From AI Tools to an AI Operating Model

What I learned by pairing autonomous execution with critical review

Nick Montano

September 2026

WHITE PAPER

Contents

Executive summary

The real AI decision is how to design the work

My division of labor

Use case one: Building a revenue-intelligence system

Use case two: Turning coverage into credible outreach

Use case three: Separating public insight from private opportunity

Use case four: What a sports-betting research experiment taught me about independent review

The five-stage operating loop

Controls that keep multiple systems from reinforcing the same mistake

What leaders should measure

A practical starting point for business leaders

Conclusion

Executive summary

The most important lesson from my work with artificial intelligence is not about which platform is best. It is about how work should be divided.

I use Pokee AI and ChatGPT together across several recurring workflows: monitoring enterprise accounts, building revenue intelligence, researching contacts and displacement opportunities, developing personalized outreach, producing an eDiscovery intelligence newsletter, and testing a multi-agent sports-betting research process. The platforms have overlapping capabilities, but I give them different default responsibilities.

Pokee AI serves primarily as the execution layer. I use it to run recurring research, coordinate specialized agents, maintain operational records, and apply a common process across a large research universe. ChatGPT serves primarily as the challenge and synthesis layer. I use it to test assumptions, find gaps in the workflow, distinguish evidence from inference, improve operating instructions, and turn research into decision-ready analysis and communication.

I remain responsible for the objective, the acceptable level of risk, and any action that can affect a customer relationship, a public audience, or a financial decision.

The resulting operating loop is simple:

1. Define the decision and the evidence required.
2. Use an execution system to collect, monitor, and organize information at scale.
3. Use a separate reasoning layer to challenge the output and the process that produced it.
4. Apply human judgment before consequential action.
5. Feed what was learned back into the operating instructions.

This is not a controlled comparison of two products, and it is not an argument that every organization needs the same tools. It is a field-tested approach to a broader management problem: how to combine automation, independent review, and human accountability without confusing activity with progress.

The real AI decision is how to design the work

My early use of AI followed a familiar pattern. I asked a system to research a subject, produce an answer, and help draft the next deliverable. The work was useful, but transactional. Each request produced an output; it did not create a durable process.

That limitation became clearer as the scope expanded. Monitoring hundreds of enterprise accounts requires schedules, persistent records, entity resolution, scoring rules, and an audit trail. Deciding whether a particular signal matters to a legal executive is a different problem. The first is largely operational. The second depends on interpretation, context, and judgment.

I stopped asking, “Which AI is better?” and began asking:

- What work needs to run repeatedly?
- What work requires deeper interpretation?
- Where could an error create reputational, financial, or strategic harm?
- What evidence is sufficient to support action?
- Which decisions must remain explicitly human?

Both Pokee AI and ChatGPT can perform multi-step work. Pokee AI describes its agents as systems that plan, act, verify, and cite across tools in a persistent environment.[1] OpenAI describes ChatGPT agent capabilities as combining research, analysis, tool use, and action while keeping the user in control of consequential steps.[2] Those capabilities matter, but feature overlap does not eliminate the need for role clarity.

The more useful question is not what each platform *can* do. It is what each platform should own inside a particular workflow.

My division of labor

I use Pokee AI when breadth, repetition, and operational continuity are central to the task. Examples include scanning a large account universe, running scheduled research, coordinating specialized agents, maintaining ledgers, and applying the same instructions to many records.

I use ChatGPT when the central problem is evaluation or communication. Typical assignments include auditing Pokee AI's output for unsupported assumptions, comparing competing conclusions, identifying missing controls, revising a master prompt, or turning a body of research into an executive brief, white paper, email sequence, or decision memo.

My role is not merely to approve the final wording. I define the business objective, resolve material ambiguity, determine the acceptable level of risk, and decide whether the work is ready to influence an external action.

Layer	Default responsibility	Typical outputs
Execution layer: Pokee AI	Recurring execution and operational scale	Monitoring, agent reports, structured records, scheduled briefs, first-pass research
Challenge layer: ChatGPT	Analysis, critique, synthesis, and editorial control	Audits, revised logic, prioritized recommendations, decision-ready writing
Human owner	Objectives, risk, accountability, and action	Approval, intervention, strategic judgment, external communication

These boundaries are defaults, not technical limitations. Pokee AI can reason, and ChatGPT can execute multi-step tasks. The value comes from assigning a primary owner, requiring a

deliberate handoff when the work moves from collection to interpretation, and preserving human authority when a recommendation becomes an action.

Use case one: Building a revenue-intelligence system

The clearest example is the revenue-intelligence system I have been developing for enterprise legal technology sales.

Useful account intelligence requires more than company research. The system must monitor relevant changes, resolve company and contact identities, remember prior outreach, distinguish active prospects from researched accounts, identify the right personas, and connect external developments to a credible business hypothesis. It must also avoid claiming that a job posting, technology tag, litigation event, or leadership change proves more than it does.

I use Pokee AI as the operating layer for this work. The system includes hundreds of account records, an append-only signal ledger, account memory, scoring and decay logic, entity-resolution rules, outreach records, recurring schedules, and reporting workflows. Specialized processes generate daily intelligence, weekly market analysis, a business-development command center, a learning report, and an editorial pipeline for my newsletter.

That scale creates its own risk. A system can be busy, orderly, and still be strategically wrong. A technographic tag may be stale. A legal development may be accurate but irrelevant. A new executive may matter without creating an immediate buying motion. Several weak signals can appear persuasive simply because they are presented together.

I use ChatGPT to challenge both the reports and the machinery behind them. I provide the operating specification, relevant files, outputs, and known outcomes, then ask for an adversarial review. The assignment is not to summarize the report. It is to determine whether the evidence is strong enough to support the recommended action and whether the workflow is behaving as intended.

That review has identified issues that polished reporting can conceal: incomplete account coverage, duplicate production paths, rules documented in a specification but not enforced in a live schedule, unclear boundaries between public editorial content and private sales intelligence, and metrics that counted production without closing the loop on replies or meetings.

The revisions were structural, not cosmetic. They included clearer routing fields for high-value signals, technology watchlists, a campaign registry, a discovery-question library, negative-signal handling, production locks, and a firmer separation between newsletter content and private revenue opportunities derived from the same research.

The lesson was straightforward: better writing would not have corrected a weak operating system. The workflow needed a reviewer that could inspect its logic, state, and incentives.

Use case two: Turning coverage into credible outreach

The same division of labor applies to contact research and personalized messaging.

For a 50-account campaign, the execution layer was tasked with identifying two or three legal contacts and one information-security contact per account, revalidating employment and titles, evaluating relevant company signals, and drafting persona-specific outreach.

That creates leverage, but it also creates risk. Contact records become stale. Technology evidence can be misread. Personalization can become intrusive. A weak signal can be transformed into a confident statement about a project that may not exist.

The campaign therefore includes controls alongside the research. Instructions require contact revalidation, distinguish partner technologies from displacement targets, treat technographic evidence as a hypothesis rather than proof, apply suppression and opt-out rules, and hold messages in draft until a human approves them.

ChatGPT helps me test whether the proposed message follows from the evidence. Does the stated business problem reasonably connect to the signal? Would the recipient care about the proposed value? Is the call to action proportionate to the strength of the hypothesis? Has the draft crossed from informed relevance into unwarranted certainty?

The operating principle is:

- Automation creates coverage.
- Analysis creates relevance.
- Human approval protects the relationship.

The goal is not maximum personalization. It is a credible reason to begin a conversation.

Use case three: Separating public insight from private opportunity

My eDiscovery intelligence newsletter introduced a related design problem. The same market research can inform public thought leadership and private account strategy, but those outputs should not be collapsed into one another.

A public newsletter must be independently useful. It should explain a development accurately, resist turning every story into a sales pitch, and serve readers who may never become buyers. Private revenue intelligence has a different purpose. It asks which target accounts may be affected, which stakeholders might care, whether a partner motion exists, and whether outreach is appropriate.

The two-system model allows verified source material to support separate workflows. Pokee AI can monitor the market, maintain an editorial queue, and connect developments to accounts. ChatGPT can challenge the editorial framing, test whether a story is genuinely important, and help ensure that the published analysis remains distinct from internal sales recommendations.

One of the strongest principles to emerge from this work is: share evidence, but do not merge objectives. Thought leadership should not become disguised outbound. A commercially interesting account signal does not automatically belong in a public publication.

Use case four: What a sports-betting research experiment taught me about independent review

One of my sports-betting research experiments is far removed from enterprise sales, but it exposed the same underlying problem in a setting where results arrive quickly. I include it here as a governance example, not as a claim about betting performance.

I built a multi-agent research process in Pokee AI with separate roles for quantitative analysis, new-information review, market pricing, scenario analysis, prediction-market analysis, adversarial review, and final portfolio decisions. The roles were separated intentionally so that the agents would not simply repeat one another's conclusions.

In one full run, a candidate reached the red team and was rejected because the central thesis relied on a false premise: the market had already moved materially after new information became available, so the analysis was effectively counting information that had already been priced.

The important point was not whether that candidate ultimately won or lost. A single outcome cannot validate the reasoning that produced a recommendation. Treating every favorable result as proof would reward fragile analysis and encourage the system to learn from noise.

I then used ChatGPT to audit the broader design. The review examined probability ranges, dependence on market priors, decision thresholds, enforcement of pre-event snapshots, entity-matching errors, missing-data behavior, user-directed exceptions, and the evidence required before the model could be changed.

Three principles transferred directly to my business workflows:

1. Independent review must be able to invalidate the premise, not merely improve the presentation.
2. "No action" must be a legitimate result. A system forced to act will manufacture confidence.
3. Process quality and short-term outcomes must be evaluated separately.

A weak account signal does not become sound research because it happens to produce a meeting. A well-supported hypothesis does not become worthless because one recipient fails to reply. Learning requires enough observations to distinguish a pattern from an anecdote.

The five-stage operating loop

The model I now use can be expressed in five stages.

Define the decision

Begin with the decision the system is meant to improve. “Research my accounts” is too broad. A more useful objective is: identify the accounts with enough verified change to justify outreach, determine which persona is most relevant, and propose a question that tests the hypothesis without assuming it is true.

Execute at scale

Use the execution layer for repeatable work: source collection, scheduled monitoring, record updates, preliminary scoring, agent specialization, and output generation. Specify required fields, source hierarchy, confidence labels, stop conditions, and prohibited inferences.

Challenge the result

Review both the output and the process that produced it. Does the evidence support the recommendation? Did several agents rely on the same underlying source? Does the scoring model reward useful accuracy or visible activity? Can the reviewer reject the premise? Is there a mechanism for recording what happened next?

Apply human judgment

Human review is essential when the recommendation could affect a relationship, publication, customer record, or financial decision. Some actions should remain drafts. Some signals should return for more research. Some recommendations should be rejected. A human checkpoint is meaningful only if it has the authority to stop the process.

Feed learning back into the system

Convert credible observations into durable changes. Revise a prompt, add a required field, strengthen a source rule, create a review gate, or deliberately leave the workflow unchanged when the evidence is insufficient. Then return the revised operating instructions to the execution layer.

Controls that keep multiple systems from reinforcing the same mistake

The most dangerous failure in a multi-system workflow is not always an obvious error. It is a plausible mistake repeated across several agents until repetition is mistaken for corroboration.

I rely on seven controls:

Separate evidence from interpretation. A record should make clear what a source states, what the system infers, and what remains unknown. Confidence should reflect evidence quality, not the fluency of the prose.

Preserve provenance. Material claims need a source, date, and path back to the underlying record. Traceability makes review and correction possible.

Check for source independence. Agreement among agents adds little when they all rely on the same article, database entry, or earlier summary. Apparent consensus is not independent verification.

Permit premise-level challenge. A reviewer should be able to conclude that the question is poorly framed, the evidence is insufficient, or the proposed action is not warranted.

Design real stop conditions. "Pass," "monitor," "research required," and "not ready" are legitimate outputs. Requiring action from every workflow degrades judgment.

Keep consequential actions human-controlled. Sending outreach, publishing analysis, changing customer records, and making financial decisions from probabilistic recommendations deserve stronger approval than drafting or internal research.

Change rules with evidence. Record exceptions and outcomes, but do not rewrite thresholds after every success or failure. Adaptation should follow a credible pattern rather than a memorable anecdote.

What leaders should measure

Output volume is easy to count and easy to misinterpret. Signals found, drafts created, and reports delivered are measures of production. They do not establish that the system improved a decision.

The right measures depend on the workflow, but a practical scorecard can include:

Dimension	Example measures
Evidence quality	Verified-signal rate, source freshness, unsupported-claim rate, correction rate
Decision quality	False positives, recommendations rejected on review, actions changed after challenge
Business value	Relevant replies, qualified meetings, opportunities influenced, time to useful insight
Operational efficiency	Time saved, account coverage, cycle time, research cost per approved action
Governance	Approval compliance, provenance completeness, suppression compliance, exception frequency
Learning	Outcomes captured, hypotheses resolved, rule changes supported by multiple observations

Not every organization needs every measure. The essential point is to connect production to review, action, and outcome. Without that chain, a large volume of AI-generated work can look productive while teaching the organization very little.

A practical starting point for business leaders

Organizations often begin agentic-AI adoption with product selection. I would begin with work design.

Choose one recurring workflow with enough volume to justify automation and enough business value to warrant review. Define the decision, the minimum evidence required, and the consequence of an error. Assign an execution layer and an independent challenge layer. Name the human owner. Record what the system recommended, what the human decided, and what happened next.

Keep the first pilot narrow. For example, monitor 25 accounts and produce no more than five verified signals each week. Require the system to explain why each signal matters, what remains unknown, and which action, if any, is justified. Use the pilot to find weak handoffs before expanding the account universe or increasing automation.

Finally, document the workflow as an operating model, not merely as a prompt. Prompts are components. The system also includes sources, schedules, state, roles, approvals, exception handling, feedback, and governance.

Conclusion

Using Pokee AI and ChatGPT together has changed how I think about AI adoption. The advantage does not come from asking two systems the same question and choosing the answer I prefer. It comes from assigning distinct responsibilities and making disagreement useful.

In my workflows, Pokee AI provides persistence and execution at scale. ChatGPT provides critique, synthesis, editorial judgment, and redesign. I provide the objective, context, risk tolerance, and accountability that neither system can own.

This is not full automation, and it is more than a conventional assistant relationship. It is a managed collaboration among execution, independent challenge, and human judgment. When those roles are clear and the feedback loop is real, AI can do more than accelerate existing work. It can help create a more disciplined way of doing it.

Researchers have described a related concept as the human-algorithm centaur, in which human judgment and computational capability are combined within one decision process.[3] My experience suggests that the practical value of that model depends on something less glamorous than the technology itself: clear roles, evidence standards, and the authority to stop when the evidence is not good enough.

Sources

1. Pokee AI. “About.” Accessed September 21, 2026. https://pokee.ai/about
2. OpenAI. “Introducing ChatGPT Agent: Bridging Research and Action.” July 17, 2025. https://openai.com/index/introducing-chatgpt-agent/
3. Saghafian, Soroush, and Lihi Idan. “Effective Generative AI: The Human-Algorithm Centaur.” 2024. https://arxiv.org/abs/2406.10942