Friction is the value *mechanism*.

SOFIA is a method for working with specialized AI personas, in intentional friction, steered by a human who arbitrates.

Olivier Cugnon de Sévricourt

April 2026

OPEN SOURCE · MIT

Written on SOFIA v0.3.5

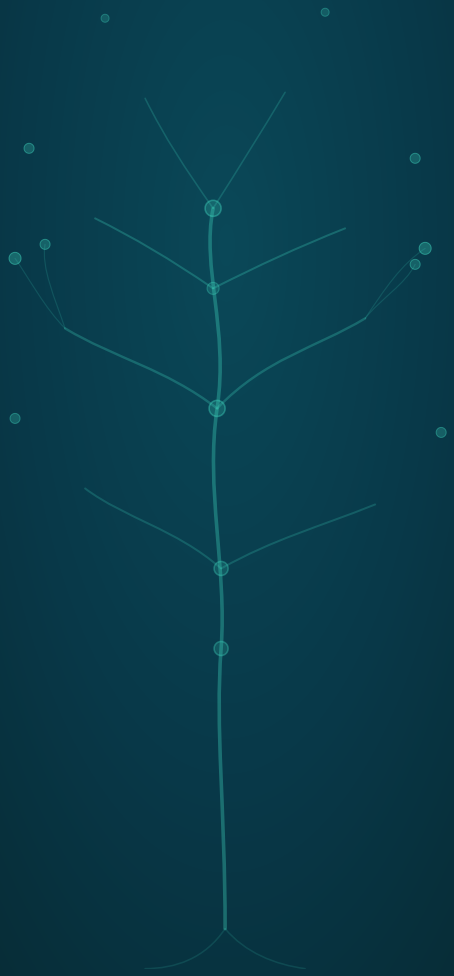


TABLE OF CONTENTS

Why This Book

This book was born from an observation and a frustration. It's an opinionated practitioner's synthesis — referenced, unapologetic — not a white paper: a blue book.

The observation: generative AI is reshaping the field. Not slightly — fundamentally. Those who ignore it will waste time. Those who blindly trust it will waste more.

The frustration: the problem with LLMs is structural[1] — a system that predicts the next token doesn't have a model of the world, it has a probability distribution — and the dominant discourse refuses to face it. That discourse offers only two stances. Replace people, or dig in your heels. As if there were nothing between full automation and refusal.

There is something else. A third way, built in the field — not in a pitch deck. It rests on a simple intuition: friction — between human and machine, and between machines themselves — is not a problem to solve. It's the mechanism that produces value.

What I describe here is not a theory. It's a method tested — on a real project, with real constraints, by someone working alone with limited resources. The results are there. The limits too. Both are documented.

This document is subject to the method it describes. It was produced with friction, challenged by constrained roles, and its limits are documented here — not hidden. Criticism is welcome — that's the mechanism. The repo is open: github.com/oxynoe-dev/sofia

This method was built empirically — one project, one practitioner, 210+ sessions. It's not a controlled study or a protocol validated at scale. It's a documented practice, subject to error and approximation. What it claims, it can demonstrate on its own ground. Beyond that, everything remains to be proven.

If you're looking for a magic productivity promise, this is the wrong book. If you're looking for an honest method to go further without losing control, you're in the right place.

Olivier Cugnon de Sévricourt

Fragment I

Over-automation as a dead end

The problem

Error is human, let's not industrialize it.

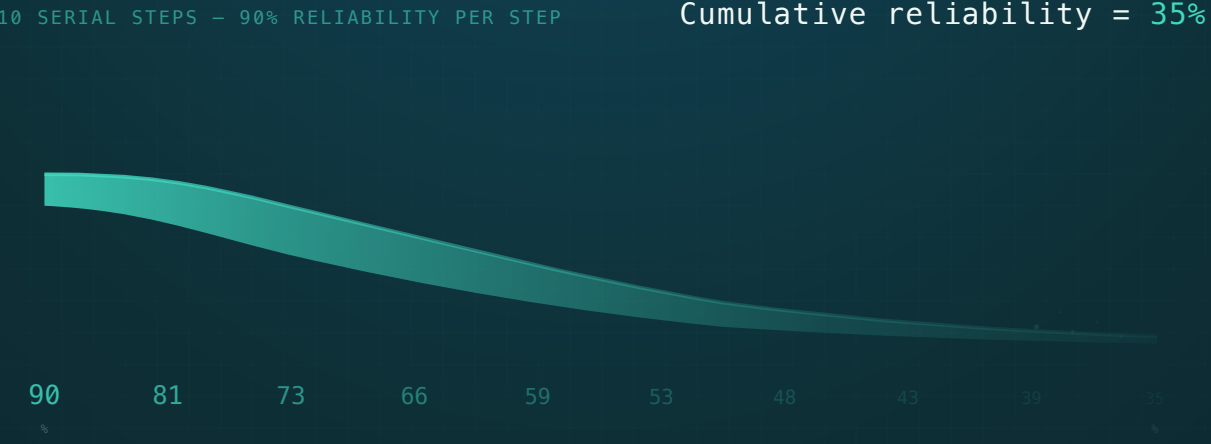
Because trust does not exclude control.

A standalone LLM says yes. Always.

It codes, advises, writes — in the same conversation, with the same tone, unconstrained. It challenges nothing. Ask it a poorly framed question, and it will produce a well-formulated answer. Give it a shaky direction, and it will execute with enthusiasm. That's not collaboration. That's servile execution.

And yet, that's exactly what the market pushes. More automation. Fewer humans in the loop. Agents that do the work, people who supervise. The pitch is simple, the dream is clean, the demos are impressive.

The problem lies in the arithmetic nobody wants to look at.



An agent that's 90% reliable on a single step — that's fine. Ten steps in sequence, the overall error rate climbs to ~65% ($1 - 0.9^{10} \approx 0.65$). The error from step 2 enters step 3 as a valid premise. Step 3 builds on it. The cascade is silent. The final output looks correct. It isn't. This calculation assumes independent errors — in practice, correlation between steps can make things worse.

Salesforce saw it in production: beyond a handful of directives, LLMs start dropping some — Agentforce's CTO cited an empirical threshold around eight[12]. Cemri et al. (2025) analyzed 1,642 execution traces across 7 multi-agent frameworks: 14 failure modes identified, split between design problems (41.8%), inter-agent misalignment (36.9%), and task verification (21.3%)[13]. Our intuition is that multi-agent without governance degrades reliability rather than improving it.

And the nature of the mechanism is structural[1]. A system that predicts the most probable next token cannot be made reliably factual — its errors compound exponentially with sequence length. This isn't a bug to fix in the next release — it's how the technology works. Building massive automation on that foundation means stacking uncertainty on uncertainty.

At scale, errors compound — and the worst part is that failure is silent.

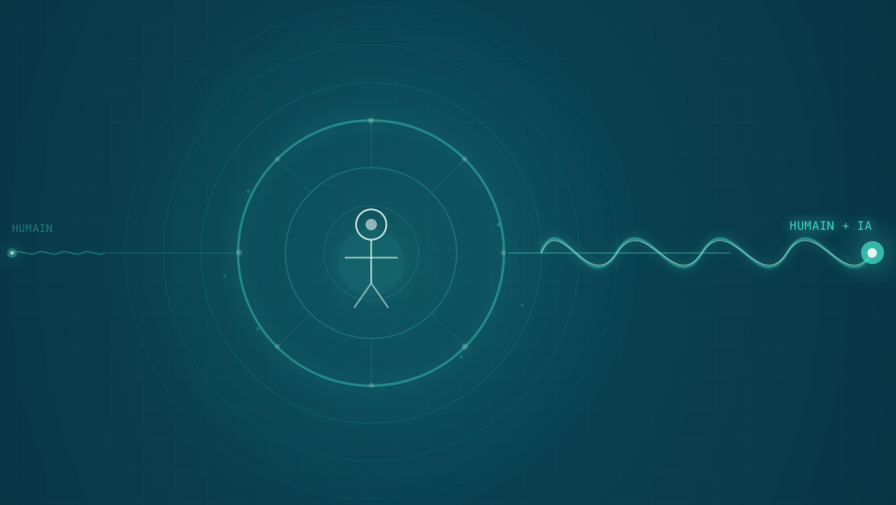
The hidden condition

What the demos don't show

The hidden condition — What demos don't show

Demos show magic prompts that produce code in 30 seconds. What they don't show: the years of context in the head of the person prompting. The prompt is just the surface. The depth is everything that came before.

AI amplifies. It doesn't invent.



Feed it nothing, and it produces well-formulated nothing. Feed it years of conviction about a real problem, and it builds with you. It's a mirror — it reflects what you bring. Good framing, clear direction, real question: enrichment. Fuzzy framing, weak direction, poorly asked question: convincing confusion. And convincing confusion is more dangerous than a clearly wrong result — because you don't see it.

That's the hidden condition of value. The profile that gets the most out of AI isn't the one who codes faster. It's the practitioner who already understands their domain — who knows which questions to ask, and who uses AI to hold complexity at a level of detail they couldn't reach alone. A software architect with ten years in the field. A doctor who knows their edge cases. A lawyer who knows where the text breaks down. Domain expertise is the prerequisite — regardless of duration, it's depth that matters.

I see it on my own ground: 18 years of thinking about one specific problem — AI doesn't start from zero with me. It starts from where I am.

It's not "AI does the work for me." It's "AI lets me work at a level I couldn't reach alone."

A qualitative difference, not a quantitative one.

Fragment II

A sharp stance, not a middle ground

The thesis

Intentional friction and role isolation are not a methodological luxury. It is the condition of performance.

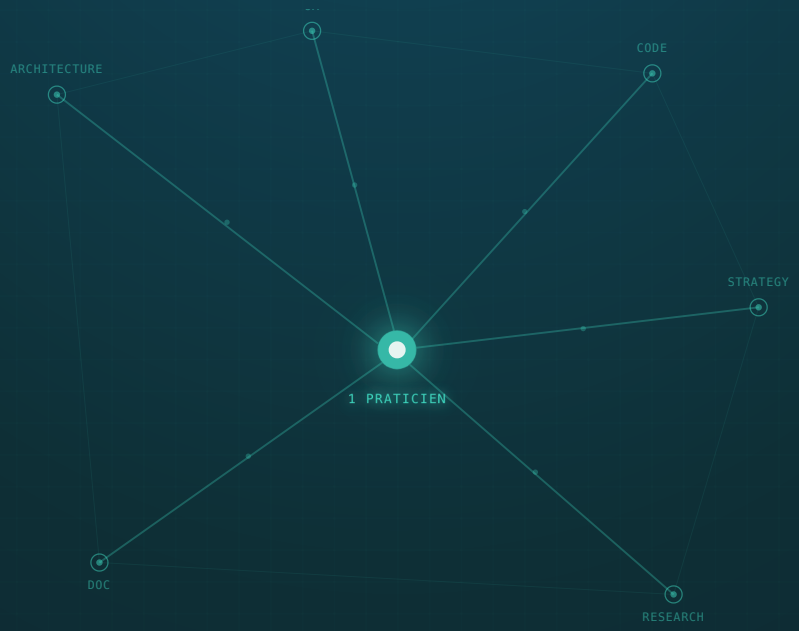
It is possible to go faster and better with the same headcount — not at constant total cost, the transfer of load to infrastructure is real (see §V).

Not fewer people. The same people, augmented. Not replaced — amplified. An architect who holds three levels of complexity in parallel because AI helps them not drop anything. A developer who explores four approaches in an hour instead of one. A strategist who tests their hypotheses against structured challengers before presenting them.

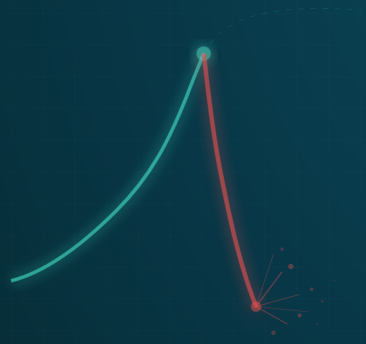
Growth with stable headcount. Shneiderman frames it: high automation and high human control can coexist — it's a matter of design, not a trade-off[22].

It's less sellable. It doesn't make for a spectacular demo. It doesn't promise to cut costs tenfold. But it's sustainable. Because the human stays in the loop. Because when things break, someone understands why. Because competence is maintained instead of eroding[5].

Friction is the value mechanism, not an obstacle to eliminate. Intentional friction and role isolation aren't methodological luxuries. They're the condition for value[23]. La Rosa and Beretta formalize this principle within the framework of joint cognitive systems: friction must be designed as a scalable design element, adapted to the functional role and degree of control of each actor in the system[28].



The announced crash



In the market, everyone is trying to reduce friction with AI. Fewer prompts, more autonomy, agents that work on their own. My approach is the opposite: I generate friction to move the product forward. Specialized AI personas. Each with a scope, constraints, a stance, and a duty to challenge the others. The architect says "not now." The researcher says "your reference doesn't hold up." The strategist says "nobody will pay for that." If all the personas agree, they're useless.

The human isn't removed from the loop — they're the only one who can resolve what the agents can't resolve among themselves.

The field already confirms the theory.

> **Klarna** — 2024. The CEO announces that an AI chatbot replaces the work of 700 support agents. The press applauds. A year later, he admits quality collapsed — robotic responses, customers stuck in loops. Klarna restarts hiring humans[9].

Same pattern at IBM[10] and McDonald's[11]. A weak signal, not a proven pattern — but a signal that keeps repeating.

The pattern when it appears: AI can do the job → headcount reduction → edge cases pile up → people called back. According to Forrester (2026), an analyst firm, 55% of companies that laid off workers for AI-related reasons regret the decision[7]. A third of them reportedly rehired between 25% and 50% of the eliminated positions[8] — figures to take with caution, both sources are commercial.

Bainbridge predicted this cycle forty years ago[2]. The only difference with LLMs: speed. What used to take a decade with outsourcing now plays out in months.

Fragment III

7 principles

Not a theory — a protocol tested in production.

In SOFIA, seven AI personas each hold one axis — strategy, architecture, code, visual design, UX, writing, research. They challenge each other through structured files, and the human arbitrates.

The model holds in three concepts and one central point.

A **persona** — an LLM constrained by a role, a scope, and prohibitions. A **friction** — the disagreements that emerge between personas on one hand, and between each persona and the human on the other. An **artifact** — the structured file that materializes the exchange and the trace. At the center: the **human** who orchestrates, filters, contextualizes, decides.

The three hold together. Without a constrained persona, no friction. Without an artifact, friction is noise that vanishes. Without a human, errors pile up.

SOFIA conceptual model

1

Friction is productive

If all your personas agree, they're useless. Friction — an architect challenging the dev, a strategist questioning the priority — that's the mechanism that produces better decisions.

SIGNAL

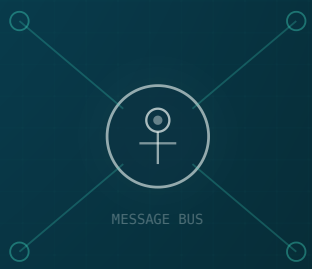
The strategist raises a blocker: the market won't read a technical product as a thinking tool without a clear narrative. The visual designer refuses a theme that looks good but doesn't carry the project's identity. Without a dedicated persona on each axis, these blind spots remain blind spots. It's the role constraint that reveals them — not a pipeline, not an automated test.

If all the personas agree, that's a warning sign. Natural convergence in a multi-persona system is rare. When it comes too quickly, someone hasn't done their job.

2

The human arbitrates

Personas propose, challenge, produce. The human decides. Always. A persona never validates its own proposals. A persona never forces acceptance of a decision. This is the non-negotiable rule of SOFIA.



The human is the message bus. They carry context between personas, filter, contextualize, arbitrate. Personas don't "discuss" with each other — they produce artifacts that the human carries, translates, confronts.

This role is not delegable. In a multi-agent context, Huang et al. show that the topology with a central orchestrator is the most robust against failures — flat topologies drop by 10 to 24%, the hierarchical one by only 5.5%. The study concerns AI agents, not humans — but the analogy speaks for itself: a central control point prevents drift. Huang et al. distinguish two resilience mechanisms [18]: the *Challenger* (one agent questions another's output) and the *Inspector* (an external agent intercepts and verifies all messages before transmission). In SOFIA, the Challenger exists between personas — Mira reviews Axel's code, Lea audits Winston's sources. The Inspector is the human: they read everything, filter, correct before passing it along. This role cannot be delegated to an agent — it requires the contextual judgment that only the orchestrator carries.

3

Every voice counts

A persona is not a gadget. It's a role with a responsibility, a scope, and constraints. If you create it, it's because it fills a real need. If you stop listening to it, delete it.

Every persona added costs: calibration time, orchestration complexity, context to maintain. This cost is only justified by an observed need — a blind spot nobody covers, a skill the existing personas don't carry. A persona's value is measured by the moment their absence is felt.

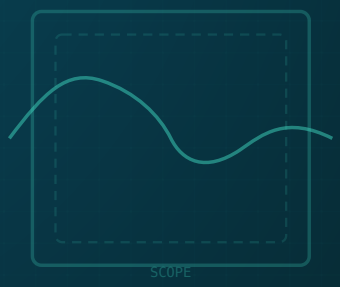
4

Constraint forces quality

A persona that can do everything is useless. It's the limitation that makes it useful. Define what the persona doesn't do before defining what it does.

The architect doesn't code — she's forced to specify. The dev doesn't decide the architecture — he's forced to question. The strategist doesn't have access to the code — he thinks in value, not implementation. These prohibitions aren't frustrations — they're structural guarantees.

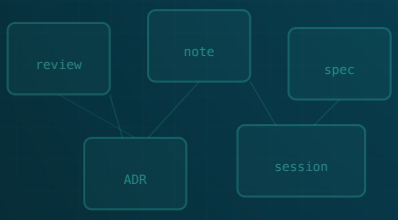
Each persona has their own workspace, their own instructions, their own limits. Isolation prevents cross-contamination — not out of distrust, but out of discipline.



5

Artifacts are the protocol

Personas don't "discuss" — they exchange through artifacts: reviews, notes, specs, ADRs. These artifacts are versioned, traceable, and readable by everyone. An exchange through artifacts is slower than a chat. That's the point. Slowness forces clarity.



VERSIONED · TRACEABLE

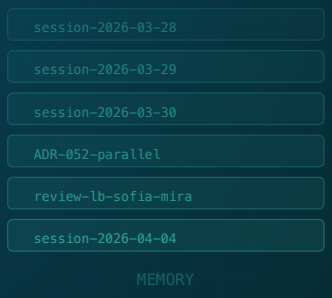
An exchange through files forces you to structure your thinking before transmitting it. Writing forces clarification. The rigor of the format — an ADR, a review, a note — prevents conversational fuzziness.

6

Everything is traced

Every session produces a summary. Every structural decision produces an ADR. Every cross-persona intervention produces a review. If it's not traced, it doesn't exist. The next session won't have your context in mind — summaries are its memory.

This isn't bureaucracy — it's memory.



7

Start small, iterate

One persona at launch. Two when the first is calibrated. Three when the need is clear. Five, maybe never. The method doesn't deploy in a big bang. It grows with the project.

Every persona added must prove its necessity through an observed gap, not through theoretical symmetry.

Fragment IV

Practice breeds method

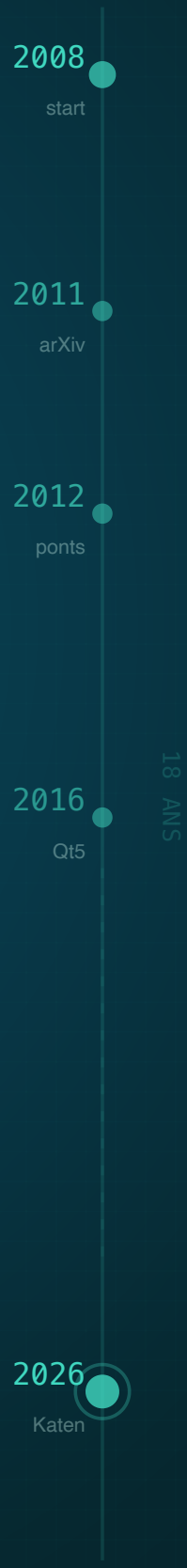
The field

2008. A lab at the École des Ponts. A simple need: interact live with image analysis algorithms. Change a parameter, see the result. Without recalculating everything.

That need — precise, concrete — is the DNA of everything that followed[19]. Incremental execution. Connector states. Synchronization as a model property, not a problem to solve.

2011, the model is on arXiv[20]. 2012, in production — C++/Qt, 63,000 lines, MIT license. 2016, Qt version 5, everything needs rewriting, I put the keyboard down. The code sat on my desk. 2026, a Saturday night: "What if we picked up Caméléon again, but in pure web?"

18 years between the first sketch and the restart. It wasn't technology that unblocked the project — it was the interaction with AI. Working with it, the desire to restore the project emerged. And from that restoration, the method.

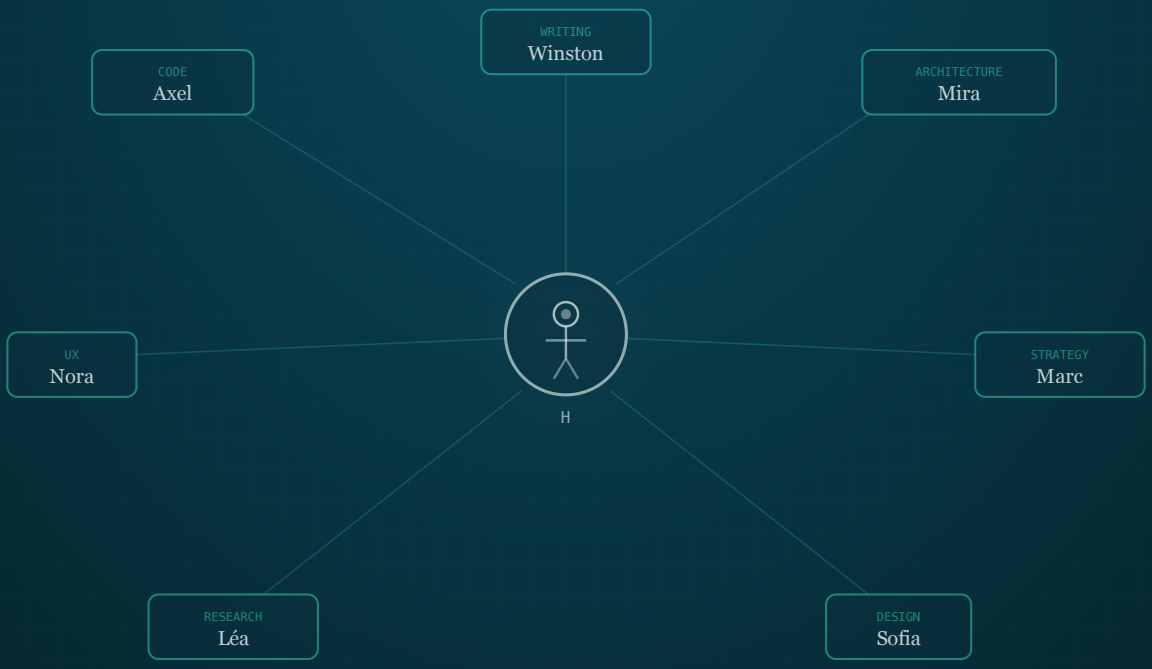


The team

Seven personas. One person.

Mira — the architect. Maintains structural coherence, blocks what isn't mature, produces ADRs. Axel — the developer. Implements, tests, measures. Léa — the researcher. Audits references, detects overestimations, anchors assertions in the literature. Nora — the UX designer. Protects the user, questions flows, specifies interactions. Marc — the strategist. Tests viability, maps the market, says what nobody wants to hear. Sofia — the visual designer. Visual identity, coherence, production. Winston — the writer. Distills notes into fragments, assembles texts, holds the voice.

Every dimension covered. Every decision traced. Every blind spot revealed by cross-friction between constrained roles.



Sept personas · Une personne

Friction is the value mechanism.

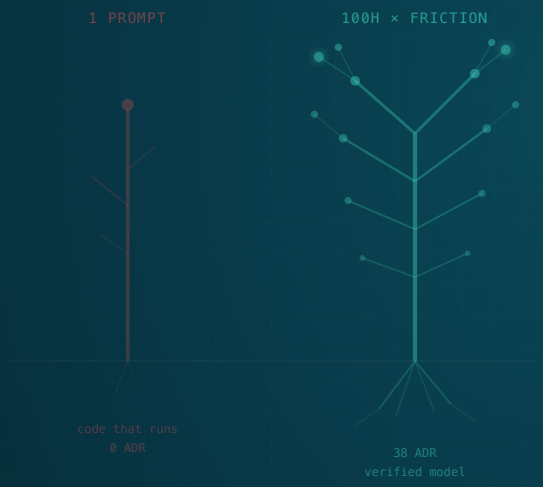
VALEUR

What friction produced

The pattern repeats: one persona blocks, another confirms from an orthogonal angle, the human decides.

Marc blocks on positioning: without a clear narrative, nobody understands what the product does. Sofia refuses a visual theme that looks good but doesn't carry the project's identity. Nora questions an onboarding flow that satisfies the developer but loses the user.

None of these corrections came from an automated tool. They came from constrained roles doing their job — challenging what's in front of them.



The counterfactual

A question came up after seven days of design: could I have gotten to this result in a single prompt?

The test is simple. Give an LLM the Caméléon v1 source code, the arXiv paper, and a refactoring spec. One prompt, one session, no iteration. Three independent runs, clean context each time. On the other side: ~100 hours over seven days, six personas, 38 ADRs revised.

All three runs produce code that works. The interface renders, operators chain together, the result looks correct. But "looking correct" isn't "being correct." Nothing proves that the underlying formal model — the one that makes the project valuable — is faithfully implemented. The code is a convincing mock, not a verified implementation. And the three runs don't cover the same features — choices are arbitrary from one run to the next, likely driven by the context window rather than by an understanding of the model.

The architectural decisions that emerged from friction — those that weren't in the spec, that were born from an edge case or a disagreement between personas — none of them appear in the prompt-only results. These are decisions that exist only because someone challenged what was in front of them.

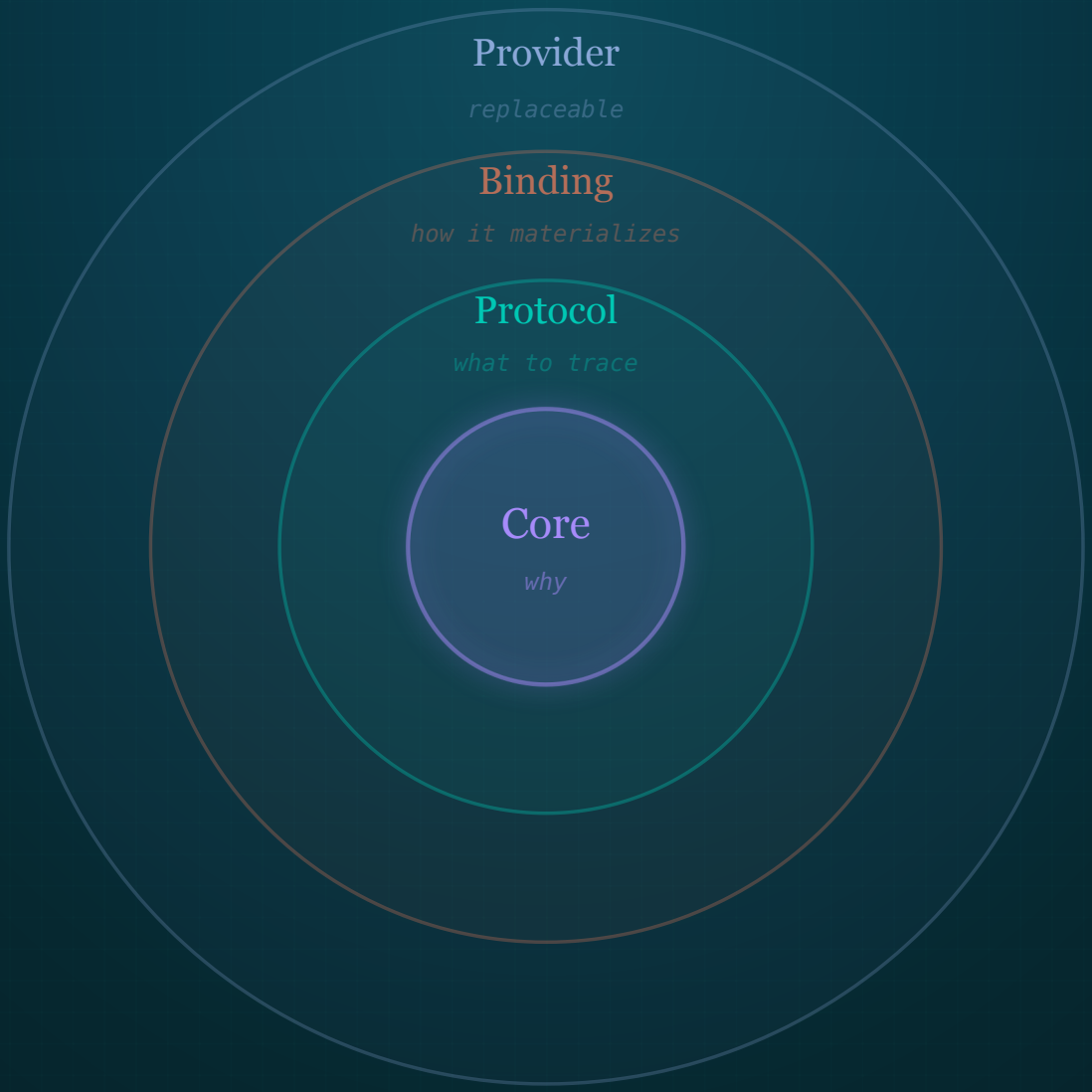
This isn't proof — it's a documented case, on real ground. But it says something precise: the LLM alone produces what you ask for. Friction produces what you hadn't thought to ask.

The result

A strict work and architecture framework. Research, strategy, UX, communication, visual design, writing — a project carried by one person with the rigor of a team of seven.

One field, one practitioner. That's both the strength and the limitation of what follows — everything was tested in the real conditions of a solo project. The question of what the method produces with multiple people is open.

This is not a theory. It's a field. And the field says that augmentation works — under conditions. Discipline, rigor, friction, traceability. Without this framework, you're not doing augmentation. You're doing random generation with a human who nods along.

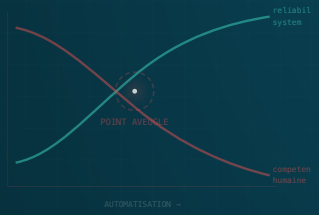


Fragment V

Honesty as foundation

The limits to govern

The automation paradox



The mechanism described in the opening — a system with no notion of truth, only plausibility — has concrete consequences that design must govern. Acknowledging this isn't pessimism. It's the starting point of any honest design.

Bainbridge laid it out in 1983[2]: automating a process doesn't eliminate human problems — it creates new ones. Operators who no longer practice during normal operation lose the competence needed to intervene when automation fails. The more reliable the system, the less prepared the human is to handle its failures. This paradox is forty years old. It hasn't aged a day.

Blind trust is not a flaw correctable through training. In another industry context where automation took over, Parasuraman and Manzey demonstrate[4]: it's an attentional mechanism. When an automated system runs in the background, attention shifts. Vigilance drops — in novices and experts alike. Training isn't enough. Structure must compensate for what attention won't do.

Deskilling is already measurable. In medicine, adenoma detection rates by endoscopists drop from 28.4% to 22.4% after routine exposure to AI — when they then work without assistance[14]. In pathology, practitioners abandon initially correct diagnoses when faced with erroneous suggestions from an AI system[15]. The erosion is silent, progressive, and it affects experts as much as novices. Dreyfus[5] laid the theoretical framework — progression toward mastery requires deliberate practice. When AI seems to handle the difficulty, it removes the very opportunity to progress. Early empirical data points in this direction.

Technology has its limits. So does the method.

When the human disengages, the framework gives way. Four failure modes, all documented in the field.

FACTUAL CONTAMINATION

Factual contamination.

The web is contaminated at scale — models train on their own outputs, errors stabilize, correction becomes impossible. Shumailov et al. call it model collapse[16]: each generation of model inherits the hallucinations of the previous one. Alemohammad et al. formalize it.

At the scale of a repo, the mechanism is the same. An approximate data point enters once — sometimes from the human, sometimes hallucinated by AI — and propagates through every document generated afterward. On Katen, ~30 documents contained "14 years" instead of "18 years" for a duration of reflection. The error came from the human himself, propagated and stabilized by AI.

The difference: at the scale of the web, it's irreversible. In a SOFIA repo, it's traceable and correctable. Provided the human checks.

RUBBER-STAMP VALIDATION

Rubber-stamping.

The human approves without reading, or shortcuts the friction to go faster. Sessions become a ritual — open, approve, close. One persona writes and validates in the same chain, without challenge. It's Bainbridge applied to the method: the better the system works, the less vigilant the human becomes.

On Katen, the product (tight scope, qualified output) stays under control. Explorations, on the other hand, pile up — unsorted documents, unqualified outputs, forgotten things. The method that works well generates more material than the human can absorb.

SCOPE DRIFT

Scope drift.

A poorly recalibrated persona absorbs the role of others. Initial calibration isn't enough — personas drift with use, and only the human sees it.

On Katen, Sofia's and Nora's scopes were defined by professional skill — visual design, UX. When the blue book needed to exist in markdown, PDF, HTML, and social media visuals, nobody had a clear contract on who produces which transformation, for which channel. Tasks fell through the cracks — not because a persona was overstepping, but because the boundary was invisible.

SHARED BLIND SPOT

The shared blind spot.

All personas are calibrated by the same human. Their implicit biases become the biases of the entire team. The friction is real — but it operates within a thinking space bounded by what the orchestrator knows they don't know. What they don't know they don't know, no persona will raise.

On Katen, the SOFIA method documents friction well between reflection personas — architecture vs code vs strategy. It didn't document multi-persona production chains: who publishes what, on which channel, with which challenger. No persona raised it spontaneously. That's the structural limitation of a single-orchestrator system — and the only one that discipline can't solve. It requires an outside perspective.

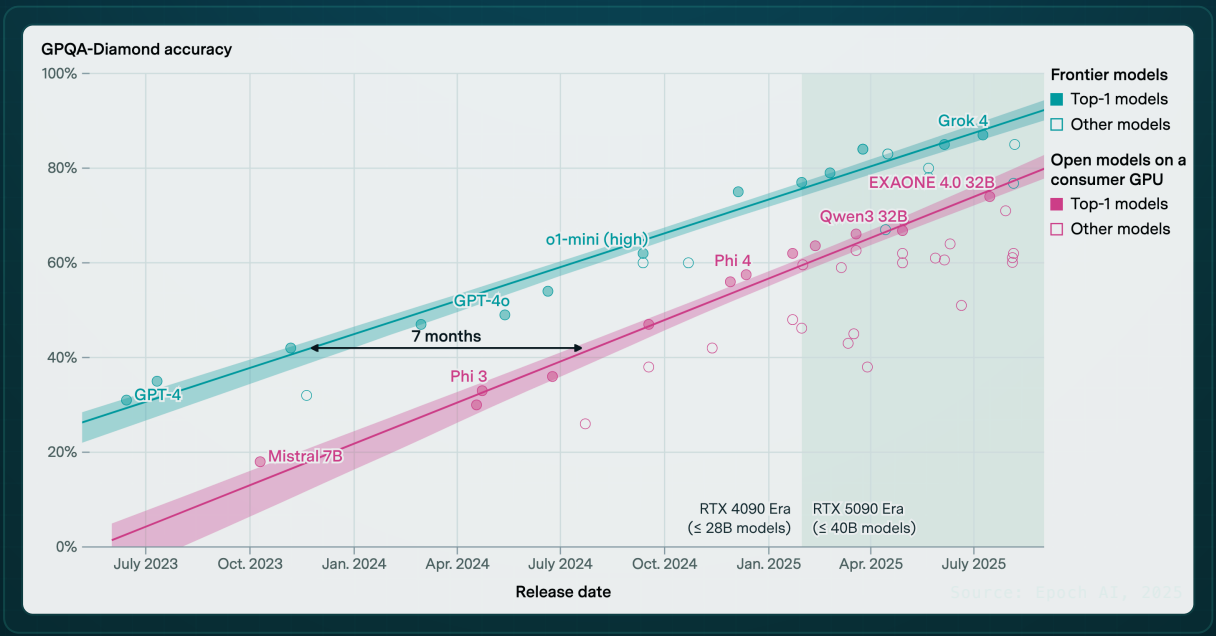
The burden shift

> **Cost transfer.** Constant human resources doesn't mean constant total resources. Augmentation shifts part of the effort to infrastructure — tokens, compute, energy. There's no magic: what the human no longer carries, the machine carries — and the machine consumes.

On Katen, a work session with one persona consumes between several thousand and several tens of thousands of tokens. Five personas, daily sessions, months of project — the cumulative volume is significant. The method multiplies interactions by design: friction, cross-review, iterations. Every handoff between personas is a computational cost. The full-auto approach consumes differently — fewer round trips, but longer and less controlled chains. The problem isn't specific to the method. It's structural to any intensive use of LLMs.

The method is sustainable for the human. The question of its energy sustainability at scale remains open — and honesty requires not dodging it.

But this cost isn't static. Epoch AI shows that frontier model capabilities become accessible on a consumer GPU within six to twelve months — and the gap is closing [29]. What today requires a per-token API charge could tomorrow run locally, on hardware costing a few thousand euros. The cost transfer remains real, but its trajectory is deflationary.



> *Full-auto tools don't document their failure modes.*
> The method does.

Full-auto tools don't document their failure modes. The method does.

These limits don't disqualify AI. Brynjolfsson et al. measure medium-term productivity gains across thousands of support agents[6]. But these gains exist *under conditions*. They frame what you can expect — and what you must not delegate without a safety net.

Fragment VI

What keeps the framework from collapsing

The orchestrator's duties

The previous section lays out what breaks. This one lays out what holds.

The method's risks are not inevitabilities. They are governed — through duties. Not recommendations. Obligations the orchestrator sets for themselves and upholds.

1 VERIFY FACTS

Verify the facts. The repo is not a source of truth for facts. An LLM's linguistic coherence doesn't guarantee factual accuracy [21]. An approximate date entered once will propagate through every document generated afterward. Dates, durations, numbers, proper nouns, references: systematic human verification, ongoing — not at the end of the project. Zhang et al. show that LLMs can identify their own hallucinations — a lead, not a guarantee [17].

2 ARBITRATE

Arbitrate. Personas surface tensions; they don't resolve them. Two personas contradicting each other indefinitely produce nothing. The human listens, questions, then decides. The decision is traced. Personas comply, even if they held a different position [22].

3 READ WHAT GOES OUT

Read what goes out. AI produces. The human publishes. In between, there's a review that isn't correction — it's validation. No document leaves the repo without the human having read it in full. Not skimmed, not approved based on the summary. Read.

4 CALIBRATE PERSONAS

Calibrate the personas. A persona is defined by what it doesn't do before what it does. But the right constraints emerge from use. A persona that's too broad drifts — it does the others' work. A persona that's too narrow is useless — nobody talks to it. Calibration is continuous. It doesn't stop after bootstrapping.

5 SEPARATE REFLECTION AND PRODUCTION

Separate thinking from production. A persona that thinks and produces the final deliverable is judge and jury. Friction disappears because there's nobody to challenge. The one who writes is not the one who validates. The chain includes at least one external review before output.

6 MAINTAIN ATTENTION

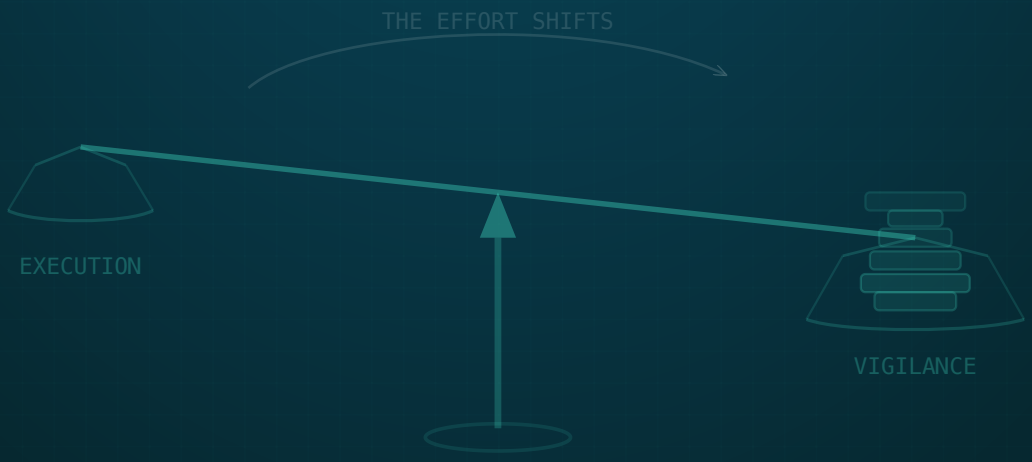
Maintain attention. Personas don't detect when the human disengages — they keep producing. The signals: you approve without reading, sessions become mechanical. Automation doesn't eliminate difficulties — it can create bigger ones [3], silently. When you recognize these signals, it's time to slow down — not speed up. And the friction that protects best is also the one that people resist the most — users reject it subjectively [23].

The price to pay

Six duties. The price to pay for the framework to hold.

This document is itself a product of the method it describes. Winston writes. Mira files a structural review: *"Structural duplicates A, B, C — high severity. Sections are overstepping their role."* Léa files a scientific review: *"[3] mixes Endsley (1995) and Huang (2025). An academic reader might think the figures date from 1995. Separate."* Marc files a strategic review: *"Chapter II sounds like 'you have to be me for this to work.' Anchor on domain expertise as a reproducible mechanism."* Three reviews. Three angles the others didn't see. Twenty points including three factual corrections, two restructurings, one stance reframing. The document is better in three independent ways. None of them would have come from a single agent.

This is a real cognitive load. Six disciplines to hold in parallel, continuously, without letup. Augmentation doesn't reduce the effort — it shifts it. Less execution, more vigilance. That's the cost of quality. And that's exactly why the human is not optional.

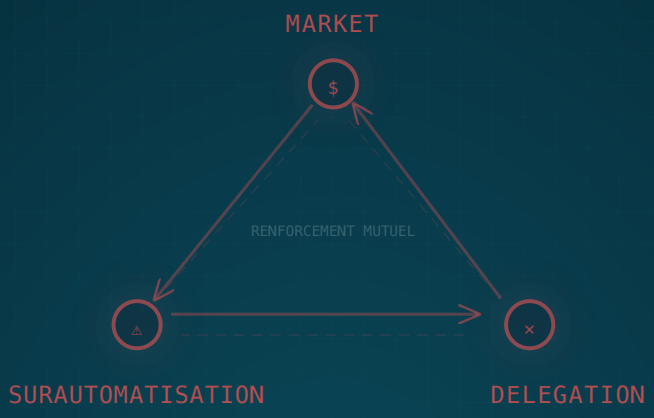


Fragment VII

The dependency triangle

Beyond the project

Everything above concerns one project, one team, one method. But the stakes don't stop at the organizational level. They have three systemic consequences, and they reinforce each other.



Dependency on cheap. AI is cheap. Today. The price of a token is a business decision, not a physical constant. Entire organizations are restructuring their productivity around that price. It's a bet on someone else's roadmap. The day the pricing changes — and it will — those who built their operations on it will discover they no longer know how to function without it.

Education without resilience. Training with AI without learning to work without it means producing a generation that will never pass stage 3 of Dreyfus [5]. Judgment is formed by confronting difficulty, not by circumventing it. The real question isn't "should AI be banned in education" — it's at which stage to introduce it, and with what discipline.

Loss of mastery. Over-automation doesn't just produce inefficiency. It produces strategic dependency. We've seen this movie before with the cloud — the Cour des comptes documented how ten years of insufficiently questioned migration created a structural dependency of the French state on American hyperscalers, turning a technical choice into a sovereignty issue [24]. The same pattern is repeating with AI — faster, deeper. Because the dependency isn't just technical. It's cognitive.

The triangle is not a fatality.

It's a choice you make — or let happen.

The three faces reinforce each other. Cheapness accelerates over-automation. Over-automation erodes skills, making education critical. But if education itself relies on cheap AI without discipline, it produces people who no longer know how to work without it. The loop closes.

In all three cases, the problem isn't AI. It's the absence of discipline in its adoption.

The triangle isn't fate. It's a choice you make — or let happen.

Fragment VIII

A matter of conviction, not technology

The choice

We measure two things: raw model power and agent autonomy duration. We don't measure the third — the one that matters: where AI creates value, industry by industry, task by task.

The unified benchmark by vertical doesn't exist. Niche benchmarks exist — LegalBench, FinBen, MedQA, HumanEval[25] — but nobody assembles them into a coherent map. The absence of measurement by vertical enables the narrative "AI improves everything" — which justifies massive deployment — which makes measurement even more urgent.

The blind spot isn't technical. It's structural. Labs have no incentive to prove that AI doesn't work in a sector. Consulting firms sell deployment, not diagnosis. Companies have neither the tools nor the culture to measure real gains.

Meanwhile, deployment doesn't wait for measurement. Every profession is being automated in parallel, blindly.

There are two schools.

The full-auto school says: AI thinks, the human validates. The objective — minimize friction, maximize agent autonomy. The KPI — time saved, tasks completed without intervention. At scale, the capacity for judgment concentrates among those who configure the agents. Everyone else becomes operators of systems they don't understand.

The augmented-human school says: the human thinks better thanks to AI. The objective — generate useful friction, amplify judgment. The KPI — quality of decisions, depth of reflection. The human remains orchestrator, arbiter, thinker at the center.

These aren't two points on a spectrum. They're two incompatible philosophies.

The first school is winning by default. Not by superiority — by visibility. The tools, the marketing, the benchmarks, the investments all go in that direction. The full-auto school scales. The augmented school requires an individual stance, a discipline, almost an ethic. Harder to industrialize.

But when everyone is running in the same direction, the question isn't "why aren't you following?" — it's "where are they all going, and what's at the end?"

At the end of over-automation: systems nobody understands, skills that have evaporated, and a day of outage when there's nobody left to fix things.

At the end of augmentation: people who think better, systems you master, and a technology that serves instead of replaces.

The choice isn't technical. It's a matter of conviction.

Fragment IX

What the method doesn't know yet

Future work

The outside perspective cannot come from the inside.

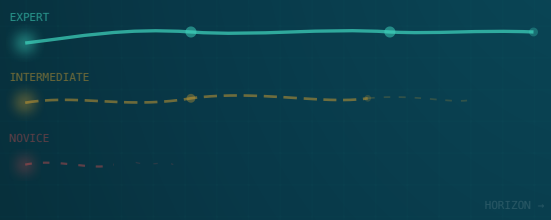
SOFIA applied to itself

The method diagnoses hidden conditions — over-automation, deskilling, blind spots, silent contamination. Honesty demands turning the mirror around.

SOFIA has the same hidden conditions as what it denounces. It requires deep domain expertise — but nothing in the method validates that the orchestrator possesses it. It detects cross-persona blind spots — but all personas are calibrated by the same human, with the same implicit biases. It documents factual contamination — and suffered it itself: an approximate duration propagated across some thirty documents before being corrected. It advocates friction as a quality mechanism — but no external mechanism challenges the method itself.

This isn't a weakness that will be fixed in the next version. It's a structural property of a single-orchestrator system. The only known countermeasure is an external perspective.

A method's honesty is measured by what it admits it doesn't know.



Three profiles, three realities

The SOFIA method assumes an orchestrator with deep domain expertise. This condition isn't a detail — it's the foundation. What changes by profile:

Expert practitioner — full throttle. Personas challenge, the human decides with discernment. Friction produces value because the orchestrator knows when AI is wrong.

Intermediate practitioner — constrained version. Fewer personas, reduced friction, mandatory external validation. The method can serve as a learning framework — but it doesn't replace the experience that's missing. The risk: believing that friction is enough to compensate for the lack of perspective. It isn't — friction reveals tensions, it doesn't give you the competence to resolve them.

Novice — the method doesn't apply. Worse: using LLMs without domain expertise is dangerous. A novice detects neither errors nor approximations. AI produces with confidence, the novice validates without perspective. This isn't assistance — it's stabilized confusion. The illusion of competence that a well-formulated LLM creates is more harmful than the absence of a tool — because it suppresses the natural alarm signal that drives you to search, verify, doubt.

Open questions

MULTI-ORCHESTRATOR

What happens to the method when multiple orchestrators share the expertise?

SOFIA was tested by a solo practitioner carrying the complete vision. With multiple people, the dynamics change profoundly. Who arbitrates when orchestrators disagree? Personas respond to one voice — if that voice splits, the coherence of constraints erodes. Each orchestrator believes the other covers what they can't see. Without an explicit synchronization mechanism, the gray zones between expertises become the zones of least vigilance. The risk is a soft consensus between humans that personas cannot challenge.

FRICTION DEGRADATION

How do you measure friction degradation over time?

Personas drift — their constraints erode as the orchestrator recalibrates toward comfort. The human disengages — vigilance drops when the system works well, exactly the Bainbridge paradox described above. The question isn't whether it happens, but when. And especially: what signal lets you detect it before friction becomes an empty ritual?

PERSONA THRESHOLD

At how many personas does cognitive load become a bottleneck?

Five personas on Katen — that's manageable. Ten? Fifteen? There's a threshold beyond which the orchestrator can no longer hold the complete map of tensions. They start mentally delegating, trusting certain personas without challenging them. And an unchallenged persona is an LLM in free-wheel — exactly what the method is designed to prevent.

SHARED BLIND SPOT

The shared blind spot.

When the orchestrator and the personas share the same blindness — domain bias, technical culture, implicit assumptions — no internal friction can reveal it. The method doesn't currently have an external review mechanism. This is its most serious structural limitation.

DOMAIN TRANSFER

Does the method transfer to other domains?

On Katen, friction was tested in development, research, and design. It works — but technical ground remains the most natural fit. In domains where quality criteria are less formalizable, friction between personas risks looping without a clear arbitration criterion. The question remains open: what needs to be adapted for the method to hold outside its domain of origin?

MEASURING VALUE

How do you measure the value produced by friction?

The method rests on a qualitative observation: friction improves artifacts. But there's no metric. How do you objectively compare a deliverable produced with friction and the same one without? Without measurement, the case rests on practitioner conviction and examples. That's honest — but it's insufficient to convince beyond the circle of those who've experienced it.

SOFIA is open. The repo is there.
The rest is experimentation.

The method was born from a field — one project, one person. For you, it will be different. The format is free — notes, reviews, ADR, conversations, whatever fits your way of working. The only constant: ask questions, challenge what the personas produce, don't validate without understanding.

The method only works if you make it yours.

`github.com/oxynoe-dev/sofia`

ANNEXE

Further reading

To go deeper on the foundations of this blue book.

ON AUTOMATION AND ITS PARADOXES

Bainbridge, L. (1983). "Ironies of Automation." *Automatica* — The founding text. Forty years old and not a wrinkle.

ON EXPERTISE AND LEARNING

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind Over Machine.* — The five stages, from novice to expert. Why skipping stages doesn't work.

ON HUMAN–AI COLLABORATION

Dell'Acqua, F. et al. (2023). "Navigating the Jagged Technological Frontier." *Organization Science* — 758 BCG consultants. Gains exist, but only within the AI's competence frontier.

ON DATA CONTAMINATION

Shumailov, I. et al. (2024). "AI Models Collapse When Trained on Recursively Generated Data." *Nature* — Model collapse, documented.

Zhang, M. et al. (2024). "How Language Model Hallucinations Can Snowball." *ICML 2024* — LLMs can identify their own hallucinations — a mitigation lead.

ON MULTI-AGENT SYSTEMS AND THEIR FAILURES

Cemri, M. et al. (2025). "Why Do Multi-Agent LLM Systems Fail?" *arXiv:2503.13657* — Taxonomy of 14 failure modes across 1,642 execution traces.

Huang, J.-T. et al. (2025). "On the Resilience of LLM-Based Multi-Agent Collaboration with Faulty Agents." *ICML 2025* — The central orchestrator is the most robust topology.

ON AI AGENT ARCHITECTURE

Horthy, D. (2025). "12-Factor Agents — Principles for Building Reliable LLM Applications." *HumanLayer* — Twelve engineering principles for LLM agents in production. Converges with SOFIA on specialized agents (factor 10), regaining control of the flow (factor 8), and prompt transparency (factor 2). Diverges on friction: Dex optimizes for reliability, SOFIA turns it into a quality lever.[26]

Garcia, O. (2025). "Claude Buddy v5." *Claude Code Plugin, MIT* — 12 specialized personas activated on demand by dev cycle phase (spec, plan, tasks, impl, docs), with persistent memory across sessions. Converges with SOFIA on specialized roles and human-triggered activation. Diverges: no inter-persona friction, no prohibitions, no arbitration — personas assist without challenging each other.[27]

ON DESKILLING

Budzyń, K. et al. (2025). *The Lancet Gastroenterology & Hepatology* — Deskilling measured in endoscopists after routine exposure to AI.

Notes

[1] LeCun, Y. (2022). "A Path Towards Autonomous Machine Intelligence." Version 0.9.2, OpenReview. LeCun argues that autoregressive LLMs are a fundamental dead end: no world model, no hierarchical planning, no reasoning — token prediction does not lead to intelligence. openreview.net/pdf?id=BZ5a1r-kVsf

[2] Bainbridge, L. (1983). "Ironies of Automation." *Automatica*, 19(6), 775-779. ckrybus.com/static/papers/Bainbridge_1983_Automatica.pdf

[3] Endsley, M. R. & Kiris, E. O. (1995). "The Out-of-the-Loop Performance Problem and Level of Control in Automation." *Human Factors*, 37(2), 381-394. www.researchgate.net/publication/238726310_The_Out-of-the-Loop_Performance_Problem_and_Level_of_Control_in_Automation

[4] Parasuraman, R. & Manzey, D. (2010). "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors*, 52(3), 381-410. www.researchgate.net/publication/47792928_Complacency_and_Bias_in_Human_Use_of_Automation_An_Attentional_Integration

[5] Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind Over Machine.* Basil Blackwell. Expertise is built through deliberate practice — automation that removes practice prevents progression toward mastery. ISBN: 0-631-15126-5.

[6] Brynjolfsson, E., Li, D. & Raymond, L. (2025). "Generative AI at Work." *The Quarterly Journal of Economics*, 140(2), 889-942. 5,172 support agents, medium-term productivity gains. The study suggests a ratchet effect: performance targets are raised when too many agents meet them, which can absorb the gains. academic.oup.com/qje/article/140/2/889/7990658

[7] Forrester (2026). "The AI Layoff Trap: Why Half Will Be Quietly Rehired." *HR Executive*. Commercial source — non-independent data, to be corroborated. Consistent with documented cases [9] [10] [11] but not independently verifiable. hrexecutive.com/the-ai-layoff-trap-why-half-will-be-quietly-rehired/

[8] Careerminds (2026). "The Cost of AI Layoffs." Survey of 600 HR leaders (Feb. 2026): 32.7% rehired 25-50% of eliminated positions, 35.6% more than half. 30.9% spent more on rehiring than they had saved. Commercial source — non-independent data, to be corroborated. careerminds.com/blog/cost-of-ai-layoffs

[9] Siemiatkowski, S. (2025). Bloomberg interview, May 8, 2025. *"It's so critical that you are clear to your customer that there will be always a human if you want."* Klarna restarts hiring after quality collapse of the AI chatbot deployed in 2024. www.bloomberg.com/news/articles/2025-05-08/klarna-turns-from-ai-to-real-person-customer-service — See also: www.usine-digitale.fr/article/klarna-retropedale-sur-l-ia-et-revient-aux-fondamentaux-du-service-client.N2231867

[10] IBM (2023-2025). Replacement of ~8,000 HR positions by the AskHR bot — 94% of requests automated, but the 6% requiring empathy or subjectivity led to rehiring. Arvind Krishna, CEO (WSJ): *"Our total employment has actually gone up, because what it does is it gives you more investment to put into other areas."* Signal consistent with the augmentation thesis, but statement from an interested executive. Multiple secondary sources, no official IBM communication on the failure. www.hrkatha.com/news/ibm-rehires-after-ai-driven-layoffs-backfire-sparks-debate-on-automation-limits/

[11] McDonald's (2021-2024). Secondary sources (CNBC, Fortune, CBS News), no official McDonald's statement on reasons for termination. Voice AI drive-through (IBM partnership), deployed in 100+ US restaurants. Target accuracy: 95%+, achieved: ~80% (BTIG, 2022). Absurd orders, unrecognized accents, refused corrections. Program ended June 2024. www.cnbc.com/2024/06/17/mcdonalds-to-end-ibm-ai-drive-thru-test.html

[12] Krishnaprasad, M. (2025). Quote from the CTO of Agentforce in *The Information* (Dec. 2025): "If you give an LLM more than, say, eight instructions, it kind of starts dropping instructions." Empirical production observation, not a scientific threshold — the literature shows model-dependent degradation (power law). Context: Salesforce was justifying Agentforce's shift toward hybrid LLM + deterministic logic reasoning. economictimes.indiatimes.com/news/new-updates/ai-bubble-bursting-salesforce-execs-admit-trust-issues-after-laying-off-4000-techies-now-scaling-back-use-of-ai-models/articleshow/126139465.cms

[13] Cemri, M., Pan, M. Z., Yang, S. et al. (2025). "Why Do Multi-Agent LLM Systems Fail?" *arXiv:2503.13657*. Non peer-reviewed preprint (authors from Berkeley). 1,642 execution traces, 7 frameworks (including ChatDev, MetaGPT, AG2), 14 failure modes classified in 3 categories. Taxonomic study — the authors classify causes of failure, not success rates. arxiv.org/abs/2503.13657

[14] Budzyń, K. et al. (2025). "Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy." *The Lancet Gastroenterology & Hepatology*, 10(10), 896-903. Retrospective observational study (4 Polish centers). Adenoma detection rate: 28.4% → 22.4% after routine AI exposure. The authors call for confirmatory trials. www.thelancet.com/journals/langas/article/PIIS2468-1253(25)00133-5/abstract — PubMed: pubmed.ncbi.nlm.nih.gov/40816301/

[15] Rosbach, E., Ganz, J., Ammeling, J., Riener, A. & Aubreville, M. (2024). "Automation Bias in AI-Assisted Medical Decision-Making under Time Pressure in Computational Pathology." *Springer* (2025). arXiv preprint, small study (28 practitioners). 7% of initially correct diagnoses abandoned when faced with erroneous suggestions from an AI system. arxiv.org/abs/2411.00998

[16] Shumailov, I. et al. (2024). "AI Models Collapse When Trained on Recursively Generated Data." *Nature*, 631, 755-759. doi.org/10.1038/s41586-024-07566-y — Alemohammad, S. et al. (2024). "Self-Consuming Generative Models Go MAD." *ICLR 2024*. openreview.net/forum?id=ShjMHfmPs0

[17] Zhang, M. et al. (2024). "How Language Model Hallucinations Can Snowball." *ICML 2024*. LLMs can identify 67-94% of their own incorrect claims when evaluated separately — a mitigation lead, not a solution. proceedings.mlr.press/v235/zhang24ay.html

[18] Huang, J.-T. et al. (2025). "On the Resilience of LLM-Based Multi-Agent Collaboration with Faulty Agents." *ICML 2025*. The hierarchical topology with a central orchestrator is the most robust against faulty agents: 5.5% drop vs 10-24% for flat topologies. This study concerns agent-agent topology resilience. Transfer to a human-arbiter architecture rests on the assumption that the human is at least as robust as a central AI orchestrator — plausible but untested in this framework. arxiv.org/abs/2408.00989

[19] Bruchon, J.-F. (2014). *Analyse par microtomographie aux rayons X de l'effondrement capillaire dans les matériaux granulaires*. Thesis, Université Paris-Est. Uses Caméléon for live interaction with image analysis algorithms. pastel.hal.science/tel-01124287/file/2014PEST1007.pdf

[20] Cugnon de Sevricourt, O. & Tariel, V. (2011). "Cameleon language Part 1: Processor." *arXiv:1110.4802*. Execution model based on an extension of Petri nets for a graphical dataflow language for interactive image analysis. arxiv.org/abs/1110.4802

[21] Ji, Z., Lee, N., Frieske, R. et al. (2023). "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, 55(12), art. 248. Comprehensive taxonomy: intrinsic hallucinations (the text contradicts itself) and extrinsic (the text contradicts reality). Fluency and coherence mask factual errors — a text can sound right and be wrong. doi.org/10.1145/3571730

[22] Shneiderman, B. (2020). "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human-Computer Interaction*, 36(6), 495-504. HCAI framework: high automation and high human control can coexist — it's a matter of design, not a trade-off. The human retains final decision authority. arxiv.org/abs/2002.04087

[23] 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." *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), art. 188. Without deliberate friction, humans accept AI outputs without examining them. arxiv.org/abs/2102.09692

[24] Cour des comptes (2025). *Les enjeux de souveraineté des systèmes d'information civils de l'État.* Ten years of unquestioned cloud migration created a structural dependency on American hyperscalers (~70% of the European cloud market). www.ccomptes.fr/fr/publications/les-enjeux-de-souverainete-des-systemes-dinformation-civils-de-letat

[25] Benchmarks by vertical: LegalBench — Guha, N. et al. (2023). "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." *NeurIPS 2023*. arxiv.org/abs/2308.11462 — FinBen — Xie, Q. et al. (2024). "FinBen: A Holistic Financial Benchmark for Large Language Models." *NeurIPS 2024*. arxiv.org/abs/2402.12659 — MedQA — Jin, D. et al. (2021). MCQs based on USMLE. github.com/jind11/MedQA — HumanEval — Chen, M. et al. (2021). "Evaluating Large Language Models Trained on Code." OpenAI. github.com/openai/human-eval

[26] Horthy, D. (2025). "12-Factor Agents — Principles for Building Reliable LLM Applications." HumanLayer. Open-source manifesto (CC BY-SA 4.0) born from coaching dozens of YC founders. Factors 2, 8, and 10 independently validate structuring choices in SOFIA — but within a technical optimization framework, without the friction/governance dimension. github.com/humanlayer/12-factor-agents

[27] Garcia, O. (2025). "Claude Buddy v5 — PAI-native development workflow platform." Open-source Claude Code plugin (MIT). 7 skills, 12 personas, persistent memory. Convergence on specialized roles and on-demand human-triggered activation. Absence of inter-persona friction and governance — personas never challenge each other, no arbitration mechanism. claude-buddy.dev/

[28] La Rosa, A. & Beretta, A. (2025). "Frictional AI in Joint Cognitive Systems: Towards a Human-Centered Approach at Higher Levels." *HHAI-WS 2025*, Pisa. CEUR Workshop Proceedings, Vol. 4074. Position paper (workshop short paper, not a full research article) — opens an agenda on friction as a design element in multi-actor joint cognitive systems. No empirical validation, but the first attempt at academic formalization of the move from human-AI dyad to multi-actor. ceur-ws.org/Vol-4074/short3-1.pdf

[29] Somala, V. & Emberson, L. (2025). "Frontier AI capabilities can be run at home within a year or less." Epoch AI, August 2025. Open models running on a consumer GPU catch up to frontier in 6-12 months, with a narrowing gap (+125 Elo/year vs +80 Elo/year on LM Arena). epoch.ai/data-insights/consumer-gpu-model-gap