d—school

The Agency Continuum

A Strategic Framework for Orchestrating Human-AI Agency

Whitepaper by Samuel Tschepe
HPI d–school / February 2026

www.hpi-dschool.de

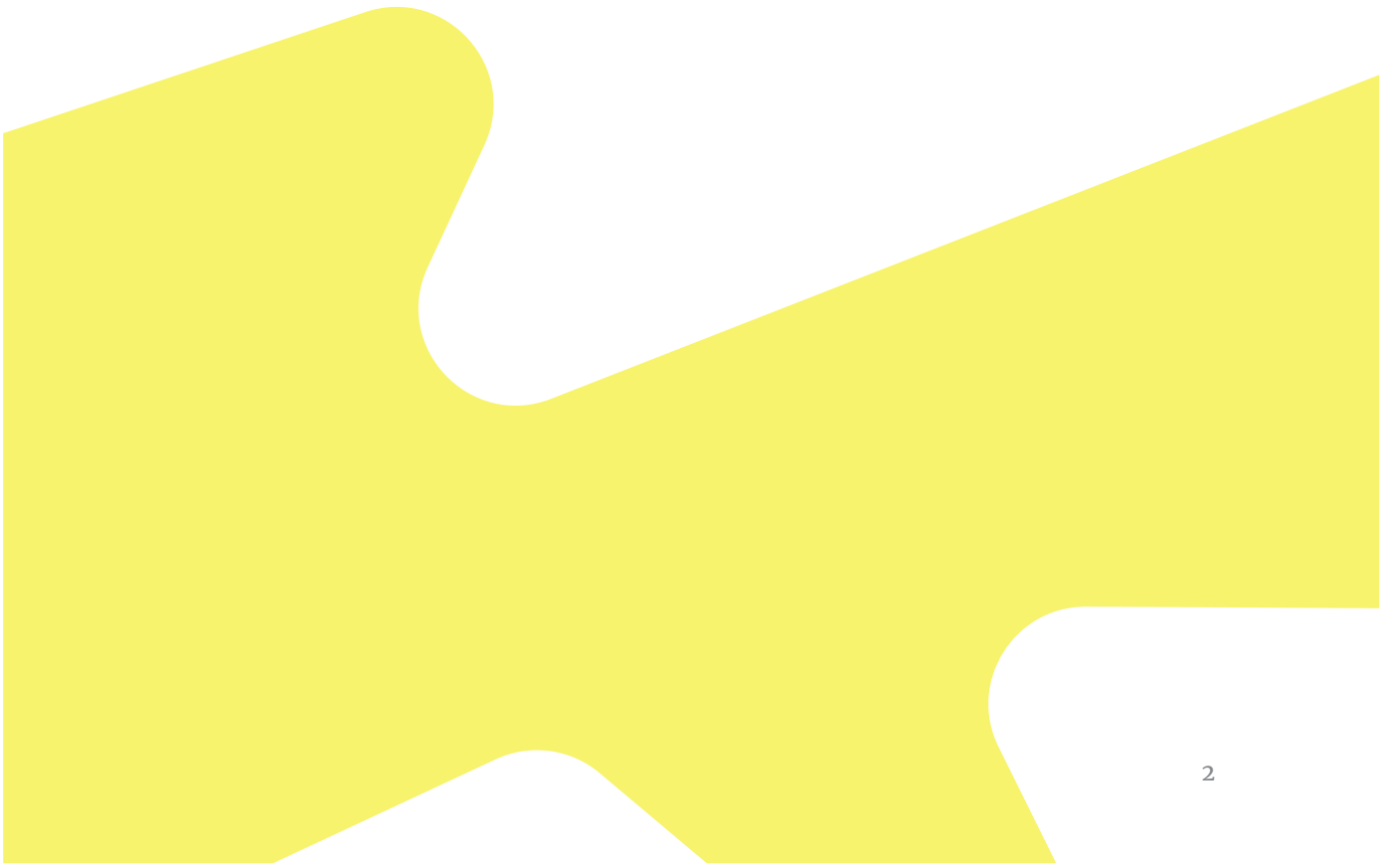
Executive Summary

By early 2026, AI has become ubiquitous. However, many organizations feel less empowered than expected. Work intensifies, output homogenizes, accountability blurs, collaboration remains shallow, and core skills erode. The underlying failure is strategic, not technical: AI is still treated as a transactional tool rather than a system of distributed agency.

This paper introduces the Agency Continuum: a strategic framework for consciously orchestrating decision-making power between humans and AI. It defines five distinct modes – each with explicit strategic gains and trade-offs – and a Mode Decision Logic that enables deliberate choice and continuous correction. Together, they provide a shared language teams can use to prevent defaulting, drift, and governance blind spots.

The central shift is from "What can AI do?" to "How should decision-making power be distributed between humans and AI?" In a commoditized world, everyone has access to the same tools. The advantage lies in the design of the distribution – not the capability of the technology.

The Agency Continuum offers a practical compass by making these choices explicit, discussable, and adjustable in everyday work. It reframes AI adoption as a question of leadership capacity in hybrid systems, rather than of technology deployment alone.



Content

I. The Context: Five Crises of Orientation

For decades, we've been trained on deterministic software: input leads to predictable output. You click "save," the file saves. You enter a formula, the spreadsheet calculates. This is how software has worked for an entire generation of knowledge workers.

Generative AI breaks this pattern in a way that is easy to underestimate. From the user's perspective, its behavior is effectively non-deterministic: the same prompt can yield different outputs, and confident-sounding responses may be hallucinations. Even when a model is technically deterministic under fixed parameters, its probabilistic training and sampling mean that the mapping from input to output is opaque and variable in everyday use. The familiar intuition – that identical inputs will reliably produce identical results – no longer holds.

This fundamental difference explains why familiar approaches fail. The crises that follow are symptoms of that mismatch.[1]

Crisis 1: The Fatigue Crisis

The promise was productivity. AI tools would handle the tedious work, freeing us for strategy, creativity, the work that matters.

The reality is the opposite. In a 2024 global study of 2,500 workers by the Upwork Research Institute, 77% of those using AI reported that it *increased* their workload rather than reducing it. Nearly half (47%) felt unable to achieve the productivity gains their organizations expected, while 40% said their company demanded too much of them in relation to AI adoption – contributing directly to overload and burnout.

The trajectory is worsening. Fortune's analysis of S&P Global Market Intelligence data shows that organizations scrapping the majority of their AI initiatives surged from 17% in 2024 to 42% in 2025. Quantum Workplace's 2025 Employee Experience report reveals that frequent AI users report significantly higher burnout rates (45%) compared to infrequent users (38%) or non-users (35%). The more you use AI, the more exhausted you become.

It's not just frontline workers. An EY survey of 500 senior executives found that roughly half observed declining organizational enthusiasm for AI, with leaders themselves reporting fatigue from the "persistent onslaught of information and developments tied to AI." Wiley Workplace Intelligence identifies this as a "cascade crisis": continuous AI-related changes triggering stress, exhaustion, and declining performance when organizations proceed without clear policies, adequate training, or strategic frameworks.

While methodologies differ across these studies, the pattern is consistent: across multiple large-scale surveys, intensive AI use correlates with higher perceived workload and burnout. This does not mean AI cannot improve productivity – controlled studies show substantial gains in specific contexts (Brynjolfsson et al., 2023; Noy & Zhang, 2023). But these gains require deliberate deployment. Without strategic frameworks, the default outcome is exhaustion, not efficiency.

[1] This paper offers a synthesized view, aggregating patterns from multiple workforce studies, technical AI research, governance reports, and organizational surveys. The goal is practical clarity, grounded in the best available data.

Why is this happening? AI has transformed knowledge workers from makers into managers. Consider how work used to flow. Writing an important report meant clearing your schedule, entering a flow state, and producing. Hard work, but focused – one continuous cognitive arc. You were the maker.

Now? You prompt. The AI generates a draft in ten seconds. Then your real job starts. Your job becomes forensic analysis: Is this fact true? Did it just hallucinate that quote? Why is the tone so weirdly enthusiastic? You spend the next 45 minutes checking, correcting, re-prompting. It's constant context-switching. Not creating, but supervising. And supervising an entity prone to confident hallucinations is incredibly taxing. That constant vigilance, that low-level anxiety that paragraph seven contains a subtle error you'll be blamed for, burns more cognitive energy than writing the thing yourself.

Recent multi-turn hallucination benchmarks confirm that this is not a marginal issue: even state-of-the-art models still produce unsupported factual claims in roughly one third of complex, citation-based conversations even when equipped with web search, and around 60% of the time without it (Fan et al., 2026).

It's like having an intern who graduated top of their class at MIT – and is also drunk. Brilliant, but unreliable. You can't take your eyes off them for a second.

This is the Supervisor's Tax, and we're paying it all day. Teams adopt tools reactively, chasing the next capability without changing the fundamental workflow – piling more drunk interns onto their cognitive load.

This isn't an inherent property of AI itself. The fatigue crisis emerges from organizational design failures: reactive adoption without strategic integration, inadequate training, and the absence of clear frameworks for when and how to use AI tools.

Crisis 2: The Generic Drift

When everyone uses the same models trained on increasingly similar data, everyone produces increasingly similar output.

One 2025 analysis estimated that around 74% of newly created websites may contain AI-generated text (WINSS Solutions); a striking figure, though AI detection methods remain imprecise. The technical mechanism is called "model collapse": when models train recursively on AI-heavy data, they enter a degenerative feedback loop. Rare but essential patterns disappear, diversity metrics decline, and outputs converge toward homogenized mediocrity. The WINSS analysis suggests the data foundation for future models may already be saturated with AI outputs – meaning future models will be trained on the bland outputs of current models, accelerating the drift toward average.

The Stanford AI Index 2025 documents shrinking performance gaps among frontier models, and separate research using embedding similarity metrics confirms that different LLMs often produce highly overlapping responses to identical prompts. This isn't coincidence; it's the predictable result of training on overlapping datasets and optimizing toward statistical probability. Note the paradox: individual outputs remain probabilistic and unpredictable, yet aggregate outputs converge. Organizations face both challenges simultaneously – they

cannot predict what any single response will say, but they can predict that responses across models will lack diversity.

The symptoms are everywhere: marketing campaigns that blur together, product descriptions swappable between competitors without detection, strategy documents that read like variations on a template. Three companies in the same sector launch nearly identical messaging in the same week because they all prompted the same foundation models the same way. Beyond business contexts, social media platforms overflow with what critics call "AI slop": low-quality generated content, synthetic images, and deepfakes that blur the line between authentic and artificial.

Call it the "beige-ification of human output" – a digital landfill of statistically probable mediocrity.

The economic cost is brutal: when differentiation erodes, competitive advantage collapses. What once required hours and had scarcity value now takes seconds and is abundant. And worthless. Organizations find themselves in a race to the middle, producing faster mediocrity while struggling to create anything genuinely distinctive.

Differentiation remains possible through proprietary data, strong editorial judgment, or domain expertise that shapes AI outputs rather than accepting defaults. But it requires deliberate strategy, not passive adoption.

The irony cuts deep: AI was supposed to amplify creativity. Without thoughtful intervention, it homogenizes it.

Crisis 3: The Governance Crisis

Who's accountable when the AI acts autonomously? In most organizations, the question has no clear answer.

The 2025 Agentic Systems Governance Report provides stark numbers: 54% of organizations now deploy at least one autonomous AI agent in production, but 72% have no formal oversight model. The gaps are systematic – 81% lack documented governance for machine-to-machine interactions, 76% have no audit trail for agentic decisions, and 62% report at least one agent-driven incident in the past 12 months.

IBM's "AI at the Core 2025" survey reinforces the picture: nearly three-quarters of organizations admit their AI risk and governance frameworks provide only "moderate or limited" coverage. The International Telecommunication Union's Annual AI Governance Report is blunt: governance mechanisms lag several years behind technological deployment. The 2026 International AI Safety Report reinforces this assessment: "The capabilities of the best AI systems improve significantly month-to-month, while major legislation typically takes years to draft, negotiate, and implement. This mismatch means that the AI landscape can change while policy processes unfold, making it difficult to design policies that address emerging risks and are robust to future changes" (International AI Safety Report, 2026).

So, the scale of deployment is accelerating faster than governance can follow. In January 2026, McKinsey & Company CEO Bob Sternfels revealed the firm now operates roughly 25,000 AI agents working alongside approximately 40,000 human employees, aiming for a

1:1 pairing. How does an organization govern decision-making when its workforce composition shifts this dramatically, this quickly?

Recent enterprise data suggests how quickly this challenge is scaling. Nearly three-quarters of companies expect to use agentic AI at least moderately within the next two years, yet only about one in five report having a mature governance model for autonomous agents (Deloitte, 2026). This gap between planned deployment and oversight capacity turns agentic systems into a structural governance risk rather than a merely technical one.

At the same time, emerging work on agentic AI workflows and agent infrastructure shows how networks of AI agents can coordinate tasks, call tools, and act across organizational boundaries with limited direct human oversight. This shifts the governance challenge from individual systems to the emergent behavior of interacting agents embedded in technical and organizational infrastructures – and makes the question of who designs and monitors these decision architectures even more acute.

The incidents aren't science fiction. They're real. Take the following two examples: In 2024, Air Canada was ordered by the British Columbia Civil Resolution Tribunal to honor a bereavement discount that its website chatbot had invented – after the airline argued, unsuccessfully, that the bot was a "separate legal entity" responsible for its own actions. The tribunal ruled that the chatbot was part of Air Canada's website and that the company was accountable for its misinformation, turning a flawed prompt flow into a legal and reputational liability.

By early 2026, the open-source agent project now known as OpenClaw (previously Clawdbot and Moltbot) had pushed these risks to a new scale. The agent runs locally on users' machines, connects to everyday apps such as WhatsApp, Signal, or Slack, and can manage emails, calendars, shell commands, web scraping, and multi-step workflows on their behalf. Within weeks, a rapidly growing ecosystem of community-built skills turned OpenClaw from a single assistant into de facto infrastructure: agents could not only act inside one user's environment, but also share playbooks and automation patterns across a wider network. Commentators describe it as a "swarm substrate": the technical plumbing and social layer together allow risky defaults, insecure skills, or overly broad permissions to spread quickly – long before most organizations have formal processes for vetting agent capabilities, default access rights, or escalation paths.[2]

When a human makes a catastrophic error, accountability is usually clear. When an AI agent does, responsibility becomes diffuse. Is it the designer who set the goals, the manager who deployed the system, or the vendor who built the model? In many organizations, no one can answer that question – until regulators, courts, or angry customers force the issue.

It's like engaging autopilot without a flight plan. Organizations are learning through expensive failures that autonomy without accountability is untenable. Many are learning too slowly.

[2] Learn more about these examples here (Air Canada) and here (OpenClaw). Access Feb 2nd, 2026.

Crisis 4: The Collaboration Collapse

According to research by Jeremy Utley at Stanford, based on 5,000 anonymous survey responses, 70% of workers don't know how to apply AI to their work – and this percentage is increasing despite training investments. Only 8% use AI as a sparring partner for iterative dialogue. More than half stop after just one or two prompts, never experiencing what emerges through sustained interaction.

A 2025 systematic review in Frontiers in Computer Science, analyzing 105 empirical studies, reached a similar conclusion: "Human-AI collaboration is not very collaborative yet." Most interactions follow one-directional patterns. AI makes recommendations; humans simply accept or reject them, with little support for genuine, bidirectional exchange.

Studies on human-AI communication show that perceived AI agency and anthropomorphism strongly shape trust and how people evaluate chat quality. Systems that feel like quasi-human partners often invite unearned deference, while tools that appear low-agency are treated as vending machines. Both patterns reinforce the imagination ceiling: they encourage people either to overtrust AI as an authority or to underuse it as a sparring partner, instead of engaging in sustained, critical dialogue.

Anthropic's Economic Index (2026) adds an important distinction. Roughly 52% of AI interactions are technically "collaborative" in the sense that they involve iteration rather than single-shot prompts. But iteration alone is not partnership. True collaboration requires using AI as a challenger and co-thinker. Only 8% do that (per Utley). The majority iterate on shallow tasks, accepting outputs without intellectual friction.

The barrier isn't skepticism; only 4% doubt AI's value. It is what Utley calls the "imagination ceiling." People struggle to envision AI as a teammate because they have only experienced it as a vending machine: insert a prompt, receive an output, walk away. "What is the capital of France?" "Summarize this PDF." Input and output.

This pattern is not unique to AI. Many organizational cultures train people to work transactionally – receive task, deliver output, move on. The imagination ceiling for AI collaboration may reflect a broader imagination ceiling for collaboration itself.

But the real value emerges in a fundamentally different interaction. Instead of using AI to retrieve answers, it can be used to challenge assumptions, stress-test strategy, or surface adversarial perspectives. Such prompts are uncomfortable by design. They invite criticism and create cognitive friction. That discomfort is precisely why so few people use them – and why those who do extract disproportionate value.

It is like having access to a world-class coach and using them only to keep score.

The organizational symptom is a lack of shared vocabulary. One team avoids AI entirely ("we don't trust it"), while another accepts every output without critique ("the AI knows best"). Neither extreme works. Both reflect the absence of frameworks for genuine partnership. The hidden cost is unrealized potential at massive scale: organizations sitting on strategic capability they do not know how to access.

Crisis 5: The De-Skilling Crisis

AI doesn't replace human thinking. It amplifies it. And that's precisely the problem.

The Anthropic Economic Index reveals a very strong correlation ($r \approx 0.9$) between the sophistication of a user's prompt and the sophistication of AI's response. Research on AI as a cognitive amplifier confirms the mechanism: "The same AI feature – deference to user feedback – amplifies expertise for those who possess it and amplifies misconceptions for those who lack it." Experts spot when AI generates plausible but incorrect information. Novices lack the knowledge base to notice. This connects directly to the oversight crisis discussed earlier: novices face the worst of both worlds. They must constantly verify AI outputs – which is exhausting – but lack the expertise to verify effectively. Maximum effort, minimum protection.

Let's look at an example, consider two analysts with the same AI tools. Analyst A has strong statistical fundamentals. She spots a confounding variable the AI missed, re-prompts with controls, and through several iterations produces robust analysis while sharpening her judgment. Each interaction makes her better. Analyst B learned analytics primarily through AI assistance. He accepts outputs without deep scrutiny, lacking the foundation to spot issues. Six months later, he's more dependent on AI and less capable of evaluating its outputs. Each interaction makes him worse. The difference isn't the AI. It's the foundation each brought to it.

This is the competence-output loop. The quality of your inputs determines the quality of AI outputs, which shapes your future capability to generate quality inputs. The loop operates in both directions:

Virtuous cycle: Strong domain expertise → sophisticated prompts → high-quality outputs → refined judgment → even stronger expertise.

Vicious cycle: Weak domain foundations → shallow prompts → mediocre outputs → uncritical acceptance → skill atrophy → further decline.

So, there is a fundamental problem here: if you haven't mastered the craft yourself, you lose the ability to judge whether the AI's output is actually good or just 'good enough'. The very experience of struggling to produce something – the frustration, the iteration, the hard-won mastery – is what builds the judgment to evaluate it.

The crisis compounds at the organizational level. As AI takes over tasks that junior employees once learned from, companies reduce entry-level hiring and remove the first rung of the ladder. At the same time, existing staff de-skill through overreliance. The pipeline for developing expertise erodes from both ends. AI was supposed to augment human capability. Left unmanaged, it hollows out the foundation on which capability is built.

Large-scale enterprise surveys point in the same direction. Many organizations expect double-digit shares of jobs to be fully automated within just a few years, yet the vast majority have not fundamentally redesigned roles or career paths around AI capabilities. Instead, most focus on raising AI fluency through training while leaving underlying job architectures largely untouched – precisely the combination that accelerates the competence-output loop in the wrong direction: more automation, less deliberate skill formation.

The Common Root[3]

Five symptoms, one underlying condition: we're asking the wrong question.

"What can AI do for me?" frames AI as a tool to extract value from – a vending machine that dispenses outputs. That framing produces the crises we've examined.

A different question reframes the entire relationship: *"How should decision-making power be distributed between humans and AI?"*

That question shifts everything. It moves from optimizing a tool to designing a system, from implementation to deliberate orchestration. This reframing is urgent because this transition is fast. Previous technological disruptions unfolded over decades; this one unfolds in years. It affects not physical but cognitive labor. And the tools improve recursively – creating dynamics without historical precedent.

Change is inevitable, and often beneficial. What matters is whether it is shaped deliberately or allowed to happen by default. The organizations that thrive will be those that consciously design how decision-making power flows between humans and machines.

The next section explores what that shift requires. The framework that follows makes it operational.

[3] A note on evidence and limitations seems to be important here once more: The five crises synthesized in this section draw on heterogeneous empirical sources: global workforce surveys, governance reports, controlled experiments, and case studies from predominantly knowledge-work contexts in North America and Europe. The pattern that emerges is consistent – intensive AI use is correlated with higher perceived workload and burnout, generic drift, governance gaps, superficial collaboration patterns, and skill erosion – but it remains correlational rather than strictly causal. Sectoral and regional differences, as well as rapid changes in AI tooling and practices, may attenuate or amplify specific effects. The aim of this synthesis is therefore not to claim universal, quantified prevalence, but to provide a practically useful pattern language for phenomena that recur across multiple large-scale studies and organizational settings.

II. The Shift: From Implementation to Orchestration

If the transactional model is broken, what replaces it? The answer requires rethinking a concept we usually take for granted: agency itself.

Redefining Agency

Agency is the capacity to make decisions that shape outcomes. This definition is deliberately functional rather than ontological; it focuses on observable decision-making authority, not on questions of consciousness or intentionality. Whether AI systems possess "genuine" agency in a philosophical sense – or whether humans do, given that cognition is itself a physical process – is a question this framework does not need to resolve.

What matters practically is that AI outputs are probabilistic, context-adaptive, and often opaque. This unpredictability makes conscious agency distribution strategically consequential.

In human-AI collaboration, agency is not binary – human OR machine. That's 2023 thinking. It's obsolete. The new reality: Agency is distributable. It exists on a continuum. At one end, 100% of decision-making power rests with humans. At the other, AI acts autonomously within human-defined boundaries. Between these poles lie configurations where power is shared, shifted, or strategically allocated.

This insight – that agency can be consciously distributed – is the foundation of everything that follows.

The Discovery of Meta-Sovereignty

If agency can be distributed, a deeper question emerges: who decides how it is distributed?

Consider a familiar organizational pattern. A competent leader does not personally approve every operational detail. They design the conditions under which decisions are made, delegate authority where appropriate, and retain responsibility for the overall system. Control is not exercised through constant intervention, but through deliberate design.

This distinction reveals two fundamentally different forms of sovereignty:

- **Operative Sovereignty:** the authority to make a specific decision.
- **Meta-Sovereignty:** the authority to decide who or what is empowered to make that decision[4].

When a human delegates a task to an AI system, they do not inherently lose control. They exercise meta-sovereignty – provided the delegation is conscious, bounded, and reversible.

[4] This notion of meta-sovereignty is distinct from current debates on digital or AI sovereignty, which focus on who controls data, infrastructure, and models at the level of states or platforms. Here, sovereignty is explicitly about organizational and individual authority over how decision-making power is distributed between humans and AI in concrete workflows. It asks a different question: not "Who owns the infrastructure?" but "Who designs and owns the decision architecture of work?"

These conditions are easier to meet for processes than for capabilities. A workflow can be redesigned; a skill that atrophies through disuse cannot be instantly restored. This asymmetry is why the de-skilling crisis is particularly dangerous: delegation of cognitive work may be practically irreversible even when formally reversible. The loss of control occurs not through delegation itself, but through defaulting: allowing tools, incentives, or urgency to determine agency distribution implicitly.

Most current failures in human–AI collaboration are failures of meta-sovereignty. Agency is shifted without being designed. Decisions about speed, scale, and autonomy are made by convenience rather than intent. Work on ethical responsibility in hybrid systems – such as the ERDEM and HAIG frameworks (2024) – addresses this gap at the level of principles and governance; meta-sovereignty, as used here, is the operational counterpart: the concrete authority to decide, and revise, who or what makes decisions in a given workflow.

From Implementation to Orchestration

Traditional software encouraged a transactional mindset: inputs produced predictable outputs, and improvement meant optimizing those transactions for speed or cost. Generative AI breaks this model. Optimization alone is no longer sufficient.

What is required instead is orchestration: the deliberate design of decision architectures that determine when humans lead, when machines assist, and when autonomy is acceptable.

This reframes the core question of AI adoption. It is no longer "How can this system perform better?" but "How should decision-making authority be distributed in this situation?" Every redistribution of agency carries consequences. These are not technical side effects; they are strategic choices. Orchestration makes those choices explicit.

Why a Strategic Framework Is Necessary

Organizations already distribute agency – constantly and often unconsciously. Tools are adopted, workflows evolve, and autonomy expands without explicit decisions about authority. Over time, these implicit choices harden into structural patterns that are difficult to reverse.

Recent survey data suggests that this pattern is widespread. Roughly one third of organizations report using AI to deeply transform products, processes, or even business models, another third are redesigning key processes around AI without changing the underlying models, and the remaining third apply AI only at the surface with little or no change to existing processes (Deloitte, 2026). In other words, most organizations are still optimizing around existing decision architectures rather than consciously redesigning who decides what, and when.

Without a shared framework, agency distribution remains invisible. Teams do not lack effort or intelligence; they lack a common way to recognize, discuss, and intentionally design how decisions are made. A strategic framework is therefore required not to prescribe a single "correct" configuration, but to make these choices visible, discussable, and

deliberate – and to ensure that judgment is exercised consciously.

Existing research on human-AI agency – design spaces such as Holter and El-Assady (2024), organizational perspectives like Krakowski (2025), governance frameworks for responsibility distribution – provides rich conceptual foundations but remains primarily descriptive or operates at policy level. The Agency Continuum translates these insights into a small set of discrete, named modes with explicit trade-offs, coupled with a decision logic that teams can apply in everyday work.

This orientation complements regulatory efforts such as the EU AI Act, which classifies AI systems by risk level and stipulates oversight requirements. Those regulations determine *which* applications require scrutiny; they do not specify *how* decision-making authority should be allocated inside teams and workflows. Most routine knowledge work falls outside the Act's scope entirely, yet still benefits from explicit agency design. The question this framework addresses is orthogonal: not "Is this application high-risk?" but "Who should hold decision-making power – and under what conditions can that change?"

The framework is designed primarily for knowledge-work contexts where organizations have discretion over how human and machine decision-making is configured – strategy, product development, creative work, analytics, operational decisions. High-risk applications governed by sector-specific regulation require additional safeguards; in those domains, the framework complements rather than replaces formal requirements.

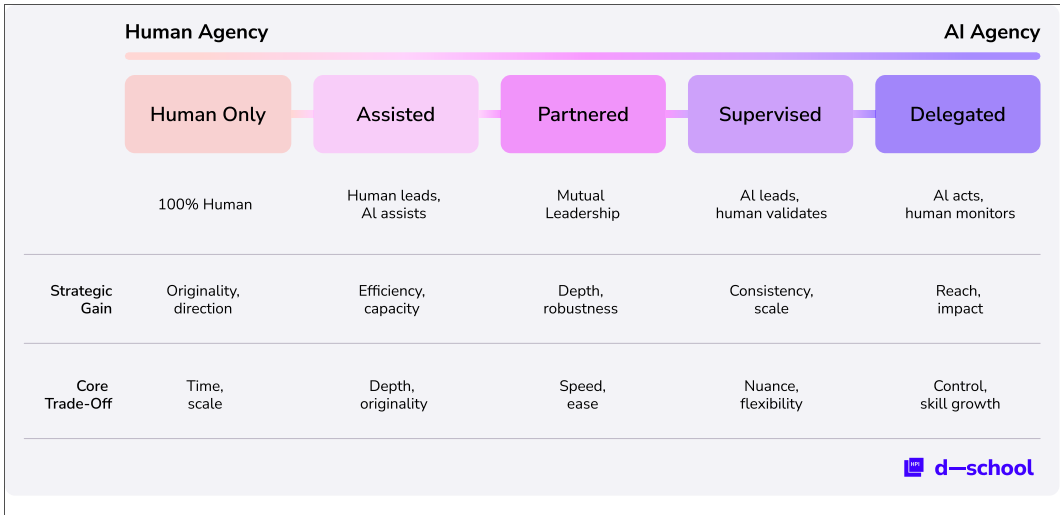
With this positioning in mind, the next section introduces the Agency Continuum itself: five modes of strategic agency that make the question of "who decides" operational.

III. The Framework: Five Modes of Strategic Agency

If agency can be consciously allocated, it requires explicit configurations. The Agency Continuum translates this principle into five operational modes. These are not maturity stages – Mode 5 is not "better" than Mode 1. They are strategic options, each suited to different contexts, each carrying distinct advantages and costs. The skill lies in choosing the right mode for the situation at hand.

The Continuum at a Glance

The following framework maps five configurations of human-AI agency, each defined by where decision-making power resides.



Each mode offers something valuable. Each mode costs something real. There is no free lunch on the continuum – only conscious choices about which value to prioritize and which cost to accept.

The five modes align with established governance patterns – human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC) – but extend them by specifying not only where decision-making power resides, but also the strategic gain, core trade-off, and (in Section IV) the preconditions and warning signs for each configuration.[5]

Now let's examine each mode in detail: what it is, when to use it, what it offers, what it costs, and the critical skill it requires.

[5] The framework synthesizes insights from multiple research traditions. Suchman's work on situated actions (2007) established that human-machine boundaries emerge from context rather than fixed properties – hence five modes rather than one prescription. Deci and Ryan's Self-Determination Theory (1985) explains why protecting human-only work matters for motivation and well-being. Frankfurt's concept of second-order volition (1971) provides the philosophical basis for meta-sovereignty. Recent work on human-AI agency and mixed-initiative systems (e.g., Holter & El-Assady, 2024) maps how control can be distributed in technical systems; the Agency Continuum shifts this lens from system design to organizational strategy.

Mode 1: Human Only – *The Sanctuary*

Configuration: Work without generative AI. Human expertise, intuition, and judgment drive all decisions.

When to Use: Work where strategic direction must be established before execution begins. Decisions carrying moral weight where human accountability is non-negotiable. Situations requiring deep empathy and direct human connection. Building foundational skills that will enable evaluation of AI outputs later. Defining what success means before optimizing for it.

Strategic Gain: Originality and direction. Work that could only have come from this person, this team, this organization.

Core Trade-off: Time and scale. Human Only is slow. It does not scale. That is the price of originality.

Why It Matters: This mode protects what cannot be delegated: the formation of judgment through direct experience, the capacity to verify AI outputs later, and the differentiation that comes from work untouched by statistical averaging.

The rationale is not that humans decide better – often they don't. It is that humans decide *differently*. Diverse humans make diverse errors; identical models converge toward identical ones. And judgment develops only through the act of deciding. Human Only mode protects the conditions under which good judgment forms.

Therefore, the most counterintuitive but strategically critical advice is sometimes: *don't use the tech.* This is not a mainstream view in efficiency-obsessed markets; it is a deliberate counter-model that protects the space where distinctive human competence is built. This is not Luddism. Mode 1 is consciously choosing not to use a tool for a specific task because the task requires it – because the struggle itself builds capability.

Empirical evidence supports this logic. A 2026 study, for example, found that developers learning with AI assistance scored 17 percent worse on conceptual understanding, without meaningful time savings (Shen & Tamkin, 2026). The cause was not AI itself, but delegation of problem-solving. The effect extends beyond novices: a clinical study reported that experienced radiologists' ability to detect tumors without AI dropped by approximately 6% after several months of AI-assisted diagnosis (International AI Safety Report, 2026). The skill erosion affects both formation and maintenance, which is why the conclusion suggests itself that when skill formation is the objective, working without AI is investment, not rejection.

However, there is an elephant in the room: who pays for the Sanctuary? If Assisted mode produces a "good enough" strategy document in ten minutes versus three days of expensive human labor, the market incentivizes speed. This is why protecting the Sanctuary is a leadership responsibility. Organizations that let market pressure erode Mode 1 will find themselves unable to produce anything distinctive – and eventually, unable to evaluate whether their AI-generated work is any good.

The Critical Skill: Sense-making. Discerning what truly matters through direct human perception. The patience to sit with uncertainty rather than rushing to an AI-generated answer.

Mode 2: Assisted – *The Accelerator*

Configuration: Human maintains full decision-making authority. AI provides targeted support – generating drafts, processing data, creating variations. The human directs; the AI executes specific sub-tasks.

When to Use: Content generation at volume. Data processing and summarization. Research synthesis. Routine communication. Repeatable tasks with clear quality standards. Any situation where the task is well-defined and the question is "How can we do this faster?"

Strategic Gain: Efficiency and capacity. More output in less time.

Core Trade-off: Depth and originality. Assisted mode optimizes what exists. It rarely creates what doesn't.

Why It Matters: This mode addresses the Fatigue Crisis – if used deliberately. The danger is making it the default for everything. Mode 2 excels at freeing cognitive capacity for higher-value work. It fails when applied to tasks requiring originality (use Mode 1) or depth (use Mode 3). Most fatigue originates here: applying Assisted mode to tasks too complex for simple assistance, then paying the Supervisor's Tax in endless verification.

The key insight: Assisted mode works best when you could do the task yourself but choose not to for efficiency reasons. If you couldn't do the task yourself, you cannot effectively direct the AI or evaluate its output. You become dependent on AI while unable to assess whether it is serving you well.

The Critical Skill: Directive clarity. Being a good boss to your AI. Vague instructions produce mediocre outputs. Directive clarity means knowing exactly what you want before you prompt.

Mode 3: Partnered - *The Crucible*

Configuration: Human and AI collaborate iteratively. The human brings context, domain expertise, strategic judgment. AI brings pattern recognition, breadth, tireless iteration. Neither could achieve the outcome alone.

When to Use: Complex problem-solving where neither human nor AI could succeed independently. Quality optimization when "good enough" isn't good enough. Strategic analysis requiring synthesis of multiple perspectives. Any situation where the answer requires genuine intellectual friction.

Strategic Gain: Depth and robustness. Outcomes better than either human or AI could produce alone.

Core Trade-off: Speed and ease. Partnered mode is cognitively demanding. It requires time and energy.

Why It Matters: This mode breaks through the imagination ceiling by establishing AI as a genuine thinking partner. The difference from Assisted mode is fundamental: in Assisted mode, you prompt and receive. In Partnered mode, the conversation goes deeper: "Here's my strategic plan. Tell me the three biggest reasons it will fail." The AI responds, you push back, the AI reanalyzes. This is dialectic.

True partnership requires multiple rounds of meaningful iteration. If you prompt once and accept the answer, you are in Assisted mode. If the process feels effortless, you are not doing Partnered mode correctly. The value comes from friction[6].

A critical caveat: genuine sparring requires a distinct perspective. If you prompt generically, you receive generic responses – an echo chamber, not a challenger. Breaking out requires providing specific proprietary context. But here is the paradox: the more context you provide, the more the AI can tailor responses to what you want to hear. You must actively prompt for disagreement, for the case against your position.

The Critical Skill: Collaborative thinking. Understanding what you uniquely bring and what AI uniquely brings. Intellectual humility combined with intellectual confidence.

Mode 4: Supervised - *The Gatekeeper*

Configuration: AI generates complete outputs. Humans review systematically, validating against explicit criteria before outputs go live. The AI produces at scale; the human ensures quality at checkpoints.

When to Use: Scaling proven processes. Content production at volume with consistency requirements. Any context where tenfold output is needed while maintaining quality standards. Situations where the process is well-established and the question is "How can we scale this?"

Strategic Gain: Consistency and scale. High volume without proportional increase in human effort.

Core Trade-off: Nuance and flexibility. Scaled processes optimize for the common case. Edge cases suffer.

Why It Matters: This mode shifts the human role from maker to curator. Instead of writing one email, the AI writes five hundred – and your job is to verify them before they are sent. But this only works with explicit criteria. If you cannot articulate what makes a good email, you have no business operating in Supervised mode. The failure to define is the failure to scale.

Mode 4 requires criteria discipline: the willingness to invest time upfront defining exactly what success looks like, so that review becomes evaluation against standards rather than intuition. Clear criteria enable faster review, more consistent quality, and the eventual ability to delegate the review itself as standards mature.

Effective supervision also depends on verifiability. Outputs that can be checked quickly against clear standards – code that compiles, translations in a language you speak, calculations that can be probed – allow genuine quality control. If verification requires expertise

[6] A common concern: if both lead, who is accountable? The answer: the human – always. Partnered mode is not co-leadership with shared authority. The roles are complementary, not symmetrical. AI expands the space of possibilities – surfacing alternatives, challenging assumptions, stress-testing logic. The human navigates that space – judging which insights matter, weighing trade-offs, making the call. The back-and-forth may feel like dialogue between equals, but the human remains the author of the decision.

the reviewer lacks, or takes as long as producing the output yourself, Supervised mode risks becoming rubber-stamping under a different name.

The Critical Skill: Quality discernment. Translating tacit quality standards into explicit, teachable criteria. "I'll know it when I see it" does not scale.

Mode 5: Delegated - *The Autopilot*

Configuration: AI operates autonomously within defined constraints. Humans set boundaries, establish success metrics, monitor performance, intervene only on deviations. The human designs the system; the AI runs it.

When to Use: Algorithmic systems requiring speed beyond human capacity. Personalization at massive scale. Real-time optimization. Monitoring and alerting. Any context where the volume or velocity of decisions exceeds human processing ability, and the stakes of individual decisions are manageable.

Strategic Gain: Reach and multiplication. Impact at scales impossible through human effort alone.

Core Trade-off: Control and learning. Autonomous systems operate beyond direct oversight. Skills that are never exercised atrophy.

Why It Matters: Here, meta-sovereignty becomes concrete. You are not making each decision. You are designing the system that makes decisions on your behalf. Every action the AI takes, you own – because you designed the conditions. Think of it like an aircraft autopilot: the system flies, but a pilot sits in the cockpit, monitoring instruments, ready to intervene. The autopilot extends capability, but the pilot never stops being responsible.

Mode 5 is not "set and forget." It requires ongoing monitoring, regular audits, and clear tripwires – rules that tell the system when to stop and escalate. The test for Mode 5 readiness: "If this AI makes a mistake, can I live with the consequences?" If the answer is no, pull back to Mode 4.

Delegation also requires that outcomes remain legible. If you cannot detect when the system fails – because failures are subtle, delayed, or require expertise you lack – you cannot delegate responsibly. The question is not just "Can I live with errors?" but "Will I know when errors occur?"

The Critical Skill: System design. Defining constraints, establishing tripwires, building monitoring, accepting accountability for everything the system does.

In increasingly agentic AI ecosystems, the stakes of this system-design role become even higher. Autonomous agents that can call tools, trigger other agents, or modify their own objectives can amplify both value and risk at system scale. Delegated mode therefore requires not just local monitoring, but infrastructure-level tripwires and observability: logging, auditing, and coordination mechanisms that make failures detectable and recoverable across agent networks, not only within a single application.

Delegated mode represents the furthest point at which human meta-sovereignty still exists. Beyond this, agency is no longer orchestrated; it is abandoned.

The Constant Role: The Orchestrator

Across all five modes, one role remains constant: the orchestrator. Whether working alone in Human Only, iterating with AI in Partnered, or designing autonomous systems in Delegated, the human remains responsible for consciously allocating agency, shaping the relationship, and owning the outcome. The skills shift – from soloist to conductor to composer – but the accountability never transfers.

Knowing the modes is the first step. Choosing the right one for a given situation is where strategy becomes practice. The following section provides a decision logic for making that choice systematic; and for recognizing when a choice is no longer working.

IV. The Compass: Choosing the Right Mode

The previous section defined what each mode is. This section addresses when each mode fits – and how to choose.

Mode selection is where strategy meets practice. The right mode, chosen deliberately, prevents the crises described in Part I. The wrong mode, chosen by default, produces them. The difference is rarely effort or intent. It is method.

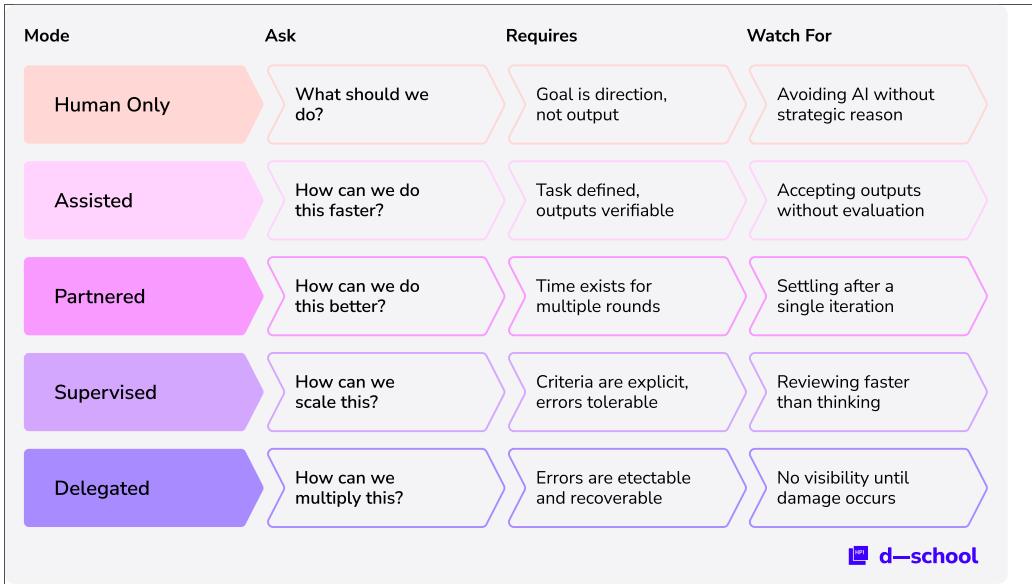
The Mode Decision Logic provides that method. It structures mode selection as a three-step process and provides a systematic response when things go wrong.

The Mode Selection Logic

The logic consists of three steps: **Ask, Requires, Watch For.**

- **Ask** identifies the question you are trying to answer – and points to a candidate mode.
- **Requires** checks whether the preconditions for that mode are actually met.
- **Watch** For monitors execution for signs that the mode no longer fits.

The following table integrates all three:



The table serves as a diagnostic tool. It identifies fit between situation and mode – not preference, but appropriateness.

Step 1: Ask - *Identifying the Real Question*

Every mode answers a different question. The first step is identifying which question your situation actually requires.

This sounds obvious. In practice, it is routinely skipped. Teams default to "How can we do this faster?" when the real need is "What should we do?" They ask "How can we scale?" before asking "How can we do this well?" The mismatch between question and mode is the origin of most failures.

Human Only answers: *What should we do?* Use this mode when strategic direction must be established – when the goal is clarity, not speed.

Assisted answers: *How can we do this faster?* Use this mode when the task is defined and the goal is efficient execution.

Partnered answers: *How can we do this better?* Use this mode when quality matters more than speed, and the problem benefits from genuine intellectual friction.

Supervised answers: *How can we scale this?* Use this mode when a proven process needs to reach higher volume without proportional effort.

Delegated answers: *How can we multiply this?* Use this mode when impact must extend beyond what any human process could achieve.

Before selecting a mode, pause and ask honestly: What question am I actually trying to answer? The honest answer points to a candidate mode.

Step 2: Requires – *Validating the Preconditions*

Each mode has preconditions. If they are not met, the mode will fail regardless of how appealing it seems.

Human Only requires that strategic direction is genuinely the goal – that the situation calls for establishing clarity before optimizing execution. If direction is already clear, Human Only becomes inefficient resistance to useful tools.

Assisted requires that outputs can be verified quickly by the user. If verification takes as long as production, the efficiency gain disappears. If verification requires expertise the user lacks, the user becomes dependent on outputs they cannot evaluate. This is the path to the De-Skilling Crisis.

Partnered requires sufficient time for multiple rounds of genuine challenge. Partnership is not a single prompt-response cycle. It is iterative refinement through friction. Without that time, Partnered collapses into Assisted – first drafts accepted as final.

Supervised requires explicit quality criteria and defined error tolerance. Without explicit criteria, review becomes intuition at scale – which is rubber-stamping by another name. Without defined error tolerance, the human cannot know when to intervene.

Delegated requires that errors are detectable, recoverable, and that consequences are acceptable. If failures are invisible until damage is done, or if consequences exceed acceptable bounds, delegation becomes negligence. The question is not just "Can I live with errors?" but "Will I know when errors occur?"

The *Requires* step is a filter. It validates whether the candidate mode can actually succeed in the current situation. Unmet preconditions do not mean "try harder." They mean "wrong mode."

Step 3: Watch For – *Recognizing Drift*

Even appropriate initial choices can drift. Circumstances change. Time pressure mounts. What began as genuine partnership becomes acceptance of first drafts. What began as supervised review becomes mechanical approval.

Warning signs indicate that a mode no longer fits. Recognizing them requires ongoing attention, not just initial selection.

In Human Only, watch for avoiding AI from habit rather than strategy. The mode is correct when direction genuinely needs to be established. It is misused when it becomes resistance to change – protecting comfort rather than building capability.

In Assisted, watch for accepting outputs without meaningful evaluation. If you find yourself approving AI work you could not have produced or verified yourself, you have slipped from using the tool to being used by it.

In Partnered, watch for settling after a single iteration. The phrase "looks good to me" after one round is the death of partnership. True collaboration requires friction. Comfort is a warning sign, not a success indicator.

In Supervised, watch for reviewing faster than thinking. If approval takes seconds when genuine evaluation would take minutes, the human checkpoint has become theater. You are rubber-stamping, not supervising.

In Delegated, watch for having no visibility until damage occurs. If failures surface only as surprises – angry customers, broken systems, reputational harm – monitoring has failed. You are not delegating. You are abdicating.

When warning signs appear, the response is not to try harder in the same mode. It is to reassess and adjust.

When Modes Fail: Patterns and Corrections

The crises described in Part I emerge from predictable mode failures. Understanding these patterns enables faster recognition and deliberate correction.

Pattern: Acceleration without Direction Using Assisted mode when Human Only is required. The team prompts for speed before establishing what they actually want. Output volume is high, but conviction is low. Everything feels productive, but nothing feels right.

The correction: Stop prompting. Define intent, success criteria, and non-negotiables without AI assistance. Return to Assisted only after direction is clear.

Pattern: Accepting "Good Enough" on Hard Problems Staying in Assisted mode when Partnered is required. Complex, strategic work receives first-draft treatment. Outputs are accepted not because they are excellent, but because they are fast.

The correction: Introduce deliberate friction. Require iterative challenge – counter-arguments, stress-tests, explicit disagreement – before accepting any output as final.

Pattern: Scaling Without Standards Entering Supervised mode without explicit criteria. Review becomes intuition at volume. Quality varies. Failures emerge downstream, often invisibly at first.

The correction: Return to Partnered mode. Develop explicit, teachable criteria through iteration. Answer the question "What does good look like?" in operational terms. Only then re-enter Supervised mode.

Pattern: Delegation Without Detection Operating in Delegated mode without monitoring or tripwires. The system runs autonomously. Errors accumulate unnoticed. Responsibility becomes diffuse. Surprises replace oversight.

The correction: Pull back to Supervised mode. Add human checkpoints. Define what signals would indicate failure. Build detection mechanisms before restoring autonomy.

Pattern: Abandoning the Sanctuary Eliminating Human Only mode entirely under efficiency pressure. Every task gets AI assistance. Judgment atrophies. Evaluation capability erodes. The organization becomes unable to distinguish good AI output from bad.

The correction: Explicitly protect non-augmented work. Create time and space for human-only sense-making. Treat it as capability investment, not inefficiency.

The Fallback Principle

When *Requires* is not met, or when warning signs persist, the appropriate response is adjustment, not abandonment. The goal remains; the mode adapts.

The principle is simple: when in doubt, move one mode toward human control.

- Assisted failing → Human Only (build evaluative skill first)
- Partnered failing → Assisted (accept less depth, or create time)
- Supervised failing → Partnered (develop explicit criteria)
- Delegated failing → Supervised (add human checkpoints)

This is not retreat. It is recognition that modes have requirements. Moving toward human control preserves optionality while conditions for more autonomous modes are established.

Example: Product Positioning

A product team needs new market positioning. The deadline is tight. The instinct is familiar: prompt an AI for options, pick one quickly, move to execution.

This is Assisted mode by default – and it is the wrong choice.

Ask: The team treats the question as "How can we do this faster?" But the real question is "What should our position be?" That is a Human Only question. No amount of AI-generated options will help if the team doesn't know what they're looking for.

Requires: Is direction genuinely unclear? Yes – the team has no shared conviction about what the positioning should achieve, who it's for, or what success looks like. The precondition for Human Only is met. The precondition for Assisted – verifiable outputs – is not, because without clear direction, there is no basis for evaluating options.

The team recognizes the mismatch. They pause AI use and spend a difficult afternoon in debate. What does the market actually need? What can the company credibly claim? Where is the differentiation? The conversation is slow, sometimes contentious, but it produces something AI cannot: shared conviction about direction.

Now the question shifts. The team asks: "How do we stress-test this direction?" This points to Partnered mode.

Requires: Does sufficient time exist for multiple rounds? The team allocates two days for iteration. They proceed with Partnered mode – challenging the positioning with AI, exploring objections, pressure-testing assumptions. The AI surfaces weaknesses the team hadn't considered. The team pushes back on AI suggestions that miss context. After several rounds, the positioning is sharper than either could have made it alone.

After validation, the question shifts again: "How do we communicate this at scale?" This points to Supervised mode.

Requires: Are criteria explicit? Not yet. The team cannot clearly articulate what "on-brand" means in operational terms. Rather than proceeding anyway, they return to Partnered mode to develop explicit criteria. What words are in-bounds? What tone is required? What claims are off-limits? Once criteria are explicit, they re-enter Supervised mode with a foundation for genuine quality control.

Watch For: During scaled production, review times accelerate. The team notices they are approving assets in seconds. This is a warning sign – reviewing faster than thinking. They pause, recalibrate the process, and restore meaningful evaluation before continuing.

The sequence took longer than the original "prompt and pick" plan. It also produced positioning the team believed in, communication that held together at scale, and quality that survived contact with the market.

The difference was not effort. It was method.

V. From Framework to Practice

Understanding the Agency Continuum is only the first step. Value is created when the framework shapes real decisions – under time pressure, under uncertainty, and under competing incentives.

This section addresses how the framework is intended to function in practice. Not as a compliance mechanism or governance overlay, but as a shared orientation tool for teams navigating complex human–AI work.

What the Framework Is Designed to Do

The Agency Continuum is designed to serve three distinct purposes.

First, it creates a **shared language.** Without explicit vocabulary, teams talk past each other. One person says "let's use AI" meaning assisted drafting; another hears autonomous execution. Explicit modes eliminate this ambiguity.

Second, it forces **conscious choice.** By framing work in terms of decision-making authority rather than tools, the framework interrupts default behavior. It makes agency distribution a deliberate act rather than a side effect of convenience.

Third, it protects **human judgment as a strategic asset.** By explicitly reserving space for human-only and partnered modes, the framework counters the tendency to optimize away the very capabilities organizations depend on for differentiation.

In hybrid human-AI systems, leadership shifts from supervision to architecture: designing who decides, under which conditions, and with what constraints. The framework does not aim to standardize behavior. It aims to make judgment unavoidable.

Applying the Framework as a Team

For teams, the framework functions best as a lightweight coordination mechanism rather than a formal process. It translates the Mode Decision Logic into a small set of shared questions that teams can use to align before and during a project. Before beginning a task or project, teams can align by answering four questions:

1. **What question are we actually trying to answer? (→ Ask)**
   Is the primary need direction, speed, quality improvement, reliable scale, or systemic reach?
2. **Are the preconditions for that mode met? (→ Requires)**
   Based on the required outcome, where should decision-making authority sit on the continuum?
3. **What warning signs should we watch for? (→ Watch For)**
   What concrete signals would indicate that this configuration is no longer serving the work?
4. **What would trigger a shift toward more human control?**
   At what point should authority be redistributed, and on what basis?

These questions are intentionally simple. Their value lies not in precision, but in making assumptions explicit. When teams share an understanding of which mode they are operating in – and why – they coordinate more effectively and avoid working at cross-purposes.

Importantly, the framework does not assume static mode selection. Many tasks will move across modes as they evolve. The risk is not movement, but unconscious drift – remaining in a mode long after its strategic fit has expired.

Agency Awareness as a Leadership Discipline

Before mastering individual modes, leaders must develop the capability that activates everything else: Agency Awareness, the ability to recognize which mode fits which context, and the discipline to choose deliberately rather than defaulting.

The crisis comes from drifting. The solution is steering.

Some of us become control freaks, staying in Assisted mode for everything because we don't trust the system. Others push too far in the opposite direction, defaulting to Supervised or Delegated modes – the *"just let the bot handle it"* impulse. Both are failures of agency awareness.

Agency Awareness means taking the wheel; even if you put it back on autopilot five minutes later. The point is not constant control, but conscious choice. It means catching yourself in moments like: *"I'm defaulting to Assisted because it's easy, but this task actually requires originality. I need to close the laptop and switch to Human Only."*

Or the reverse: *"I'm spending hours writing routine emails when I should have designed a Supervised system and freed my time for more important work."*

The skills this requires – setting clear goals, giving effective feedback, defining success criteria, knowing when to intervene – are not new. They are fundamental management skills. As Ethan Mollick observes (2026), "the skills that are so often dismissed as 'soft' turned out to be the hard ones."

This emphasis on skills is echoed at policy level: the World Economic Forum and McKinsey argue that 'brain skills' – adaptability, complex problem-solving and communicative judgement – are central to capturing the benefits of AI and must be actively cultivated across organizations (World Economic Forum & McKinsey Health Institute, 2026).

So, when everyone has access to an army of tireless AI agents, advantage no longer comes from technical cleverness or prompt engineering. It comes from knowing what good looks like; and being able to articulate it clearly enough that even a machine can deliver it. That capability is not technical. It is judgment.

The Discipline Paradox

Here's the uncomfortable truth: The framework requires discipline. But the crises we've described – especially fatigue – have already eroded the cognitive capacity for discipline. We're asking overworked, burnt-out employees to make careful, conscious choices about

mode selection. That's a heavy ask.

This is why agency awareness cannot remain an individual responsibility alone. Individual discipline cannot overcome organizational dysfunction. If the company rewards speed over quality, individuals will default to Mode 2 regardless of their intentions. If leadership doesn't protect Mode 1 time, market pressure will eliminate it.

Protecting agency is ultimately a leadership responsibility: creating structures, incentives, and cultural permission to choose the right mode, even when it's slower. The path of least resistance will always be to let the machine do it. Leaders must make the deliberate path viable.

The Orchestrator's Question

In 2026, everyone has access to excellent AI. In many knowledge-work contexts, Claude, ChatGPT, Gemini etc. are rapidly becoming commodities, like electricity. Your competitor can prompt the same models you can.

What cannot be commoditized is judgment: knowing when to protect human agency, when to partner, when to scale, when to design autonomous systems. The organizations that thrive will not be those that use AI the most. They will be those that orchestrate agency most strategically.

But this practical advantage points to something deeper.

The concept of second-order volition – having authority over your own will – raises a question we can no longer avoid. The ultimate risk of the de-skilling crisis isn't just that we forget how to write good code or structure a good argument. It's that we lose the ability to consciously decide how we work. And if we lose that, do we eventually lose the ability to decide what we want?

Consider the trajectory. If the system suggests our goals for the quarter, and the system executes all the tasks to achieve those goals – where are we in that equation? What part do we play? The risk is the atrophy of human will itself.

Which brings us to the question at the heart of this framework. The final question it asks is not about productivity or efficiency.

It's this: Are you the orchestrator of this relationship – or just part of the performance?

Remember where we started: the promise of productivity, the reality of exhaustion. The math didn't work because we were solving the wrong equation.

The right equation isn't about tools. It's about agency. It's about sovereignty. It's about the conscious, strategic distribution of decision-making power between humans and machines.

Master that equation, and the math starts working again.

References

Empirical Research Cited

Anthropic. (2026). Anthropic Economic Index, January 2026.

ASGR. (2025). Agentic Systems Governance Report.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at Work. NBER Working Paper.

Crowston, K. et al. 2025. Deskilling and upskilling with generative AI systems. arXiv preprint.

Dykes, B. 2026. Why AI's Productivity Promise Falls Apart Without Human Expertise. Forbes, 27 Jan.

Deloitte. (2026). State of AI in the Enterprise: The Untapped Edge. Deloitte AI Institute, January 2026.

EY. (2024). Pulse Survey of Senior Executives on AI Implementation.

Epstein, Z. et al. (2025). AI as Cognitive Amplifier. arXiv preprint.

Fan, D., Delsad, S., Flammarion, N., & Andriushchenko, M. (2026). HALLUHARD: A Hard Multi-Turn Hallucination Benchmark. arXiv:2602.01031.

Fortune / S&P Global Market Intelligence. (2025). Analysis of AI Initiative Outcomes.

Hildebrandt, C. et al. (2025). Human-AI collaboration is not very collaborative yet. Frontiers in Computer Science.

IBM. (2025). AI at the Core: CIO Perspectives on Risk and Governance.

International AI Safety Report. (2026). International AI Safety Report 2026. Chair: Yoshua Bengio. February 2026. https://internationalaisafetyreport.org

ITU. (2025). Annual AI Governance Report.

Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science.

Quantum Workplace. (2025). Employee Engagement Trends: Employee Experience.

Section. (2026). The AI Proficiency Report: Leaders think their AI deployments are succeeding. The data tells a different story. January 2026.

Shen, J. H., & Tamkin, A. (2026). How AI Impacts Skill Formation. arXiv:2601.20245.

Stanford University. (2025). AI Index 2025 Report.

Sternfels, B. (2026). McKinsey CEO remarks on AI agent deployment. January 2026. Online.

Upwork Research Institute. (2024). Future of Workforce Report: AI and Productivity.

Utley, J. (2026). The Imagination Ceiling: Why 70% of Your People Are Stuck on AI. Stanford d.school.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour.

Wiley Workplace Intelligence. (2025). The Cascade Crisis Report.

WINSS Solutions. (2025). Model Collapse and Data Saturation in AI Systems.

World Economic Forum & McKinsey Health Institute. (2026). The Human Advantage: Stronger Brains in the Age of AI. Insight Report.

Theoretical Foundations

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Plenum.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. Journal of Philosophy, 68(1), 5–20.

Suchman, L. A. (2007). Human-machine reconfigurations: Plans and situated actions (2nd ed.). Cambridge University Press.

Agency Distribution and Human-AI Collaboration

Holter, S., & El-Assady, M. (2024). Towards Agency in Human-AI Collaboration: Design Space and Emerging Patterns. Computer Graphics Forum, 43(3).

Krakowski, S. (2025). Human-AI agency in the age of generative AI. Information and Organization, 35(1).

Mollick, E. (2026). Management as AI Superpower. One Useful Thing (Substack).

Zhang, Y. et al. (2024). Exploring Collaboration Patterns and Strategies in Human-AI Co-creation through the Lens of Agency. Proceedings of the ACM on Human-Computer Interaction.

Governance and Responsibility Frameworks

Eitel-Porter, R. et al. (2024). HAIG: A Human-AI Governance Framework for Trust and Utility. AI and Ethics.

Köbis, N. et al. (2024). Distributing Ethical Responsibility in Hybrid Human-AI Systems (ERDEM). Nature Machine Intelligence.

High-Level Expert Group on AI. (2019). Ethics Guidelines for Trustworthy AI. European Commission.

HPI d–school – Creating the future through innovation and Design Thinking at the Hasso Plattner Institute

To learn more just visit our website:
www.hpi-dschool.de