

You Are Paying Twice

Once for the system that produces the work. Again for the expert who has to read all of it. Only one of those was in the business case.

The AI was supposed to remove the work.

It moved it.

Production is faster than it has ever been. Your agents generate plans, drafts, recommendations and actions at a rate no team could match. That part worked exactly as promised.

And every one of those outputs lands in front of a person who has to decide whether it is right — because in your domain, a mostly correct answer is not a partial win. The residual uncertainty is precisely where the liability sits.

So you are running two payrolls against one workflow. The bottleneck did not disappear. It relocated, from production to supervision, and the second line item never appeared in the case that justified the first.

The AI system that produces the work

BUDGETED · APPROVED

The expert who must read every output

UNBUDGETED · PERMANENT

This is not autonomy. It is supervised automation, and it costs more than what it replaced.

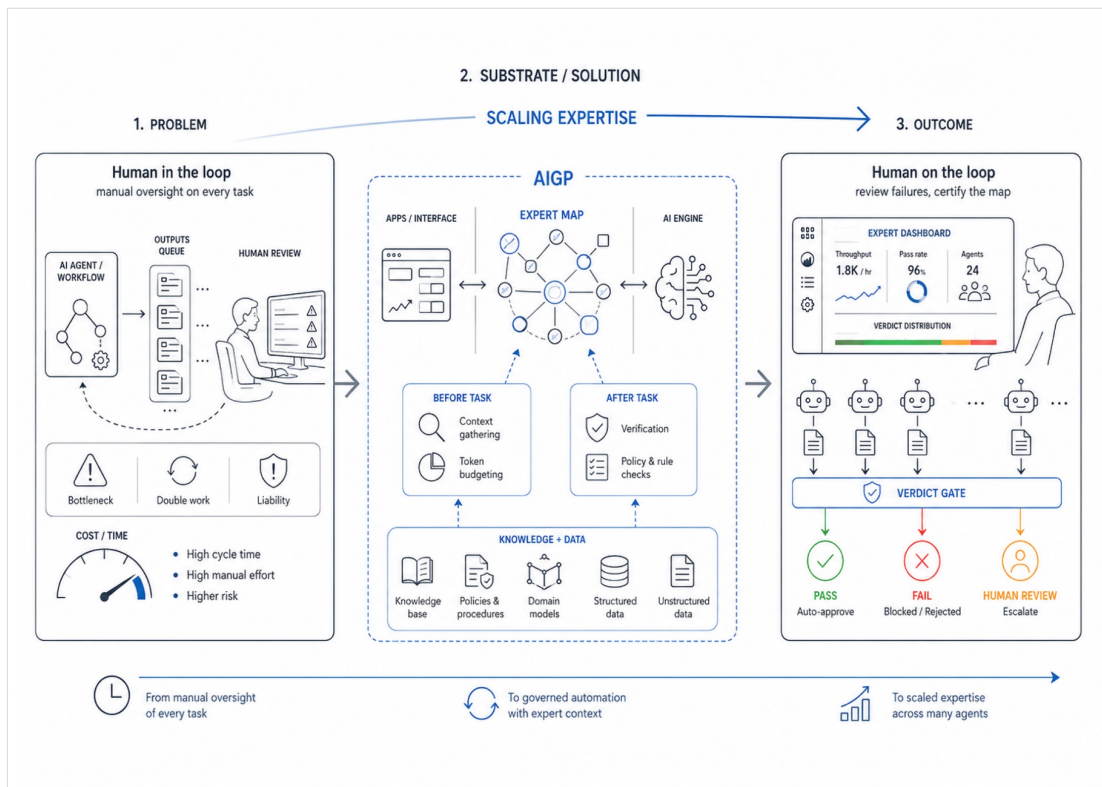
These are not two design positions

Human-in-the-loop and human-on-the-loop are usually presented as a choice: pick the posture that matches your risk appetite. They are not a choice. They are stages along a governed trajectory.

The question was never where the human sits in relation to the AI system. The question is **what must be true** before that human can move from inspecting individual outputs to governing the system that produces, evaluates and improves them.

That movement cannot happen by trust alone. It cannot happen by removing reviews, accepting more risk, or waiting for better models to make oversight unnecessary. Each of those is a way of arriving at the destination without earning it, and each fails in the same place — the first time something goes wrong and nobody can say why it was permitted.

It requires a substrate capable of scaling expertise. Nothing else moves you along the line.



SCALING EXPERTISE: FROM MANUAL INSPECTION, THROUGH GOVERNED AUTOMATION WITH EXPERT CONTEXT, TO SCALED EXPERTISE ACROSS MANY AGENTS.

02 · THE PROBLEM

Oversight does not scale

Agentic AI generates work at speed, but generation is not accountable execution. A model can produce an answer, a plan, a recommendation or an action. It does not, on its own, provide the three things an expert provides: **correctness, efficiency and accountability**.

When those qualities still require a human to review every result, AI does not scale. It shifts the bottleneck from production to supervision — and supervision is the more expensive half, because it consumes the scarcest thing in the organisation.

The problem is not that humans are in the loop. **The problem is that the loop has no mechanism to absorb, structure and reuse what the human knows.** Every task returns to the expert as another output to inspect. The same judgment is applied to the same class of problem, hundreds of times, and evaporates on each occasion.

The expert remains the bottleneck because expertise remains tied to individual acts of review. Nothing accumulates.

You are not short of expertise. You are throwing it away one output at a time.

03 · THE MISSING SUBSTRATE

Four things that will not carry expertise

Most AI systems are built around models, prompts, tools and logs. All necessary. None sufficient — and it is worth being precise about why each one fails, because each is currently doing duty as the answer somewhere.

Prompts	Too local. They capture task instructions but not the durable structure of a domain — and they are only as good as whoever last edited them.
Logs	Too passive. They record what happened. They do not, by themselves, determine what should be learned from it.
Guardrails	Too narrow. They block known violations. They do not define what correct work looks like across a domain, which is the part your expert actually holds.

Human review

Too expensive as a permanent operating model. It preserves accountability only by keeping expertise locked in manual inspection — which is the arrangement you are paying for now.

What is missing is a governed substrate that holds expert knowledge in a form agents can use, systems can evaluate against, and humans can certify over time.

Expertise must become executable.

04 · THE SOLUTION

Instrument, map, compose

AIGP turns expertise into an executable governance layer. Three movements, in order — and the order matters, because each one is only possible once the previous is in place.

01

Instrument

AIGP sits between applications and AI systems as a governed execution layer, introduced by a code change or an endpoint swap. From that point, AI calls are checked before execution and recorded afterward. You gain a traceable record of what the AI attempted, what was allowed, what was blocked, and why.

This does not yet make the system human-on-the-loop. It creates the conditions for movement. Without instrumentation there is no reliable memory of behaviour — and without behavioural memory there is no way to learn from use, identify recurring failures, or determine where expert attention is genuinely needed rather than merely habitual.

02

Map

Expertise should not live in prompts. It needs to become a structured map of the domain: what matters, how things relate, what correctness looks like, what evidence is required, and when uncertainty must be escalated.

The map changes the agent's behaviour on both sides of the task. **Before**, the agent walks the map to gather the right context, constrain the work and set an appropriate budget for effort — rather than overloading the prompt with everything available. **After**, the output is evaluated against the map. Not a vague score, not a model-generated self-explanation, but a governance verdict.

PASS	FAIL	HUMAN REVIEW
Auto-approve	Blocked or rejected	Escalate

03

Compose

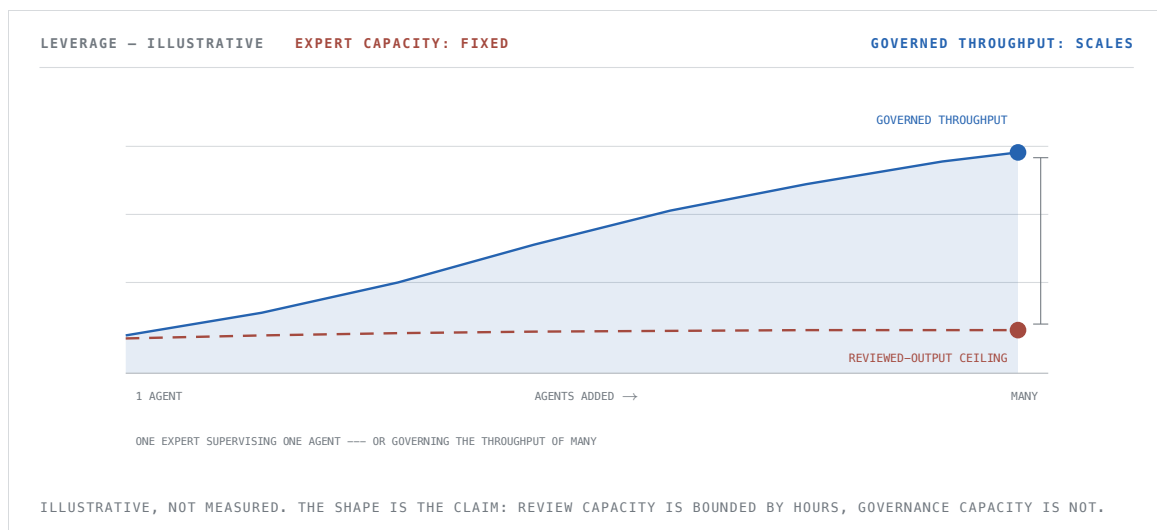
Once expertise is captured as a map, it becomes reusable. The same structure generates training data, defines benchmarks, guides model selection and turns failures into improvements. Every failure the system catches becomes a signal: refine the map, correct the knowledge base, expand the benchmark, change the prompt, replace the model.

The system that governs the work also sharpens the expertise required to perform it. That is the compounding step, and it is the one that no amount of review effort produces on its own.

Inspection becomes stewardship

The operational change is simple to state and hard to overstate. The expert stops reviewing every output and starts governing the substrate that produces and evaluates them.

Reviews every result	→ Reviews failures
Rewrites prompts	→ Certifies the map
Corrects the same error repeatedly	→ Improves the knowledge substrate
Sits inside the execution loop	→ Governs the learning loop



This is the practical meaning of scaling expertise. One expert no longer supervises one agent. **One expert governs the throughput of many.**

Not removal. Relocation.

The promise of agentic AI is usually framed as replacing human work. That framing is too crude, and in serious domains it is also wrong. The goal is not to remove the human from responsibility. It is to move human expertise to where it has the greatest leverage.

Human-in-the-loop keeps the expert close to the output. Human-on-the-loop brings the expert closer to the governing structure. The first protects quality through inspection; the second protects it through design, certification, feedback and continuous improvement.

Organisations do not get there by skipping oversight. **They get there by making expertise reusable** — and until they do, every additional agent they deploy makes the supervision problem worse rather than better.

Human-on-the-loop is not the removal of the human. It is the scaling of human expertise.

07 · RESEARCH GROUNDING

This is not a novel claim

The argument that these are points along a trajectory rather than alternatives is consistent with a substantial body of existing work.

LEVELS OF AUTONOMY

Autonomy is not binary. It varies with context, system capability, risk, reliability and the distribution of control between humans and machines — the premise the whole trajectory rests on.

HUMAN FACTORS

Automation does not eliminate the human problem. It shifts human work from execution to supervision, creating new risks of overreliance, monitoring failure, complacency and decision bias.

GENERATIVE AND AGENTIC AI

These intensify the concern, because humans are asked to supervise outputs produced faster than they can be meaningfully inspected. Which is the condition described on page one.

COGNITIVE FORCING FUNCTIONS

Overreliance falls when systems force analytical engagement with a recommendation. Human review cannot be assumed — it has to be architected.

HUMAN-AI COLLABORATION

Successful systems are not produced by stronger models alone. They require human-centred design, clear task allocation, appropriate feedback and deliberate structuring of the relationship.

STRUCTURED DOMAIN KNOWLEDGE

Knowledge graphs and executable knowledge graphs support the claim that expertise can be externalised into inspectable, reusable, operational forms — out of fragile prompts and into structures that support retrieval, reasoning, benchmarking and evaluation.

Together these support the central position: human-on-the-loop is not achieved by removing humans from AI work. It is achieved by scaling the expertise that makes AI work accountable.

**CLOSING POSITION**

Human-on-the-loop is not a destination reached by trusting AI more. It is a capability reached by making expertise executable, reusable and governable.

The trajectory begins with instrumentation, progresses through domain mapping, and matures into composition. At each stage the human remains essential; what changes is the form of participation. The expert is no longer consumed by every transaction. The expert governs the conditions under which transactions occur.

Until that substrate exists, every agent you add is another output queue, and another claim on the one resource you cannot hire your way out of.

The shift is from human labour as a bottleneck

to human expertise as a scalable control surface.

SELECTED REFERENCES

- Abraham, S., Carmichael, Z., Banerjee, S., VidalMata, R., Agrawal, A., Al Islam, M. N., Scheirer, W., & Cleland-Huang, J. (2021). Adaptive autonomy in human-on-the-loop vision-based robotics systems. arXiv:2103.15053. <https://arxiv.org/abs/2103.15053>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, 3(2), 74–99. <https://doi.org/10.5898/JHRI.3.2.Beer>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. <https://doi.org/10.1145/3449287>
- Luo, Y., Yu, Z., Wang, X., Zhu, Y., Zhang, N., Wei, L., Du, L., Zheng, D., & Chen, H. (2025). Executable knowledge graphs for replicating AI research. arXiv:2510.17795. <https://arxiv.org/abs/2510.17795>
- O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, 64(5), 904–938. <https://doi.org/10.1177/0018720820960865>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A. E., Sellen, A., & Rintel, S. (2024). Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interactions. arXiv:2402.11364. <https://arxiv.org/abs/2402.11364>
- Stewart, I. A., et al. (2026). Knowledge graph-guided agentic AI for cross-domain applications. arXiv:2602.07491. <https://arxiv.org/abs/2602.07491>
- Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., Titterton, V., Malreddy, S. V., Aggarwal, R., Xu, Y., Ding, L., Mehta, J., Grinnell, N., Liu, L., Zhong, S., Gandamani, D. N., Tang, X., Ghosalkar, R., Shen, C., Shen, R., Hussain, N., Ravichandran, K., & Davis, J. (2024). A survey on human-AI teaming with large pre-trained models. arXiv:2403.04931. <https://arxiv.org/abs/2403.04931>

SOURCE NOTES

Bainbridge's classic work supports the claim that automation can exacerbate rather than eliminate problems involving the human operator, especially when the human remains responsible for abnormal conditions. Parasuraman and Riley raise the concern that automation use can become misuse, including overreliance and monitoring failure. Beer, Fisk and Rogers argue that autonomy should be understood in levels rather than as a binary. Buçinca, Malaya and Gajos show that cognitive forcing functions can reduce overreliance on AI recommendations, although users may dislike the friction.

Recent work on generative AI and human-AI teaming extends these human-factors concerns into modern AI systems. Recent knowledge-graph work supports the claim that structured knowledge can provide a reusable substrate for agents, retrieval, reasoning and evaluation.

The leverage chart is illustrative rather than empirical. It renders the paper's claim about the shape of the relationship — bounded review capacity against unbounded governance capacity — not a measured dataset.

SCALING EXPERTISE · FROM HUMAN-IN-THE-LOOP TO HUMAN-ON-THE-LOOP AS A GOVERNED TRAJECTORY

POSITION PAPER / CONCEPT NOTE · CAUSUM RESEARCH · JULY 2026 · ORIGINALLY DEVELOPED AT KANJANI AI RESEARCH