

We Now Have an “Agent-Principal” Problem

By Rahul Jindal

We call the thing that does the work an agent, and then we reassure ourselves that whatever it takes over, the agency stays with us. It's a strange thing to believe once you look at it. An agent, by the ordinary meaning of the word, is something that has agency. We picked a name for the machine that already holds the very thing we're claiming to keep. A lot of the confident writing about where humans fit in the AI era is built on exactly that mistake, and it doesn't hold up.

So it helps to borrow an older idea, the one economists have used for a hundred years: the Principal and the agent. Not human against machine, but a Principal who wants something done and an agent who does it. Look at the relationship that way and you notice something AI has quietly changed about a very old problem.

For a century the Principal-agent problem has been about keeping the agent honest. The whole point of the field was one worry: how do you get someone working on your behalf, someone with their own interests, their own laziness, information you can't see, to actually do what you need instead of what suits them? All the attention pointed at the agent. AI takes that particular worry off the table. Not because the agent is flawless. It will hallucinate, call a job done when it isn't, tell you what you want to hear. But none of that is self-interested. The agent has no agenda. It isn't covering for itself or angling for a promotion, and its mistakes are honest ones, out in the open, with no hidden party working against you. So the old question hasn't gone away. It has changed direction. We used to ask whether the agent could be trusted with the Principal's work. **The question now is whether the Principal is worthy of the agents.**

“Worthy” is a word that ends up on posters, so pin it down. **A worthy Principal is one who makes the fleet worth more than it would be on its own.** Not more supervision. More value: your real contribution to the thing that pays for all of you, human and machine alike. That's the whole of it.

A fleet needs fewer Principals than it needed people to do the work it just absorbed. You can hear that as a threat, and plenty will. But the seat isn't scarce like a lottery ticket. It's scarce like skill. It goes to whoever builds what it takes to hold it, which means you can earn your way into it.

The old craft of managing an agent came down to three things: watching it, motivating it, and knowing more than it did. AI reverses all three, and each reversal asks something new of you.

Watching

You used to manage by watching. Reviews, one-on-ones, dashboards, trust but verify. Most of that was really about distrust: catching the agent slacking or hiding something. There's none of that to catch now. The agent still makes mistakes, but they're in the open, and checking for them is mostly mechanical. What no dashboard can show you is whether the work was worth doing at all, and for that there's no dashboard above you. You're at the top. For most of a career, someone above you catches your worst calls before they land; here, no one does. What the seat asks of you instead is a feel for what actually matters, and the nerve to commit to it long before anyone can tell you whether you were right.

Motivating

You used to get an agent moving by lining its interests up with yours: pay, promotion, culture, a little pressure. None of that touches a system that wants nothing. You can't inspire it, can't guilt it, can't win it over on a bad Monday. The whole management reflex counts for nothing here, and what's left is saying exactly what you mean. That sounds easy and isn't. Most of the goals we hand each other are vague, and they only work because another person quietly fills in what we left out. The fleet won't. It gives you precisely what you said, at scale, which is usually how you find out what you actually said. Very few people who spent years learning to motivate have ever had to be that exact.

Knowing

You used to work partly in the dark, because the agent always knew things you didn't; the person doing the job knew whether they'd actually tried. Now it runs the other way. The

agent can know more than any of us and keeps none of it back. The only blind spot left in the room is yours.

I've watched this go wrong. A capable, experienced operator hands a hard piece of analysis to the fleet, gets back something fast and clean, and ships it. A month later they're in a meeting being asked why, and they can't say, because they never held the reasoning, only the answer. **Understanding used to arrive as a byproduct of doing the work.**

This time the fleet did the work and the understanding never showed up. They handed off the task and, without meaning to, handed off the knowing that used to come with it. In that meeting they still have the title. They've stopped doing the job it's there for.

That's the trap, in one motion. Doing the work was how we came to understand the work. The fleet does the work now. Let it take the understanding too and you slowly lose the depth that made you worth listening to, and the seat goes to whoever didn't. Staying the person who understands the goal better than anyone, even after you've stopped doing the thing that used to teach you, is the hard part of this whole era.

“The question used to be whether the agent could be trusted with the Principal's work. Now it is whether the Principal is worthy of the agents.”

The seat

Put the three together and “worthy” finally has a shape:

1. You decide what's worth doing.
2. You say what you mean clearly enough that a fast, literal machine can't run off in the wrong direction with it.
3. You stay close enough to the problem to lead people and systems that know more than you do.

Manage all three and the fleet really is worth more with you than without you, which is the only honest way to be worth your seat.

You'll keep hearing that agents still need us. They do, and it shouldn't reassure you as much as it's meant to. A fleet needs a Principal the way a company needs the right person in its one seat that matters, which is to say not many, and not you by default. The agent was the hard problem for fifty years. That one's solved. What's left open is you.

Sources: the principal-agent problem in economics traces to Stephen Ross, “The Economic Theory of Agency: The Principal's Problem” (American Economic Review, 1973), and Michael Jensen and William Meckling, “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure” (Journal of Financial Economics, 1976). The reversal argued here, that AI moves the scrutiny from the agent to the Principal, is my own.

Take it with you

Email this as a LinkedIn pack

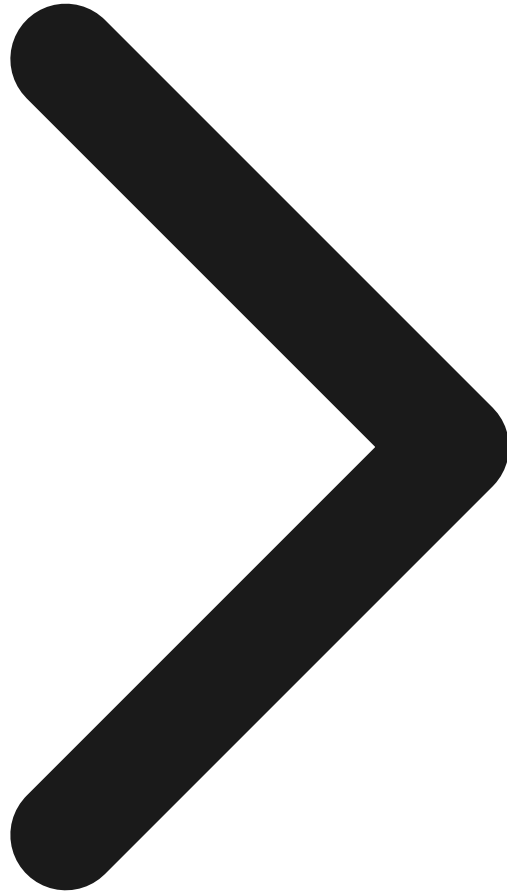
Get a feed-ready LinkedIn post (under the 3,000-character cap), a long-form LinkedIn article version, the hero image, and an editable document version of the full essay, delivered to your inbox. Ready to post.

Email this

Related Insights

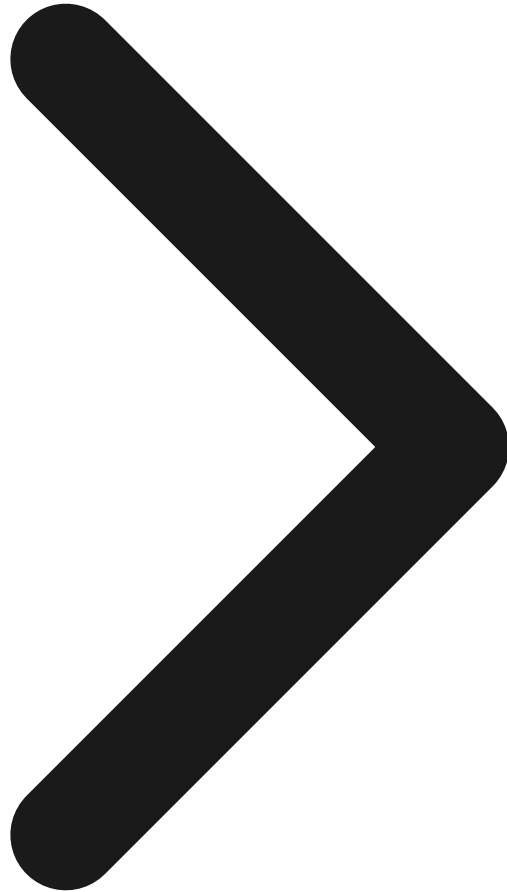
Playbook9 min

The Agentification Playbook



Strategy 5 min

AI Transformation Is Not a Technology Problem



Framework 6 min

What Is Organizational Metabolism?

