ARTIFICIAL INTELLIGENCE

Why AI Agents Need an Identity, Not Just Instructions

By Björn Ingenleuf and Vladimir Lukic

ARTICLE JUNE 17, 2026 12 MIN READ

Two retailers deploy AI agents built on the same underlying model, configured to the same standards of helpfulness, accuracy, and compliance. Two customers reach out with similar requests: returning a product well outside the standard window due to a personal circumstance.

In each case, the agent's initial response is technically sound. The AI presents the company's policy clearly, essentially telling the customer what they already know. But to the customers, the rest of the interaction could not be more different.

At the first retailer, the agent has been built to reflect a culture of process and efficiency. It applies the returns policy precisely and closes the interaction. The agent at the other retailer has different cultural roots, reflecting the importance of top-flight customer service and brand loyalty. It acknowledges the customer's strong ties to the company and finds a way to make an exception for the return.

> Over time, across millions of interactions, AI agents express a company's purpose, values, and culture.

Nothing is wrong with either answer. But the difference in customer perception is huge. Over time, across millions of interactions, that difference compounds. One answer builds consumer relationships; the other delivers consistency, but at a cost.

The difference comes down to culture—and whether organizations deliberately embed it into the AI systems acting on their behalf. As AI agents become more autonomous, companies will need to define far more clearly how their purpose, values, and culture are expressed in every interaction they enable.

Shifting from Execution to Representation

When systems act on behalf of organizations, they don't just deliver outputs or measurable KPIs. Their actions are the sum result of choices the company made in implementing AI: how the agent responds to customers, what it prioritizes, when it makes exceptions, and how it handles ambiguity. These are not purely technical questions. They are cultural ones.

Yet most agents today are still built to optimize for performance: speed, accuracy, efficiency, compliance. They are aligned to metrics, not to meaning. The result is a growing disconnect. Organizations invest heavily in defining their purpose, values, and brand. But the systems acting on their behalf often remain at best neutral to these principles—or, at worst, at odds with them.

Over time, this creates a subtle but powerful form of drift: customer interactions become consistent but not distinctive; decisions become efficient but not principled; experiences become scalable but not meaningful. The system works. But it no longer reflects who the organization is.

> When agents are deployed without cultural grounding, organizations scale sameness.

Consider a financial services firm that has spent years building a culture of empathetic, human-centered advice. Its advisors are trained to slow down, listen, and treat difficult conversations—such as a client facing debt or a family navigating a bereavement—with care. Then the firm deploys an AI agent to handle first-line customer interactions. The agent is accurate, fast, and compliant. But when a customer reaches out in financial distress, it responds with a technically correct summary of options and a link to the relevant policy page. The expertise transferred. The humanity did not.

When agents are deployed without cultural grounding, organizations scale sameness. In a world where more and more interactions are mediated by AI, this creates a new competitive risk: the erosion of identity. If every organization relies on agents that behave the same way, differentiation disappears—not at the brand level, but at the critical crossroads where customers decide whether to continue to engage with the company.

Coding Organizational Identity

For decades, culture inside organizations has been expressed through stories, institutional knowledge, leadership styles, and shared norms. People absorbed it through experience: how decisions were made, what behaviors were rewarded, when exceptions were allowed, and what mattered under pressure.

Over time, those internal behaviors shape organizational identity: how a company is consistently experienced by customers, employees, partners, and the market. Culture is what produces that consistency.

> If agents are going to represent companies, then the company's identity has to be translated into a language that systems are comfortable with.

But nonhuman actors require a more explicit and programmed set of instructions. If agents are going to represent companies, then the company's identity has to be translated into a language that systems are comfortable with.

Organizations already understand this principle in human terms. Companies do not hire solely for competence. Instead, they also evaluate whether people will fit into the culture comfortably: Do they have the right personality and skills to handle customers, collaborate with colleagues, and exercise judgment in ways that align with the organization's vision, values, public-facing presence, and brand? Yet, AI systems are often implemented almost entirely around capability and performance, with far less attention paid to whether they reinforce the organization's identity in practice.

If organizations want AI systems to reflect their identity consistently, then culture can no longer remain implicit. It has to be embedded directly into how systems operate, make decisions, and interact with people. In practice, that means turning organizational identity into operational logic:

- **Corporate purpose** becomes the foundation from which the system operates—an organization's timeless *why*, not a decision rule or task directive. A health care company focused on helping people live healthier lives, for example, would design agent interactions so that even routine administrative tasks reflect empathy, clarity, and patient well-being, not just procedural accuracy.
- **Corporate values** become how agents make decisions and manage tradeoffs. They shape what the system prioritizes when goals conflict, rules are ambiguous, or exceptions arise. For a company that emphasizes fairness, that could mean applying consistent reasoning across all customers, avoiding exceptions for high-value accounts that would not be made for others.
- **Corporate behaviors** become interaction protocols: how agents communicate, escalate, collaborate, and respond emotionally. They determine not just what the system does, but how it does it. A company that values empathy might encode behaviors that require the agent to acknowledge a customer's emotional state before attempting to solve their problem.

To see what this looks like concretely, consider a premium hospitality brand that believes every guest should feel genuinely welcomed in a personal way, not just efficiently served. In human

terms, this is manifested in the way staff remembers a returning guest's preferences or notices when a guest seems stressed and helps them out proactively.

In an AI design layer, these interactions must be captured in a set of explicit agent behaviors: examine the full context of the hospitality activity and the guest's needs before addressing the problem; never close an interaction without confirming that the guest is satisfied, not just answered; and escalate to a human whenever emotional signals are present.

This is not an exercise in adding more rules or constraints. It is instead about determining how systems interpret situations, balance tradeoffs, and behave when judgment is required. Every system optimizes for something.

How Does Culture Become Computable?

This shift to a more "cultured" AI introduces a new kind of work. Call it translational: turning meaning into mechanism, belief into behavior, and identity into systems.

Values are not rules. "Act with integrity" or "put the customer first" are not instructions a system can execute. They are beliefs and principles formed through thousands of human decisions, shaped by context, exception, and precedent. The challenge of making culture computable is, at its core, the challenge of making tacit knowledge explicit.

Several technical approaches are beginning to emerge to tackle this issue.

The first is **prompt and system architecture**. This involves structuring the inputs that govern agent behavior to include not just task instructions but explicit direction about the organization's identity and how it wants the agent to behave: what the agent should optimize for, what it should refuse, and how it should reason through ambiguity and conflict. For a company whose identity is built on trust, that last instruction might mean "When in doubt, say so. Never project false confidence just to close an interaction cleanly." This is the fastest approach to deploy because organizations can more easily see how the AI is being guided and adjust its behavior over time, making it a practical starting point for many organizations.

A second possibility, and one that provides deeper encoding, is **fine-tuning and value-weighted training**. This method adjusts models using selected examples that show how the organization's values are applied in real situations. Not abstract principles, but behavioral patterns. A luxury retailer, for example, might use hundreds of scenarios to demonstrate how its best advisors handled difficult situations—what they decided as well as how they reasoned, what tone they used, and which exceptions they were willing to make.

Moreover, training examples do not have to come from designed scenarios. Call center transcripts, executive interviews, internal communications, and other records of everyday work can often provide a richer picture of how the company actually operates. But not everything an organization does should be reinforced. The goal is to teach the system the behaviors and judgments that best reflect who the organization wants to be.

A third option, still experimental, focuses on **rule-based governance layers**. This approach incorporates automated review systems that evaluate whether agent responses reflect the organization's values before the responses are delivered. In effect, they function as a real-time cultural audit and guardrail. A health care organization, for instance, might identify responses where overly efficient or impersonal language appears in emotionally sensitive situations, helping prevent cultural drift before it reaches a patient or family member.

In practice, implementing these approaches is often more nettlesome than the technology itself. To do so, organizations first determine the distinctive aspects of their culture and identity they want to preserve, how those qualities can be expressed in operational terms, and which examples best represent them in practice.

Most organizations have never had to make their culture explicit enough to be operationalized. It has always lived in people, in relationships, in the unwritten rules of how things get done. Translating that into something a system can learn from requires a different kind of work: part strategic, part cultural, part curatorial, and above all, deeply human. The challenge is twofold: building the system and ensuring the system accurately reflects the organization behind it.

Beyond Technology

Technology alone will not solve this problem. Organizations first have to do the harder work of defining what their values actually mean in practice: under pressure, at the edges, and in situations where tradeoffs are unavoidable.

That raises broader organizational questions. How should companies measure whether agent behavior consistently reflects the organization's values over time? What metrics should be used to identify cultural drift or reinforce cultural coherence across thousands or millions of interactions? And who inside the company manages AI agents: brand leaders, technology teams, operations, sales, strategy—or some combination of all five? Today, in most organizations, the responsibility remains unclear.

Culture that cannot be described with enough specificity to teach to a human will not survive translation into a system. Organizations that solve this problem in a disciplined and rigorous way will gain an advantage that extends beyond operational performance. Their systems will not

simply function effectively. They will behave in ways that consistently reflect the organization itself.

Organizations that succeed in the next phase of AI adoption will not be those that simply deploy the most advanced systems. They will be those whose systems consistently reflect the organization's identity, values, and way of operating. In a world where agents increasingly act on behalf of companies, culture can no longer remain implicit or confined to human behavior alone. It must be translated into the systems that increasingly act in our place.

The defining question is no longer just what an organization stands for. It is whether the systems acting on its behalf behave accordingly.

Authors

Björn Ingenleuf

Brighthouse Group Creative Director
Berlin

Vladimir Lukic

Managing Director & Senior Partner; Global Leader, Tech and Digital Advantage
Boston

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change — bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.