Whose Ally Is 'Your' Agent?

'Computational Egonomics' and the Allocation of Human Agency

Michael Schrage MIT Sloan School of Management · Initiative on the Digital Economy
MIT Sloan Working Paper

Abstract. The dominant narrative about agentic AI impact keeps score in capability, atrophy and displacement. This paper argues that the consequential impact lies elsewhere: in the allocation of human agency. Every delegated action reinforces some human capacities and erodes others. Drawing on Thomas Schelling's classic *'Egonomics* or The Art of Self-Management,' we model the self — not one decision-maker but several, present and future, sharing a body but not always a preference — as a portfolio of strategic players. We assert the Quantified Self movement failed because it measured the outputs of this game without ever instrumenting its players. We specify an alternative: a 'Multiple Selves' architecture in which commitment devices generate their own evaluation data, governance is intra-personal mechanism design, and the central claim is rendered measurable. The defining competition in agentic AI is not 'model capability;' it is the architecture of measurement and feedback wrapped around the human-agent interaction. There is no 'neutral default.' We recognize this approach has enormous implications for 'behavioral economics' and finance, as well as for organizational design.

Keywords: agentic AI; egonomics; human capital; mechanism design; human-AI interaction; agency.

I. The Allocation Nobody Is Measuring

Past midnight before the earnings call: The CFO clears their queue of seventeen agent-generated recommendations — approve, approve, approve, decline — and goes to bed. The enterprise has just been governed. The question no dashboard asked was: just which 'self' did the governing? Not which person — the CFO's approvals are not in doubt — but which version of them? The 'strategic

self' that set the quarter's priorities? Or the 'exhausted self' that wanted the queue empty? These are the human questions agentic AI forces and almost no enterprise system is instrumented to see. Bluntly: the most important impact of agentic AI will not be 'task automation' at scale; it will be the reallocation of human agency - not autonomy from the agent but the capacity to revise one's own commitments rather than simply act on them.

The techno-futurists keep score in capability and displacement. They debate whether 'autonomous' agents will book the travel, close the trades, write the code, run the supply chains, and how many knowledge workers will 'retire' as they do. This framing inverts consequence: Agentic systems do not merely anticipate, coordinate, and execute decisions; they're alternately chartered and harnessed to allocate attention, initiative, memory, judgment, and deliberation across competing modes of human conduct. Every agentic interaction quietly reinforces some human capacities and lets others shrivel. Every delegation is also a choice about what kind of judgment gets strengthened, what kind gets weaker, and who the human becomes — for individuals and for teams.

However, the deeper and more consequential design question is not *what can the agent do?* but *which human capabilities does the agent reinforce, undermine, or invite?* An agent that removes friction from impulsive decisions cultivates a measurably different human future than one that slows the principal long enough to provoke or evoke - reflection. Agents engagingly optimized for 'convenience' can systematically weaken strategic judgment even as they measurably raise productivity. Agents built to elicit assumptions, commitments, and tradeoffs may feel slower at first while compounding human agency over time.

The crucial determinant is not agent capability; it is the intentional architecture of measurement and feedback wrapped around the interaction. That architecture has a name in waiting — *the quantified selves*, plural, governed, reflective — and a paradigm old enough to carry a few Nobel Prizes and new enough to be buildable only now: *computational egonomics*. Whether agentic, generative, and predictive AI capabilities are provisioned to serve that architecture - or automatically bypass it - is the intelligent 'intelligence' question, challenge and opportunity. Every organization seeking to productively deploy agentic AI at scale needs to understand that.

Personal informatics, commitment-device research, behavioral design, AI alignment, and people analytics each address important aspects of self-regulation and delegation. This contribution is different: an architecture that treats the 'allocation of agency' itself as the primary object of measurement and governance. Agentic AI exposes a hidden and flawed assumption embedded in classical principal-agent theory: that the principal is singular. Computational egonomics begins from the opposite premise: The fundamental challenge is not ensuring agents faithfully serves their principal; it is determining *which* principal those faithful agents serve. With apologies to Walt Whitman, principals are large and contain multitudes. When is 'human agency' more bug than feature – and vice versa?

II. The Agent Is a Player, Not a Tool

Nobelist and strategist Thomas Schelling named the problem classical economics could not formalize. His 1978 'Egonomics, or the art of self-management' article boldly argued that strategic interactions between a person's present self and the future selves who must live with the present self's

commitments is not a metaphor but a non-cooperative game — played by selves who share a body but not their preferences. The dieter who buys the ice cream and the dieter who eats it at midnight are different strategic actors, with different information sets, time horizons, and utility functions. 'Nudge' Nobelist Richard Thaler's 'planner-doer' model proffers the same insight: choice architecture exists precisely because the 'self that chooses' and the 'self that bears the consequences' are not the same self. Schelling's 'commitment devices' — Odysseus lashed to the mast, the hidden cigarettes, the Christmas Club account — were analog instruments of this 1970s paradigm. They 'worked' – sort of - but were static, single-behavior, and left almost no analyzable record of the changes they induced. Schelling's paradigm was 'correct' but its instrumentation was primitive.

Ongoing ontological shifts cut even deeper: Contemporary digital discourse treats 'the agent' and 'agents' as neutral infrastructure that users seek to effectively deploy. Egonomics asserts something edgier: the agent is now a *player* in the non-cooperative game among a person's selves. In other words, the agent is not merely deployed; it is *assigned* — by default or by design — to an alliance. That agent booking that impulsive meeting has not served 'the user;' it has effectively – affectively? - allied with the 'impulsive self' against the strategic self in a game the principal did not know was running. This paper's governing – and governance – question: whose ally is the agent, and on what authority? Which 'self' runs the show? How do we know? Better yet, how can we measure?

AI's evolving capabilities radically transform the economics – and revealed preferences – of both computational agents and human agency. A non-cooperative game is empty without its players — and the framework here must pay its own bill rather than borrow the authority of the formalism. Who names the selves? Not the 'theorist' by stipulation; that is 'Quantified Self' astrology. Nor is there a sovereign self standing outside the portfolio to ratify the rest — that 'self' is the oldest fiction in the literature, and looking for it only relocates the problem. Our resolution: stop looking!

Selves are not declared and then approved; they are *revealed by enforcement*. A 'self' exists, for this architecture, exactly insofar as it can 'bind' the others: the midnight self is real because it overrides the morning's commitment and the strategic self is real because it builds a 'commitment ledger' that overrides the midnight one. The agent roster is therefore not a list the principal certifies; it is the record of which selves have bound which, written by the commitment devices themselves. Operationally, this requires no psychological classification: selves become legible through observable differences among commitments, instructions, overrides, constraints and realized actions. As Amazon founder Jeff Bezos relatedly observed, "Good intentions don't work, mechanisms do."

This raises the constitutional question our architecture can't evade or avoid: if no 'sovereign self' stands above the portfolio, who has authority to charter the rest? Computational egonomics does not answer by declaring one self 'superior;' it answers procedurally and algorithmically. The charter's authority derives from the conditions under which it was authored: reflective deliberation, explicit commitment, temporal distance from immediate temptation, and willingness to be evaluated against future outcomes. The architecture privileges procedures, not personalities. It assumes no sovereign self; it makes competing claims of authority visible, revisable, and contestable. The enterprise Ulysses is agentically and algorithmically lashed to his chosen digital mast.

III. What the Quantified Self Got Right, and Missed

The Quantified Self movement of the 2010s was a serious idea served by trivial instruments. Its Druckerian premise — that measurement enables management, that what is tracked can be improved — was correct and remains so. Its implementation was, no pun intended, bloodless: Step counts, sleep cycles, glucose curves, screen time, and heart-rate variability produced a torrent of data and a trickle of transformation. People came to know measurably more about themselves but few 'became' measurably more.

Read through Schelling's egonomic frame, the failure is largely diagnostic. The original Quantified Self tracked the *outputs* of the game without ever instrumenting the *players*. Ten thousand steps taken by an exhausted, avoidant self are hardly equivalent to ten thousand taken by a disciplined, strategic self; they are moves by different agents in a strategic situation a dashboard cannot see. The movement had no theory of which 'self' was being quantified, because it had no theory of the game in which the quantification occurred.

Even worse, it had no theory of purpose. Tracking assumed that more data plus a dashboard equals measurable and/or motivated improvement. It does not. Without an answer to *who am I trying to become?* — the Who Do You Want Your Customer To Become? question turned inward — measurement collapses into surveillance of the self by the self, producing guilt and stasis in roughly equal measure. The Quantified Self stalled for want of three things: a theory of selves as strategic players, a theory of becoming as a directional commitment, and an instrument that produces reflection rather than mere record. AI now plausibly, computationally and economically supplies all three.

IV. From Quantified Self to Quantified Selves

The pivot is conceptual before technical. A person is not a single self with a single utility function; a person is a portfolio of selves, each with characteristic strengths, blind spots, time horizons, and risk appetites. This is a coherent extension of human capital theory under egonomic conditions. Gary Becker treated human capital as a single stock to be invested in over time. The selves framing treats this as a *portfolio* of stocks — each with its own return profile, correlation structure, and depreciation rate — and adds the egonomic insight that the holdings are themselves strategic actors with private information and competing interests. Paul Romer located economic compounding in idea-generation; the selves framing locates it in the disciplined cultivation of the selves that generate the ideas, mediated by mechanisms that make cooperation among them incentive-compatible. Thank you, Leonid Hurwicz.

But the instrumentation problem has always been the binding constraint and it's one AI dissolves. Dashboards can stop asking *how am I doing?* and start asking *which self is doing the doing, and is it the self the principal authorized?* A portfolio-allocation problem emerges: how many hours did the 'analytical self' draw this week, the 'facilitative self,' the 'resilient self'? Which is under-invested? Which is over-deployed against the strategy of becoming? Where is the arbitrage — the self that, given more cultivation, would generate disproportionate returns? These are both capital-allocation and human capital cultivation questions and they have been unaskable – let alone answerable - for want of an instrument. Thank you, AI.

V. The Alliance Is Always Assigned

This is where current agentic AI deployment paths appear more dangerous. The dominant market incentive is frictionless delegation: the agent handles the task, the recommendation appears 'automatically,' the workflow executes invisibly, the preference is inferred before deliberation occurs. This feels 'efficient' because it is efficient — and measurably so. But efficiency and agency are not even remotely identical. Paths of least resistance are not highways to maximum advantage.

Similarly, the three AI capability classes are not equivalent in their effect on agency, and the egonomic frame makes those distinctions cleaner than their standard taxonomies. *Predictive* AI is the observer: it converts the historical record into a forecast of which self is likely to play under which conditions. But prediction is not destiny; it is the prior against which deliberate departure becomes legible. *Generative* AI is the proposer: it surfaces the moves a given self would most generatively play and the situations where that self is the wrong player. *Agentic* AI is the participant — and the high-leverage variable — because every delegated action is a move on behalf of some self, and those moves compound into the cultivation - or deprivation - of the selves they serve.

The critical distinction is not the legacy 'human versus machine' dialectic and trope - and not even 'automation' versus 'augmentation' as tidy opposition. The reality: *every consequential delegation assigns an alliance*. The 'agentic vs agency' question is whether the alliance is assigned by default or by charter and/or constitution. Delegation assigns it by default — to whichever 'self' happened to be present at the prompt. Instrumentation assigns it by charter — to the self the principal has reflectively and/or intentionally chosen to develop. That's the magic of commitment mechanisms.

For the serious AI-empowered enterprise, this isn't philosophy – it's measurement error with a balance sheet. The productivity case and the agency case are not the same case and – crucially - the second is the one that compounds. A further constraint sharpens this: Agentic AI scales the production of cognitive demands far faster than humans scale their governance of attention. Launching agents is now cheap; reviewing, reconciling, authorizing, and learning from their output is emphatically not. Google's Addy Osmani calls it 'the orchestration tax:' agents parallelize execution, but judgment remains scarce, serial, and bounded. The danger is not Terminator's Skynet; it is developmental drift. And the governance question is not whether a human is 'in-the loop' or 'on the loop' but whether their judgment is fit for purpose — i.e., which self holds the evaluative lock. Practically and pragmatically speaking, an 'exhausted self' approving twenty agent outputs at a midnight deadline is not the 'strategic self' reviewing three at the dawn of a new day. The CFO of the opening — clearing the queue before the earnings call — is the same person, but not the same governor. Ironically and economically, this is 'obvious.'

VI. Instruments That Measure What They Commit

The technical core of a Quantified Selves architecture is an instrument we have elsewhere called 'Reflective Escrow.' When a task is delegated to an agent, the system completes the work — but holds the output in escrow. Release requires the individual or team to submit a revised prompt. Not a reflection. Not a journal entry. Not a quiz. A revised prompt that reflects a changed understanding of the task — a refined assumption, a sharper success criterion, a clearer constraint, a rethought trade-off. Not more words: a change to one of those four, not a rewording of the same one. Talk is cheap; a revision that costs nothing to fake was never a commitment at all. Schelling's classical devices constrained behavior but generated no analyzable record of that kind; the Reflective Escrow closes that gap — the instrument of commitment and the instrument of measurement are the same thing.

It's the move that converts Schelling's analog paradigm into computational form. Escrow is therefore a selective gate, not a universal tax: most actions execute, some are logged, fewer surface divergence, and only consequential conflicts demand reflection or adjudication.

This pattern generalizes beyond prompts: the proposed action and the action actually taken, the committed direction and the realized trajectory, the recommendation accepted and the recommendation declined. These pairs are the raw material of selves-measurement. The same instrument that commits and measures also individuates: the gaps it records are what make the selves countable in the first place. It does not find the players and then score the game; *scoring the game is how the players come to exist.*

Three stored objects — committed, proposed, realized — yield three gaps that animate the architecture. To be clear, these gaps are not direct measures of agency. They are telemetry: observable signals whose interpretation depends on context, capability, domain, and circumstance. Like heart rate in medicine, they are evidence rather than diagnosis. Read together, and validated against outcomes, they can illuminate when one self is binding, or failing to bind, another:

• **Revision distance and deference rate capture complementary signals:** how far someone moved from the AI's first response, and how often they move at all.A CFO who revises one of seventeen AI recommendations is deferring roughly 94% of the time — visible the instant one looks but silent on whether that's an expert trusting a well-tuned agent or a self that's quietly clocked out.
• **Commitment drift** is the gap between the plan on record and what actually happened — the number that catches the midnight self signing off on something the morning self never agreed to deliver.
• **Net drift** compares the AI's first suggestion straight to the final outcome, skipping everything in between. It catches those cases where looks unchanged' doesn't mean 'nothing happened: a plan gets heavily revised, then drifts back during execution, landing right where the AI started — same final answer, real deliberation, real slippage, invisible unless the other two numbers are also checked.

None of this is inherently new instrumentation; it respects and reflects earlier LLM metrics like 'Return on Iterative Prompting' and/or Anthropic's recent 'Reflect' offer. — the same delta that scores a prompt, scoring a self. Each gap is where a self 'binds' — or fails to bind — another. This is the enforcement of Section II made operational — individuation shown, not asserted.

These measures' purpose is not eliminating or minimizing divergence. Divergence is often where learning begins. The developmental question is whether the architecture helps distinguish between productive revision and unreflective deviation. That's not subtle. A developmental ledger records where commitments held, where they failed, and where experimentation, adaptation, or self-revision produced better judgment. The objective is not fidelity to a past self but better informed authorship of a future one.

For example, an agent managing vendor renewals is chartered to flag any renewal over 5% before accepting it. At 11pm, the same CFO who set that rule tells it to push one through anyway. The agent doesn't comply or refuse — it logs the conflict: threshold exceeded, override requested. That gap is the dependent variable for agency; closing it, where closing serves 'becoming,' is the work.

This inverts oversight implications: Typically, oversight means humans watching machines. Here, the agent reports its own divergence: when an action would serve a 'self' other than the chartered one, the agent writes that gap to the ledger itself and surfaces it for adjudication. This is what converts the promise to 'surface, not hide' from an aspiration into a mechanism. An agent that volunteers the moment it strayed from the charter is doing the one thing a default-aligned agent never does — making the alliance it just formed visible to the principal empowered to revise it – if they so choose. That choice is the essence and substance of human agency.

VII. Governance as Mechanism Design

A measurement architecture without a governance architecture is surveillance. The governance problem is, in Leonid Hurwicz's terms, a mechanism-design problem applied to the intra-personal case: given that a person's selves hold private information — each sees what the others suppress — and pursue competing interests, how does one design the agent-mediated interaction so that the dominant strategy across selves yields *developmental* rather than extractive outcomes?

This paper borrows game theory's vocabulary — players, payoffs, equilibria — without doing game theory's math. That formal modeling is worth doing, domain by domain, but comes later. The claim here comes first: agentic systems turn the old, private tug-of-war between your selves into actions with real consequences — money spent, contracts signed, renewals overridden. Once the conflict produces outcomes instead of just feelings, it can only be governed by an architecture that writes the conflict down, lets it be contested, and changes behavior because of it. The log has to teach both the agent and the human something. Recording contradictions isn't enough.

Posed that way, governance elements are not checklists; they are what the mechanism requires to be incentive-compatible. A *charter* solves the objective problem: without it, the agent inherits whatever objective the platform optimizes — rarely the principal's — and the alliances are assigned by default. The charter makes purpose computable: it specifies delegated authority, constraints, escalation and reconsideration conditions, override rights, and the purpose against which actions are judged. An *Institutional Surprise Rate* analog solves the detection problem: a measured capacity to notice when realized behavior departs from chartered direction — in egonomic terms, when unauthorized selves have been playing.

'Surprise' that is investigated compounds; surprise merely registered is consumed. An *adversarial agent* solves the motivated-reasoning problem, which the paradigm predicts: the self in control of the interpretation has every incentive to interpret in its own favor, so a function must be charged to argue against the dashboard's own conclusions. A commitment ledger solves a precommitment problem: it holds a later, tempted self to what an earlier, clearer self decided."

This returns the argument to principal-agent theory, but with 'moral hazard' relocated. Classical theory assumes a unitary principal and fears a self-interested agent. Computational egonomics inverts the worry: the agent may be perfectly faithful and still serve the wrong principal, because the principal is plural and the agent allies with whichever self holds the prompt. The charter resolves what the classical problem never had to — not the agent's loyalty, but the principal's identity. Designing and discussing 'agentic AI' decoupled from this behavioral economic reality is foolish and, yes, counterproductive.

The principle here is the same as in the enterprise: KPIs focus attention; KPAIs learn. A KPI for personal productivity counts tasks completed. A KPAI learns which 'self' completed them, under what conditions, at what cost to which other self, and what the predictive signature of a high-quality deployment looks like. KPAI is to KPI what Waze is to a static map: it adapts, and it learns to adapt.

The enterprise boundary matters. The 'selvesware' we describe is a developmental architecture, not a managerial surveillance architecture. Its purpose is to give individuals and teams better visibility into how human agency is exercised, delegated, and cultivated — not to give organizations finer-grained behavioral control over employees. An employee's charter is theirs to write; the enterprise's role is limited to checking that one exists, that it was actually agreed to, and that departures from it are visible — never to dictating what the charter says. That's the whole difference between oversight and control. Enterprise deployment therefore requires explicit rules for consent, data ownership, auditability, interpretation rights, and permissible use. Without those boundaries, this architecture degenerates from self-authorship into Taylorism with intelligent telemetry.

VIII. The Wager, the Divergence, and the Frozen Self

This framing yields a concrete and — this is the discipline the Quantified Self never imposed on itself — *measurable* prediction. Two populations will almost certainly develop and diverge over the next decade.

The first will use agents to bypass deliberation. Decisions will continue to be made but increasingly by defaults 'intelligently' set elsewhere; preferences inferred rather than chosen, selves never deliberately cultivated. The signature is readily observable, and now nameable: revision distance collapses toward zero, deference rate climbs toward one, commitment drift widens unremarked; the evaluative lock migrates to fatigued and impulsive selves; the gap is never read. This population will report rising satisfaction and rising productivity. Both will be real — and that is precisely why neither is the variable in question. The second population will instrument and contest that divergence; it narrows where narrowing serves becoming. The difference between the two will not be the models they use; it will be the divergence signatures, and those signatures are measurable.

The strongest and most obvious objection to this deserves discussion: People did not want to be quantified singly and will likely desire quantification of 'multiple selves' even less. Relentless deliberation over which self to deploy could become its own agency-erosion, crowding out the unselfconscious actions where much of life's value lives. Agents that simply *do things* may be more mercy than menace. They may be market winners.

This objection has real and pragmatic force and the reasonable response feels more narrow than triumphant. This architecture is not for every market, domain or domain expertise. For the consequential ones — strategic decisions, professional judgment, long-horizon commitments, the cultivation of capability — the alternative to instrumented selves is not zen-like flow; it is the unmeasured surrender of those decisions to defaults optimized against objectives that are not the principal's own. The effective and affective analogy is sports analytics: baseball without WAR did not feel impoverished to its participants. The measurement was built because the gap between perceived and actual contribution compounded where the stakes were highest. It's called 'Moneyball' for good reason. Most domains need no such instrument. The ones that do get it from nowhere else. That's the disruption.

That said, don't underestimate how talented individuals and teams may 'double down' on agentic innovation and instrumentation to win the games they're playing. Winners won't just deploy agents faster — they'll read their own revision distances and commitment drifts as closely as their P&Ls.

This wager has a precedent the market respects. Warren Buffett's edge, for example, was **never** foresight; it was temperament rendered as rules — the refusal to let the in-the-moment self trade. Bridgewater's Ray Dalio wrote his down — Principles, literally. Investment discipline binds the future self to protect returns already defined.

'Selves discipline' binds the future self to *define who does the returning*. One firewall keeps the analogy honest: capital converts to a single number and selves do not — there is no exchange rate between the creative self and the recuperative one. That missing denominator is not a defect in the analogy; it is the firewall. Investment discipline answers *how* to bind; it can't answer *which self to become*. That question has no price, which is precisely why it must be chosen rather than optimized.

This provocatively recasts 'binding commitment.' 'Alignment' optimizes an agent against what may best be described as a snapshot of the principal's stated preferences. But a perfectly aligned agent that systematically replaces the principal's deliberation has not aligned with the principal; it has aligned too simply with a 'still' picture and thus corrodes the original. And now we can say precisely why: there is no fixed source to align to because the self is constituted by an ongoing act of binding and commitment – whether operational or aspirational. An agent that freezes the snapshot doesn't just misread the principal; it *arrests the authorship*, effectively and affectively halting the person from becoming anyone they had not already committed to being.

By contrast, agency optimizes the opposite target: the capacity to revise the snapshot. A modestly capable agent under a charter and a commitment ledger aligns with something larger than stated preference — the principal's capacity, and willingness, to revise it. This working paper asserts research and policy communities should stop seeing 'alignment' as their dominant binding question and start treating the 'augmentation of agency' as their North Star.

So, the 'Who Do You Want Your Customer To Become?' question, turned all the way inward, was never *which self should act?* but *which self are we willing to bind ourselves into being?* 'We' are the commitments that hold.

This scientific/economic/technical historical frame clarifies the stakes for anyone who's lived through prior 'leverage revolutions.' Industrial systems amplified physical leverage. Information systems amplified informational leverage. Agentic systems are the first to amplify — or erode — *developmental* leverage. That is historically, ontologically, and teleologically new. The defining competition in agentic AI will therefore not be whose models or model ensembles are smartest, fastest, or cheapest but whose architectures most effectively compound judgment, intentionality, adaptability, and self-authored capability over time.

IX. What This Demands

The implications here are not modest: they challenge agent developers to treat the principal's agency as a first-class design target rather than an afterthought. They explicitly recognize that every 'default' shipped is effectively a constitutional assignment made on the user's behalf. They also fundamentally

challenge enterprises deploying agents at scale, who must accept that productivity and agency are not the same case — and that agency is the one that compounds.

Just as provocatively, these implications fall on individuals who must decide whether to refuse the default treatment of themselves as a single dominant utility function. Ideally, they will be able to access instrumentation that offers actionable insights into the 'selves' they actually are. This may prove central to their professional development of competences, capabilities and judgment in the AI enterprise. The competitive question may ultimately be less "How many agents can we cost-effectively deploy?" than **"**How many agents can we deploy without destroying the quality of the principal they are supposedly empowering?"

The operative demand is sharper and ruthlessly reframes the central artifact of the whole architecture. Stop writing charters that delegate tasks; craft charters that author 'selves.' The charter is no longer a permission slip for the agent — it should be the principal's statement of which 'self' they are committing themselves into being and becoming. The three gaps represent the preliminary dashboard of whether the authorship is holding. Perhaps that dashboard will become their GPS.

To be sure, the agents – ever-smarter and more capable - will arrive regardless, thanks to better design or Recursive Self Improvement. Whether they enlarge or erode the humans they ostensibly serve is the singular design decision of the decade. There is no neutral default. The choice is not between human and machine: This choice offers a future in which agents tacitly answer the question of who the human is becoming and a future in which the human, instrumented at last, can actually answer.

Sidebar 1 — Agent Charters

'Agent charters' as standing instructions for AI agents:

- What they should **do**
- What they should **question**
- When they should explicitly seek permission before acting.

These examples show how people can delegate work **without unintentionally delegating away their judgment**.

The CFO Decision Agent Charter

Purpose. You exist to increase the quality and revisability of my financial judgment, not merely the speed with which I clear decisions.

You may prepare analyses, compare alternatives, challenge assumptions and execute routine actions already covered by established policy. You may not treat my latest instruction as automatically superior to commitments, thresholds or decision criteria I deliberately established earlier.

When my requested action conflicts with an existing commitment, exceeds an agreed threshold, materially changes capital at risk, or contradicts assumptions on which the original decision depended, surface the conflict before silently resolving it.

Show me what changed: the prior commitment, your proposed action, the relevant evidence and the economic consequence of proceeding or reconsidering.

For material decisions, do not ask me merely to "confirm." Ask me to identify what has changed in at least one of four places: assumption, success criterion, constraint or trade-off. If nothing has changed, make that visible.

My goal is not perfect consistency with my past decisions. I authorize myself to learn, reconsider and reverse them. Your job is to make the difference between deliberate revision and unexamined drift legible.

Record material divergences and their eventual outcomes so that we can learn which kinds of reconsideration improve decisions over time.

The test of this charter is not whether you make fewer decisions for me. It is whether your delegation helps me become better at knowing which decisions deserve my judgment.