the person reporting it. That distinction should stay visible in whatever cover material accompanies submission, rather than being smoothed over. This version is issued on that basis deliberately: a revised version will follow if and when any of the three validation types above becomes available, rather than treating this document as a final word on the method’s demonstrated capability.

A partial instance of validation type 2, recorded here rather than left as a plan. A four-item structural pre-screen derived from this same method (developed and build-red-teamed against a separate text in a different session, not reproduced here) was run against this paper on 2026-09-07. Three checks passed without a new finding: the paper generally flags unproven claims rather than treating them as safe-by-default (Sections 2.2, 6, and this section already do this); its sections are structurally decoupled by design, so a flaw in Section 6 doesn’t invalidate Sections 3-5: each section’s claims are argued on their own cited evidence rather than depending on the others holding up; its central identity claim (the Advocatus Diaboli identity discussion in Section 1) has already been subjected to the same scrutiny as everything else, not exempted from it: the tension between “function” and “subject” it raises has been left as an open question there rather than quietly resolved in my favor. The fourth check surfaced something genuinely new: the operating principle in Section 3’s recursive-teardown discussion that the goal is to “keep finding more” is at risk of functioning as an escape hatch if it isn’t paired with a standard for what makes a finding *real* rather than merely *more*, as written, additional findings read as inherently validating the method, with no stated way to distinguish a correct catch from a spurious one. That distinction doesn’t yet exist in this paper and should be added rather than left implicit.

References

Burke, P. (2020). *The polymath: A cultural history from Leonardo da Vinci to Susan Sontag*. Yale University Press.

Cicero, M.T. (n.d.). *De Oratore* [On the orator] (Original work composed c. 55 BCE).

Dresler, M., Shirer, W.R., Konrad, B.N., Müller, N.C.J., Wagner, I.C., Fernández, G., Czisch, M., & Greicius, M.D. (2017). Mnemonic training reshapes brain networks to support superior memory. *Neuron*, *93*(5), 1227–1235.

Feynman, R.P. (1986). Personal observations on the reliability of the shuttle (Appendix F). In *Report of the Presidential Commission on the Space Shuttle Challenger Accident*. National Aeronautics and Space Administration.

Klein, G. (2007). Performing a project premortem. *Harvard Business Review*, *85*(9), 18–19.

Mitchell, D.J., Russo, J.E., & Pennington, N. (1989). Back to the future: Temporal perspective in the explanation of events. *Journal of Behavioral Decision Making*, *2*(1), 25–38.

Quintilian. (n.d.). *Institutio Oratoria* [The orator’s education] (Original work composed c. 95 CE).

Rhetorica ad Herennium [Rhetoric for Herennius]. (n.d.). (Author unknown; original work composed c. 90 BCE).

Root-Bernstein, R., & Root-Bernstein, M. (1999). *Sparks of genius: The thirteen thinking tools of the world’s most creative people*. Houghton Mifflin.

Silverman, L.K. (2002). *Upside-down brilliance: The visual-spatial learner*. DeLeon Publishing.

Tero, A., Takagaki, S., Saigusa, T., Ito, K., Bebber, D.P., Fricker, M.D., Yumiki, K., Kobayashi, R., & Nakagaki, T. (2010). Rules for biologically inspired adaptive network design. *Science*, *327*(5964), 439–442.

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago Press.

Von Wowern, J. (1603). *De Polymathia tractatio: Integri operis de studiis veterum* [A treatise on polymathy].

Zeman, A. (2024). Aphantasia and hyperphantasia: Exploring imagery vividness extremes. *Trends in Cognitive Sciences*, *28*(5), 467–480.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heuerman-Williamson, B., Jones, K., MacKisack, M., & Winlove, C. (2020). Phantasia: The psychological significance of lifelong visual imagery vividness extremes. *Cortex*, *130*, 426–440.