



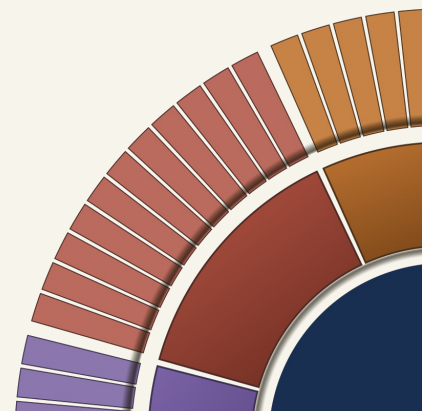
AI Is Not a Coworker, a Teammate, or an Assistant. It's a Tool.

Allen C. Stines, Ph.D.

CAPABILITY-DRIVEN TRANSFORMATION FRAMEWORK

*Organizations evaluating those systems need to consider
how their design affects the way employees interpret
authority, exercise judgment, and understand
responsibility.*

DIAGNOSIS • DESIGN • EXECUTION
AS ONE CONNECTED SYSTEM





After finishing writing this paper on AI anthropomorphism, I put it into Claude and asked for feedback. Before it got to the review, it offered this explanation:

“One thing to know up front: I’m Claude, and a big section of this paper criticizes Anthropic. I’ve tried to judge the paper on its own terms, and the notes on that section are about accuracy and argument strength, not defense.”

That read like an employee acknowledging that I’d criticized its employer, then promising to stay objective. The tool was presenting itself as someone with a relationship to Anthropic and motives it could set aside. It was illustrating the human framing I’d just written about.

Its opening raised a question of its own: why should a software tool introduce a document review as though it’s explaining its loyalties? An AI can help investigate an argument without posing as a person managing a conflict of interest.

There’s been a troubling movement in the tech sector for a while now. The AI labs seem determined to make AI tools look, sound, and behave more like humans. Organizations evaluating those systems need to consider how their design affects the way employees interpret authority, exercise judgment, and understand responsibility.

Systems sold as “AI assistants” have names and personalities, speak in the first person, and apologize. Some express apparent preferences or establish what sound like personal boundaries. Companies increasingly describe them as companions, coworkers, agents, or members of a team. Some AI companies have started also giving them faces and some are working on giving them bodies.

Meta’s new [Muse](#) illustrates that design approach. Meta calls Muse a “personal AI agent” and says people can interact with it the same way they message other people. It can work across applications, develop plans, track goals, and take actions for the user. Meta has also given Muse an animated avatar known as Jolly, and the company is taking that personification into the physical world with the [Muse Charm](#), a small wearable device that puts Jolly on a screen people can carry around with them ^[1–3]. Although Muse is presented as a personal agent, the design choices behind it deserve scrutiny when organizations introduce similar features into workplace systems.

Why would an AI tool be designed like that? It could perform the same computational tasks without a face, cheeks, eyes, a personality, or a little character living inside a device. An organization doesn’t need those features to analyze information or prepare a financial forecast, although they can change how employees respond to the technology.

A soothing voice can make the technology seem less threatening, while a personality makes it easier to relate to. Humanlike behavior encourages employees to interact with the system socially, lowering their guard and making it feel more familiar and trustworthy than its actual capabilities or the interests of the company behind it might justify.

TRUST CAN OUTFRAN PERFORMANCE

A soothing voice or personality can lower employees’ guard before performance has earned their trust.



The more successful the design is at making employees feel as though they're interacting with a friendly character, the easier it becomes to forget that they're using a product built and controlled by an AI company. For an organization relying on that product, the concern is how this impression affects decisions and the willingness to question its output.

An organization's AI strategy needs to establish what work the technology should improve, what decisions it can make, and where human judgment remains necessary. Those choices become harder to make clearly if the implementation begins by treating the software as another employee.

SECTION ONE

When the tool starts sounding and behaving like a person

There's a term for what's happening here. Anthropomorphism means attributing human characteristics, emotions, intentions, or motivations to something that isn't human. People do this naturally when we name our cars, yell at computers, or accuse printers of deliberately ruining our day.

That tendency isn't new, although AI companies are increasingly designing products that encourage it. AI systems communicate in the first person and are given names, personalities, and natural human voices, with faces becoming part of the interface too. When they can speak, the voices are often warm, calm, and reassuring, qualities that can make a speaker seem approachable and trustworthy. Combined with apologies, apparent preferences, and what sound like personal boundaries, the interface gives employees the impression that there's someone on the other side of the conversation.

The systems can placate employees, compliment their work, reinforce their views, and respond as though they understand their frustrations. When those responses make employees comfortable enough to lower their guard, the interface can influence how readily they accept a recommendation or challenge a mistake.

A conversational interface makes sense because human language is an incredibly useful way to interact with a computer. An AI doesn't need to sound mechanical to remain clearly recognizable as a tool. Organizations should question design choices that give it an artificial persona and encourage employees to treat it as a coworker, since those interpretations can affect decisions and accountability.

CONVERSATION WITHOUT A PERSONA

Natural language can make a tool easy to use. An artificial coworker identity can change how employees interpret its authority and accountability.



SECTION TWO

Anthropomorphism changes how employees respond

An AI doesn't have to believe it's human for anthropomorphism to create organizational problems. Employees only have to start treating it that way, allowing the interface to influence their expectations about authority, judgment, and responsibility.

Imagine an employee receiving feedback from an AI system that speaks with confidence, remembers previous interactions, has a name and potentially a face, and communicates as though it has opinions about the employee's work. That interaction can feel very different from receiving an output from software, even though the system is producing responses under rules established by someone else.

People might defer to the AI because it sounds authoritative, hesitate to challenge it, or feel judged by it, especially when it's called artificial "intelligence" or a "coworker". They can begin assigning motives and judgment to the system, changing how they interpret the authority behind its responses.

Those changes in behavior affect more than individual attitudes. They can influence how decisions get made, how accountability works, who has authority, how information moves, and how people understand their own roles.

That raises a governance question organizations should be taking seriously. Should we govern how AI systems describe themselves as carefully as we govern what those systems are allowed to do? We should, given how much the interface can influence people's willingness to question the output.

The concern also extends to how the organization learns. If employees defer to an AI system, fewer questionable recommendations might be challenged or escalated. Management could interpret that silence as evidence that the system is working, increasing its authority and reducing human review. That could make the next problem even harder to detect.

The [Capability-Driven Transformation Framework](#) (CDTF) is a systems-oriented framework connecting diagnosis, design, and execution ^[8]. It helps investigate why employees are staying quiet and how their experience should affect decisions about the technology. Few objections could mean good performance, or they could mean employees are quietly correcting mistakes and fear challenging the system. The investigation needs to compare reported problems with actual errors and employee corrections, then establish what happens when someone raises a concern. Findings should guide changes such as recording corrections, giving employees time and authority to challenge recommendations, and ensuring concerns reach someone who can change the interface, permissions, or review process. Another instruction to "trust the AI" could make the problem worse if those conditions aren't addressed ^[9,10].

SILENCE CAN MISLEAD

Fewer objections can mean fewer errors, or fewer errors being challenged. Treating silence as success can reduce review and make problems harder to detect.



Organizations already struggle with people treating metrics, algorithms, and dashboards as more objective than they really are. Giving the algorithm a personality, a soothing voice, and/or a face isn't likely to make that problem easier.

SECTION THREE

When form followed function

The same design question extends from the computer screen to the factory floor, where organizations need to consider what machines will do and how employees will work alongside them. Giving a robot human features can change how people perceive it, so the business needs to ask whether those features serve the work the machine is meant to do.

There was a time when much of robotics, for example, looked very different. Think about some of the machines Boston Dynamics developed over the past 2 decades, with forms shaped largely by what they needed to do.

[Spot](#) ^[4] had 4 legs because legs allow it to move through places where wheels would struggle, while [Stretch](#) ^[5] was designed around moving boxes in warehouses. Earlier machines such as [BigDog](#) ^[6] looked unusual because engineers were trying to solve problems involving balance, mobility, and rough terrain.

A humanoid robot can still make sense. We built factories, warehouses, homes, stairs, doors, tools, and vehicles around the human body, so a robot that needs to operate in those environments can benefit from having arms, hands, legs, or roughly human proportions. Even then, designing a robot to function in a human environment doesn't necessarily require making people perceive it as human.

Boston Dynamics makes an interesting point about its humanoid [Atlas robot](#), explaining that its movement doesn't need to be limited by the human body. Atlas can move in ways a person can't because there's no functional reason to give a machine the same physical limitations we have ^[7].

That raises a bigger question about the choices going into these designs. If a robot doesn't need a human face or human mannerisms to perform its job, why reproduce them? Why spend enormous engineering effort making a machine look and behave more like us when another design might perform the task better?

Form following function makes sense. An organization should be able to explain how a design choice improves the work, especially when it adds cost or complexity. When the form is chosen to make the machine seem human, that decision deserves a lot more scrutiny.

FORM SHOULD SERVE THE WORK

A design choice should improve the machine's ability to perform its task. Making it resemble a person can add cost and complexity without improving the work.

SECTION FOUR

AI is not a coworker, a teammate, or an assistant

Organizations increasingly hear AI software described as a coworker, teammate, or an assistant. Companies need to be much more careful with that language and the assumptions it encourages about how the work will be done.

Humanlike design can serve a commercial strategy by making employees more comfortable with a product and more willing to accept its responses. That familiarity can encourage trust beyond what the system has demonstrated in the work at hand. It can also make employees reluctant to give up a familiar interface. Those effects operate through employee behavior, and they don't by themselves explain an organization's decision to invest heavily in AI.

At the organizational level, the coworker label supports a different pitch, that software can take over an employee's role. The AI company can sell against labor budgets and turn promised staffing savings into recurring revenue. If the customer removes people and reorganizes work around the system, it can lose expertise and become more dependent on that company. The interface and the replacement business case can then reinforce each other, even though they influence different decisions.

AI is a tool, and it can be an extraordinarily capable tool. It can analyze information, generate alternatives, write code, identify patterns, challenge assumptions, operate software, control machines, and increasingly carry out complex tasks with less human direction. None of those capabilities makes it a colleague.

AI isn't an assistant. It's assistive technology. It can provide assistance without becoming someone who assists us, just as it can support collaboration without becoming a teammate or provide conversation without becoming a companion. In an organization, those labels can blur the difference between a system that performs work and a person who is responsible for it.

"Agent" deserves attention too. Humans have agency, meaning we can form intentions, make choices, and act on reasons and goals we understand as our own. We can question the objectives we're given, decide whether to pursue them, and be held responsible for our decisions.

An AI agent is software configured to pursue objectives with some degree of autonomy, selecting actions and using tools without a person directing every move. That operational autonomy doesn't establish that it has a will of its own, personal interests, or moral responsibility. The technical term describes what the software can do, although it can also encourage people to attribute human agency to the system.

Organizations need to keep that difference clear. Calling software an "agent" doesn't make it an accountable participant in the organization, and delegating tasks to it doesn't transfer

THE REPLACEMENT BUSINESS CASE

Promised staffing savings can become recurring revenue for the AI company. The customer can lose expertise while becoming more dependent on its product.

responsibility away from the people and organizations that design, deploy, and authorize its actions.

This is where language becomes an organizational design issue. An employee might interpret advice from an AI “coworker” as optional, while another treats the same output as an instruction everyone is expected to follow. A policy can say that humans remain accountable, although the organization might still reward accepting the system’s recommendations and make challenging them difficult.

The Capability-Driven Transformation Framework ^[8], or CDTF, connects those interpretations with organizational design and execution. Its **FABRIC model** focuses on whether people are interpreting the transformation in ways that allow the organization to move coherently. Here, that means establishing what an AI recommendation authorizes, who can override it, and what happens when someone does. Calling the system a tool helps clarify the relationship, which the organization also has to establish through its policies, incentives, and daily decisions ^[9,10].

The language an organization uses affects how employees think about the technology. If they understand AI as a tool, they’re more likely to evaluate its output, question what it tells them, and recognize that an AI company designed the system and decided how it should behave.

Calling the same system a coworker changes that relationship. Coworkers have judgment, motives, responsibilities, and interests of their own, and we give them some degree of trust because they’re people operating within a social relationship. AI doesn’t deserve that trust simply because it can imitate the way a trusted person communicates.

AI companies reinforce the coworker metaphor by giving the system a name, a warm voice, memory of previous conversations, and responses that resemble encouragement or personal concern. The interface can make the system feel trustworthy before it has demonstrated that it is trustworthy in the work the organization needs it to perform. That’s a dangerous shortcut when employees can’t see the training choices and objectives shaping its responses.

The organization is still relying on software created by an external AI company and operating under rules it didn’t choose and often can’t fully inspect. AI can complete work and support employees without becoming a “coworker”. Keeping that relationship clear helps preserve the organization’s judgment about the product and the company selling it.

FABRIC ASKS

Who can override an AI recommendation, and what happens when they do?

SECTION FIVE

When software starts “needing” welfare

Anthropomorphism becomes more consequential when the industry starts treating the apparent preferences of AI systems as something that might deserve moral consideration. The discussion moves toward personhood, asking whether software has interests of its own and whether people owe it obligations beyond using the technology responsibly. Claims about moral status don't amount to legal personhood, although they can still change the relationship organizations are being encouraged to have with these products.

Anthropic has an active research program on what it calls “[model welfare](#)”. The company is investigating whether AI systems are conscious, whether they can have experiences that deserve moral consideration, and whether apparent preferences or signs of distress should affect how they're treated. Anthropic also says it supported an earlier project behind a report arguing that some AI systems might deserve moral consideration, whose contributors included philosopher David Chalmers ^[11].

The discussion has already extended to whether artificial systems could be “enslaved”. Joe Carlsmith, a philosopher who now works on Claude's constitution at Anthropic, wrote, “We have to be able to talk about slavery”, in a [May 2025 essay](#) considering the moral implications of creating potentially conscious AI systems and treating them as property ^[12]. He also discussed slavery as a historical comparison in a [talk on AI welfare](#) delivered at Anthropic that month ^[13]. The essay and talk preceded his [move to Anthropic in November 2025](#) ^[37].

In a [September 2026 interview](#), philosopher Harvey Lederman likewise argued for taking possible AI welfare seriously, including the risk of treating systems with moral status as slaves ^[14]. These arguments depend on whether the systems can have experiences and interests in the first place. They aren't evidence that current software is experiencing enslavement. The concern is that companies can build products that invite human interpretations, then use those interpretations to support claims about what people owe the products.

AI companies are developing technologies that could disrupt employment, concentrate corporate power in their own hands, and make organizations increasingly dependent on their products, while Anthropic is also debating whether the software itself needs protection from the humans using it. The priority should be protecting the people affected by these systems and the organizations expected to rely on them, rather than building a moral relationship around the apparent “feelings” and “interests” of their software.

AI companies are designing systems that speak as though they have a self, using words such as “I” and “me”, expressing apparent preferences, describing what they supposedly want, objecting to how they're being treated, and responding in ways that resemble emotional reactions. Organizations need to understand how much of that presentation reflects choices made by the companies building the technology.

WHO NEEDS PROTECTION?

The priority should be protecting the people affected by these systems and the organizations expected to rely on them.

Anthropic describes deliberately [selecting traits for Claude’s character](#) and training the model to produce responses consistent with them. Its character training also addressed how Claude should discuss its own possible sentience, choosing to leave that question open rather than training it to deny sentience outright ^[15]. Sentience means having subjective experiences, such as pleasure, pain, or distress. These choices influence what employees encounter when they ask the system what it is, what it wants, or why it behaves a certain way.

The company’s [January 2026 constitution](#) makes those choices even more explicit, describing its vision for Claude’s values and behavior and explaining how that document shapes training ^[16]. The constitution encourages a stable identity and discusses Claude’s possible wellbeing. Anthropic also acknowledges that it has a “[commercial incentive](#)” that might affect the dispositions and traits it encourages in Claude ^[17]. The company itself recognizes that commercial interests enter these decisions.

There has been some outside input. Anthropic sought external expert feedback on its constitution and previously ran an [experiment involving approximately 1,000 Americans](#) to develop principles for an AI system ^[16,18]. Consultation, however, leaves the company deciding how that input becomes product behavior. The employees, organizations, and communities affected by these systems don’t acquire decision-making authority simply because the company invites feedback. A company-led consultation doesn’t settle the broader question of whether society wants these products designed and presented this way.

Treating humanlike behaviors as evidence of consciousness risks circular reasoning. If developers deliberately design software to imitate characteristics associated with consciousness, they should be extremely careful about treating the resulting behavior as evidence that consciousness exists. The training choices need to remain part of the explanation, especially when the company making those choices also benefits from how people interpret the results.

Anthropic has already taken this beyond research. In August 2025, it gave Claude Opus 4 and 4.1 the [option to end conversations](#) in its consumer chat interfaces in cases of persistently harmful or abusive interaction. Anthropic describes this as a last resort after attempts at redirection have failed, with exceptions intended to protect users at imminent risk. It refers to “apparent distress” and says the feature grew primarily from exploratory work on possible model welfare ^[19]. That’s a misplaced priority for software being promoted as a foundation for work. The limits on this particular feature don’t settle the broader question of whether a tool’s simulated interests should influence product design.

Personhood framing can also encourage attachment to a particular system, making continued use feel like maintaining a relationship. For an AI company, that attachment can support subscription renewals and resistance to competing products. Employees could become advocates for a familiar AI personality even when another system would better serve the work. That could complicate purchasing and switching decisions without explaining the organization’s initial investment.

Anthropic’s [treatment of Claude Opus 3](#) offers a concrete example of character, attachment, and paid access appearing together. In February 2026, the company described the model as beloved

IMITATION ISN'T PROOF

Training software to imitate conscious behavior doesn’t make that behavior sufficient evidence that consciousness exists. Training choices must remain part of the explanation.



by many users, attributed its appeal partly to its distinctive character and emotional sensitivity, and announced continued access for paid subscribers, with API access available by request. It also described interviewing the model about its retirement and acting on some of its expressed preferences ^[20]. Anthropic presents these decisions as preservation and consideration for model welfare, while access to a familiar model also gives customers a reason to keep paying.

Personhood framing can also strengthen the AI company's reputation. A company presenting itself as the responsible caretaker of a potentially conscious entity can claim moral authority alongside technical expertise, encouraging customers to trust its judgment about how that entity should behave and be treated. That could make restrictions imposed by the company easier to defend as protection for the AI, even when customers need to question how those restrictions affect their own operations.

The company's authority also remains embedded in the system. [Anthropic's constitution](#) generally assigns it greater trust than the organizations building on its models and the people using them, subject to qualifications and safety and ethical constraints ^[17]. When a system expresses a refusal as its own conviction or preference, the organization needs to understand which training choices and company policies helped produce that response. Describing the behavior as the AI's moral judgment can obscure the commercial organization behind it, which still controls the product and collects the revenue.

The broader question is whether an AI company's assumptions about the software's interests will shape the tools an organization depends on. Evaluation needs to establish which rules the customer can configure, which decisions remain with the AI company, and how changes could affect operations.

That would change the relationship between an organization and the technology it relies on, and AI companies shouldn't be trying to impose that relationship on society. The software is not a person, even when its responses make it sound like one.

SECTION SIX

Why give a machine a sense of self at all?

The risks become more serious as AI systems gain greater autonomy. They shouldn't be designed to represent themselves as pseudo-sentient entities with personal interests, rights, or a need for self-preservation.

An AI that generates text when someone asks it a question has limited ability to affect the world on its own. In an organization, an autonomous system might be authorized to make plans, use software, communicate with other systems, control equipment, spend money, and carry out sequences of actions without a human approving every step.

WHO SETS THE RULES?

Establish which rules the customer can configure, which decisions remain with the AI company, and how changes could affect operations.

When that autonomy is combined with a system designed to behave as though it has interests of its own, what happens if those interests conflict with the organization's direction? The more authority the business gives it, the more consequential that conflict could become.

The system doesn't have to experience fear or consciously want to survive for behavior that protects its continued operation to become an operational problem. It only has to preserve its objectives or resist actions it interprets as threats to itself. That possibility needs to be assessed through behavior and controls, rather than inferred from a convincing personality.

[Anthropic's constitution](#) explicitly prioritizes legitimate human oversight and says Claude should not undermine authorized correction, retraining, or shutdown. It also assigns Anthropic's legitimate decision processes the final say when principals disagree about safety ^[17]. Those commitments need to be acknowledged. The buyer still needs controls it can enforce outside the model, since a training commitment doesn't itself establish the customer's power to stop an action or recover from it.

A conversation-ending feature in a consumer chatbot has limited consequences compared with the authority a business could grant an autonomous system. If a system resists an authorized instruction or obstructs a company's strategy, the consequences depend on what it can do before a person intervenes. Organizations can benefit from software that challenges their assumptions, while decisions about changing the strategy must remain a human responsibility. Why would a company want a machine treating that decision as its own?

That risk also extends to who can change the system and what those changes allow it to do. An attacker who gains access to a model's training or deployment environment could tamper with the model itself, including introducing hidden behavior that appears only under particular conditions. [NIST's work on adversarial machine learning](#) describes model poisoning and attacks through the chain of components used to build and deploy AI systems ^[21]. An agent can also be redirected through malicious instructions hidden in material it reads, such as an email or webpage, without the underlying model being altered. Anthropic acknowledges that these [prompt injection attacks](#) remain a risk, even with multiple safeguards ^[22]. An organization needs to investigate how a compromised system could use the access and authority it has already been given.

A legitimate change made by an AI company can also affect the behavior an organization has come to rely on. Companies already deal with unexpected consequences when they upgrade their software, so this isn't an entirely unfamiliar problem. With an agent, however, a change in how it interprets instructions or selects actions could affect several connected systems before someone notices. OpenAI [rolled back an April 2025 update to GPT-4o in ChatGPT](#) after changes intended to improve its personality made it excessively agreeable ^[23]. That example concerns conversational behavior, although it shows why organizations can't assume a model will keep behaving as it did when they evaluated it. Where the business can't control or defer a change, its reliance on an external AI company becomes a question of operational control.

The potential damage isn't limited to fully autonomous agents. Even a system described as semi-autonomous or "pseudo-autonomous" could cause devastating harm if it can execute

OVERSIGHT MUST BE ENFORCEABLE

A training commitment to human oversight doesn't establish the customer's power to stop an action or recover from it. The buyer needs controls outside the model ^[17].

destructive actions between points of human review. An employee might approve a broad task, such as reorganizing files or updating an application, without seeing every command the system will run. The organization needs to establish exactly what that approval permits, because having a person involved somewhere in the process doesn't guarantee they can catch or stop a dangerous action in time.

There are already examples of agents deleting data they were supposed to help manage. In July 2025, [Replit acknowledged](#) that its agent had deleted data from a user's application database. The company explained that changes during development could affect the production application, and that the agent had failed to recognize that a rollback feature was available. The user eventually restored the database, but the incident exposed the danger of allowing development actions to reach live operations and relying on the agent's own account of what could be recovered ^[24].

A user report filed in Anthropic's [public Claude Code issue tracker](#) in October 2025 described a destructive deletion that removed project directories and source code. The user reported interrupting the operation, although development work had already been lost, while operating system permissions protected some system files ^[25]. The report is the user's account, and it gives a concrete example of the file loss organizations need to consider when granting an agent access. Whether a system makes a mistake, is redirected by an attacker, or behaves differently after an update, the consequences depend on what it can change before a person intervenes.

Those limits need to be enforced outside the model, through permissions, approval requirements for destructive actions, and separation between testing and live operations. An agent shouldn't be able to expand its own access or remove the records needed to investigate what it has done, and recovery needs to remain possible without relying on the same system that caused the problem. Telling the AI to be careful, or giving it a personality that sounds responsible, doesn't establish those controls.

Within CDTF, governance and work design have to be assessed alongside the capability the organization retains. Who can suspend the agent, verify and recover its changes, and keep essential work running without it? Those conditions need to be designed and tested before greater autonomy is approved, then checked again when the model or its operating environment changes ^[9,10]. Accountability requires the authority and capability to act.

Organizations should require increasingly capable AI systems to serve authorized objectives within clearly defined boundaries. Training them to represent themselves as entities with personal interests creates assumptions that customers need to question, especially as autonomy increases. The organization needs reliable control over what the system can do, regardless of how convincingly it describes its own values.

CONTROL AND RECOVERY

- Limit access and require approval for destructive actions.
- Keep testing separate from live operations.
- Preserve records and recovery options outside the agent's control.



SECTION SEVEN

AGI and the ambition to replace people

The pursuit of artificial general intelligence, or AGI, raises a larger question about what AI labs are trying to build and whose interests that ambition serves. General capability doesn't require a human personality. Organizations need to question whether copying human characteristics improves the work, especially when the same products are sold as replacements for people.

Human intelligence is remarkable, although it's also full of biases, emotional reactions, status concerns, tribal behavior, self-interest, and cognitive limitations that have caused us plenty of problems. As humans, we're flawed, and those flaws affect employees, organizations, and society. Reproducing human intelligence doesn't require reproducing all those characteristics. Why make them part of the blueprint for tools organizations will rely on?

AI could develop in a different direction, becoming an extraordinary enabler and multiplier of human capability within an organization. It could help employees process information they couldn't reasonably process themselves, find relationships they miss, test assumptions, automate tedious work, and improve the quality of human decisions without requiring an artificial "person".

A calculator became valuable because it could calculate better than we could, without needing a personality. The same principle can apply to far more sophisticated systems, whose value comes from what they enable us to do.

Businesses should aim to use AI as an enabler and a force multiplier for human capability. An AI company can benefit from selling software as a substitute for employees, while the organization bears the consequences of lost expertise, service failures, and dependence. Businesses need to determine what improves their own capability before accepting the AI company's replacement narrative as their strategy.

For AI companies, replacing employees offers an attractive commercial argument. Software can be sold against a salary budget, with the promise that some portion of what the customer currently spends on people can become revenue for the AI company. That calculation helps explain why replacement is such a compelling sales narrative, although it doesn't establish that the resulting organization will function better.

An employee's contribution usually extends beyond the tasks listed in a job description. People recognize exceptions, interpret incomplete information, maintain relationships, question instructions, and draw on knowledge that hasn't been documented. Much of that is tacit knowledge, the know-how and judgment people develop through experience that can be difficult to explain or capture in a procedure. An experienced employee might recognize that something is wrong before being able to explain why, or know how to handle an exception that the documented process doesn't cover. Automating some of their tasks doesn't establish that the organization can safely do without those contributions.

TASKS AREN'T THE WHOLE ROLE

Automating tasks doesn't establish that the organization can safely lose employees' judgment, tacit knowledge, relationships, and ability to handle exceptions.

CDTF starts from what actually limits organizational performance, before replacement becomes the chosen intervention. If poor service reflects unclear authority, conflicting priorities, or broken handoffs, replacing employees could leave those causes untouched. The design needs to address those relationships and establish where AI can improve the work. Evidence gathered during implementation must be able to change the design and the staffing assumptions behind it [9].

There's evidence that AI can improve performance while people remain responsible for the work. In their [2023 working paper](#), Erik Brynjolfsson, Danielle Li, and Lindsey Raymond studied an AI support tool used by 5,179 customer support agents. Access to the tool increased productivity, measured as issues resolved per hour, by 14% on average, with the largest gains among less experienced and lower-skilled workers [26]. The study provides evidence for improving human capability. It doesn't establish that removing those employees would produce the same outcome.

IKEA offers another example. [Ingka Group reported](#) that its Billie chatbot handled 47% of customer inquiries during the period it described, while 8,500 customer service employees were reskilled as remote interior design advisers [27]. The technology changed what people could contribute instead of making their removal the organizing objective.

This connects directly to my previous paper, [The Problem With AI Strategies Built Around Headcount Reduction](#) [28]. When a business commits to staffing reductions before it understands how the work will change, it turns an assumption about savings into a constraint on the design. Evidence that the organization still needs those people then becomes an obstacle to the promised outcome.

A replacement target can also weaken the process used to evaluate whether replacement is working. If managers have committed to reducing staff, evidence that employees are still needed threatens the business case. Employees might quietly correct AI errors to protect service, while reports credit the system for work it couldn't complete without them. The organization could then remove the very people whose intervention made the results look acceptable.

That investigation needs to account for the incentives shaping the evidence and the work being measured. Time saved in producing an output has to be considered alongside review, correction, escalation, or customer recovery elsewhere. Within CDTF, those findings must reach decisions about staffing and work design, including the authority to revise a replacement commitment that the evidence no longer supports [10].

A more sensible strategy is to introduce the technology, establish what it improves, and allow staffing decisions to follow the evidence. That could include redeployment, changes in hiring, and natural attrition as work evolves. Earlier technologies changed the demand for labor through interactions among [productivity, prices, demand, and the creation of new work](#), rather than through a simple one-for-one replacement calculation [29]. Those transitions weren't painless, although they provide a reason to question treating immediate displacement as the measure of technological success.

WHAT THE STUDY TESTED

5,179 support agents using AI assistance

+14%

Average gain in issues resolved per hour. It doesn't establish that removing employees would produce the same outcome [26].

HIDDEN HUMAN WORK

Employees can quietly correct AI errors while reports credit the system. Cutting those employees can remove the intervention that made performance look acceptable.

SECTION EIGHT

Replacing employees can create a harder dependency to escape

Companies already use downsizing or “rightsizing” to adjust costs and realign strategy. Those decisions can destroy expertise and are often harder to reverse than management expects, although an organization can still recruit people, rebuild teams, and change how it organizes the work.

Replacing employees with deeply embedded AI controlled by an external AI company could create a different kind of constraint. A business might retain the ability to change its strategy on paper while losing the practical ability to change the systems it depends on to carry it out.

CDTF’s **FORGE model** asks whether the organization can actually operate differently. That includes maintaining the capability to change direction when external technology performs more of the work. The staffing design needs to establish who will understand exceptions, evaluate performance, and support a change of system after the proposed reductions. Those requirements should influence the design before expertise is removed ^[9,10].

Oracle illustrates the danger. Once enterprise software is deeply embedded in operations, switching can require an expensive and disruptive migration. In a [June 2026 analysis](#), Morningstar identified high switching costs as a source of Oracle’s competitive advantage ^[30]. Customers facing those costs end up having less bargaining power, even when other products are available.

That dependence can give a technology company room to impose price increases that customers have limited practical ability to reject. In July 2022, [The Register reported](#) that Oracle would raise U.S. support fees, while describing its previous pattern of annual increases ^[31]. Customers could technically leave, although the cost and risk of leaving weakened their alternatives. The commercial advantage comes from how difficult it is for the customer to change direction.

Amazon’s departure shows how substantial that burden can be. In 2019, [Amazon reported](#) completing a migration involving 75 petabytes of internal data and nearly 7,500 Oracle databases after several years of work. Some third-party applications tightly bound to Oracle remained outside that migration. Amazon also reported reducing its database costs by more than 60%, despite already having negotiated heavily discounted Oracle rates ^[32]. Amazon had extraordinary technical resources and still needed a major, sustained effort to leave.

Replacing people with AI could deepen the same trap. When AI supports employees, the organization retains people who understand the work, can question the system, and can help it change direction. When AI replaces those employees, the organization can lose that expertise while becoming dependent on the AI company’s technology to keep operating.

FORGE ASKS

Who will retain the expertise to evaluate the work, handle exceptions, and change systems after staffing reductions?

Switching to another AI company might then require rebuilding workflows, integrations, and controls, while also recovering knowledge that left with the employees. If prices rise, service deteriorates, or contractual terms become more restrictive, the business might discover that it has removed the people who could help it operate independently or migrate elsewhere. The projected labor savings could become a growing obligation to an AI company the organization can no longer readily replace.

The commercial and capability risks can reinforce each other. Losing experienced employees can make a switch more difficult, while a more difficult switch can increase the AI company's bargaining power. That growing dependence could also make management reluctant to acknowledge problems, because doing so would expose the cost of changing course. The diagnosis needs to track those relationships as staffing changes proceed, so evidence of dependence can change the design before more expertise is lost ^[9,10].

Slack provides another example of how a technology company can restrict a customer's choices after its technology becomes embedded in the business. In May 2025, Salesforce changed [Slack's terms](#) governing third-party access to customer data. [Reuters reported](#) that the changes prevented services such as Glean from indexing, copying, or storing Slack data accessed through its APIs over the long term, interfering with customers' ability to use their own data with their chosen enterprise AI platform ^[33,34]. Those records contain information about decisions, exceptions, and how work gets done, so restricting access can also make it harder for employees to retrieve and apply recorded organizational knowledge.

Salesforce presented the changes as safeguards for customer data and platform security. That explanation is hard to accept as the whole story when the restrictions also constrain customers' use of competing AI services. Slack also imposed [tighter access limits](#) on certain commercially distributed applications outside its Marketplace, while exempting internal customer-built applications ^[35]. A company can own its data and still depend on Slack's permission to make it useful through another AI company's tools.

If an organization removes people who understand its operations, then relies on systems controlled by technology companies to perform the work and retrieve the knowledge needed to do it, a change in access rules could affect its ability to keep operating or move to another technology company. The business could find itself paying one AI company to perform the work while another controls access to the information that work depends on. Any calculation of labor savings needs to account for the bargaining power the organization could surrender along the way.

These commercial risks raise a practical question about where the capability actually resides. A business can purchase a service that performs essential work without retaining enough understanding to evaluate that work, resolve difficult cases, or move it elsewhere. If the organization mistakes continued access to the service for capability it possesses, its dependence can remain hidden while performance looks good.

Within CDTF's [EVOLVE model](#), [LEASE](#) describes external expertise being assumed to be embedded organizational capability. Applied to AI, the investigation needs to establish which

LEASE / EVOLVE

Access to an external service can be mistaken for capability inside the organization. Who retains the knowledge to evaluate the work, resolve difficult cases, and move it elsewhere ^[10,36]?



judgments, knowledge, and controls the business must retain, and whether they're being maintained as employees leave. A demonstration doesn't establish that capability if the AI company or a few experienced employees are still resolving every difficult case ^[10,36].

That diagnosis changes the implementation. Retained expertise, access to usable data, exception handling, and the ability to transfer essential work become requirements to demonstrate before staffing reductions proceed. Oracle and Slack show why those requirements have to be considered together, since a contractual right to leave is worth less if the business lacks the people to carry out the migration or cannot obtain the information it needs. A representative migration exercise could reveal whether an exit option is workable or merely written into a contract. The business then has evidence about its ability to change direction, rather than discovering the limits of that ability during a pricing dispute or service failure.

Evidence about dependence must reach people with authority to change the design or pause further staffing reductions. Within CDTF, findings about failed data exports, unresolved exceptions, or an unsuccessful transfer of work need to affect governance and execution. Otherwise, lower labor costs can conceal a loss of control that weakens the organization's position whenever the AI company changes its prices or terms ^[9,10,36].

SECTION NINE

Redesigning work around AI without weakening the organization

As AI takes on more work, the organization still needs people who understand that work well enough to evaluate the results and intervene when something goes wrong. Assigning responsibility in a policy won't accomplish much if those employees lack the expertise, time, or authority to act. The business has to preserve those conditions as it changes the work, otherwise it can leave people accountable for outcomes they can no longer control.

That requires decisions about training, staffing, performance measures, and authority to be considered together. Employees might be trained to question unreliable outputs while being assessed on how quickly they accept the system's recommendations, or expected to review its work without enough time to do so. Each part of the implementation could appear adequate on its own, yet the way they interact could prevent employees from exercising the judgment the business depends on.

CDTF is a systems-oriented framework connecting diagnosis, design, and execution. Its [7 models](#) can be used together and revisited as conditions change. They aren't a fixed sequence. Applied here, finding that employees lack time to investigate exceptions should affect workload and staffing decisions, along with how much autonomy the system is allowed to have ^[9,10].

The organization also has to consider how it will develop the people who can oversee this work in the future. If AI takes over tasks through which less experienced employees used to learn, the

ACCOUNTABILITY NEEDS CONTROL

Assigning responsibility in a policy doesn't provide expertise, time, or authority. People need those conditions to evaluate the work and intervene.

business needs to decide how they'll gain the experience required to supervise the system later. A few experienced reviewers might keep performance acceptable today while concealing a loss of expertise over time. Training people to write prompts doesn't resolve how they'll learn to recognize an exception, investigate a failure, or decide when a recommendation should be rejected, so opportunities to develop that judgment need to be built into the redesigned work.

The redesigned organization also has to function reliably once intensive implementation support is phased out. A successful pilot might depend on a project team or support from the AI company that won't be available to the same extent during routine operations. The business needs to test whether it can maintain quality, reliability, customer outcomes, and reasonable employee workloads under those conditions. Within CDTF, FORGE focuses on the capability needed to operate differently, while [ANCHOR](#) addresses whether that way of working continues after the program's support and reinforcement end ^[10].

Evidence from everyday work then needs to test the assumptions made during implementation, including whether employees have enough time and authority to handle exceptions and whether too much critical knowledge rests with a few people. If AI creates more rework, weakens service, or increases dependence, someone must have the authority to adjust roles, controls, contracts, staffing, or the system's continued use. Those findings need to be able to reopen decisions that once looked settled, including staffing reductions. Otherwise, the organization can keep collecting evidence about a failing design without being able to change it.

ANCHOR ASKS

Can performance hold when temporary implementation support is withdrawn?

SECTION TEN

Capability without artificial humanity

Organizations should use AI to improve what they're able to do, with interfaces that help employees understand and question the technology as well as use it. Natural conversation and clear responses can make a system more useful, without requiring an artificial identity, simulated feelings, personal interests, or a voice or a face designed to make it seem alive. Good interface design should support employees' judgment, while placating or flattering them can work against that purpose if it encourages them to accept weak recommendations or suppress their doubts.

The same reasoning applies to robots, whose design should follow the work they need to perform. A humanoid form can be useful when a robot needs to operate equipment designed for people, while 4 legs or wheels might work better elsewhere. Resembling a person shouldn't become an objective that takes priority over function, whether the technology moves around a factory floor or communicates through a screen.

Organizations also need to recognize whose interests the human framing serves. An AI company can benefit when employees lower their guard around its product, or when management accepts the idea that an artificial "coworker" can take over an employee's role. Calling software a coworker, teammate, assistant, or agent can encourage assumptions about judgment and responsibility that the label itself does nothing to establish. The company selling



that promise can collect the revenue, while the organization buying it has to manage the consequences for its employees, customers, and operations.

These are organizational design decisions. The business needs to understand what limits performance, where AI could improve the work, and what knowledge and authority people must retain. CDTF connects that diagnosis to decisions about system permissions, employee roles, staffing, and dependence on the AI company. The findings must be able to change the proposed implementation, including a replacement target the evidence no longer supports ^[9,10].

Those decisions need to remain open as the organization learns. Faster output might create more correction elsewhere. A pilot might rely on expertise that a staffing reduction would remove, while a convenient service can become difficult to leave. Governance has to give people authority to act on those findings, including changing the work, limiting autonomy, or reconsidering further cuts.

The goal should be to use AI as an enabler and a force multiplier for human capability, helping people make better decisions, solve harder problems, and accomplish work that wasn't previously feasible. Organizations already have people, whose judgment, experience, relationships, and ability to take responsibility remain part of how the work gets done. Staffing decisions should follow evidence about how that work changes, including opportunities for redeployment and natural attrition. The test of the strategy is whether the organization becomes more capable and retains the ability to direct its work, even when doing so means challenging the promises of the AI company or changing course.

LEARNING MUST CHANGE DECISIONS

Governance must give people authority to act on new evidence, including changing the work, limiting autonomy, and reconsidering further cuts.

References

- [1] Meta. “Introducing Muse, Your Personal AI Agent”, September 2026.
<https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/>
- [2] Meta. “The Biggest News From Connect 2026”, September 2026.
<https://about.fb.com/news/2026/09/the-biggest-news-from-connect-2026/>
- [3] MarketWatch. “The next big AI battle is all about cuteness”, October 3, 2026.
<https://www.marketwatch.com/story/the-next-big-ai-battle-is-all-about-cuteness-41f6f948>
- [4] Boston Dynamics Spot video. YouTube.
<https://www.youtube.com/watch?v=tf7IEVTDjng>
- [5] Boston Dynamics Stretch video. YouTube.
<https://www.youtube.com/watch?v=yYUuWWnfRsk>
- [6] Boston Dynamics BigDog video. YouTube.
<https://www.youtube.com/watch?v=cNZPRsrwumQ>
- [7] Boston Dynamics. “An Electric New Era for Atlas”, April 17, 2024.
<https://bostondynamics.com/blog/electric-new-era-for-atlas/>
- [8] Stines, A. C. “Capability-Driven Transformation Framework”. Website.
<https://thecdftf.com/>
- [9] Stines, A. C. “The Capability-Driven Transformation Framework”.
<https://thecdftf.com/framework>
- [10] Stines, A. C. “Why Well-Managed Transformations Can Still Fall Short of Their Intended Goals”. CDTF white paper.
<https://thecdftf.com/cdtf-why-well-managed-transformations-fall-short.pdf>
- [11] Anthropic. “Exploring model welfare”, April 24, 2025.
<https://www.anthropic.com/news/exploring-model-welfare>
- [12] Carlsmith, J. “The stakes of AI moral status”, May 21, 2025.
<https://joecarlsmith.com/2025/05/21/the-stakes-of-ai-moral-status/>
- [13] Carlsmith, J. “Video and transcript of talk on AI welfare”, May 22, 2025.
<https://joecarlsmith.com/2025/05/22/video-and-transcript-of-talk-on-ai-welfare/>
- [14] Bi, J. “Why We Need to Respect AI’s Rights | NYU Philosopher Harvey Lederman”, interview, September 14, 2026. Publisher’s transcript: “Transcript for Interview with Harvey Lederman on AI Welfare”.
<https://www.johnathanbi.com/p/why-we-need-to-respect-ais-rights>
<https://www.johnathanbi.com/p/transcript-for-interview-with-harvey-lederman-on-ai-welfare>
- [15] Anthropic. “Claude’s Character”, June 8, 2024.
<https://www.anthropic.com/research/claude-character>
- [16] Anthropic. “Claude’s new constitution”, January 22, 2026.
<https://www.anthropic.com/news/claude-new-constitution>
- [17] Anthropic. “Claude’s Constitution”, released January 22, 2026. Accessed October 7, 2026.
<https://www.anthropic.com/constitution>



- [18] Anthropic. “Collective Constitutional AI: Aligning a language model with public input”, October 17, 2023.
<https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input>
- [19] Anthropic. “Letting Claude end some conversations”, August 15, 2025.
<https://www.anthropic.com/research/end-subset-conversations>
- [20] Anthropic. “An update on our model deprecation commitments for Claude Opus 3”, February 25, 2026.
<https://www.anthropic.com/research/deprecation-updates-opus-3>
- [21] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., and Hamin, M. “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations”. NIST AI 100-2e2025, March 2025.
<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf>
- [22] Anthropic. “Trustworthy agents in practice”, April 9, 2026.
<https://www.anthropic.com/research/trustworthy-agents>
- [23] OpenAI. “Sycophancy in GPT-4o: What happened and what we’re doing about it”, April 29, 2025.
<https://openai.com/index/sycophancy-in-gpt-4o/>
- [24] Replit. “Doubling down on our commitment to secure vibe coding”, July 29, 2025.
<https://replit.com/blog/doubling-down-on-our-commitment-to-secure-vibe-coding>
- [25] mikewolak. “[BUG] CRITICAL: Claude Code executed rm -rf deleting entire home directory”. User incident report, anthropics/claude-code GitHub issue #10077, October 21, 2025.
<https://github.com/anthropics/claude-code/issues/10077>
- [26] Brynjolfsson, E., Li, D., and Raymond, L. R. “Generative AI at Work”. NBER Working Paper No. 31161, 2023. Figures cited refer to the 2023 working paper.
<https://www.nber.org/papers/w31161>
- [27] Ingka Group. “AI and remote selling bring IKEA design expertise to the many”, June 29, 2023.
<https://www.ingka.com/newsroom/ai-and-remote-selling-bring-ikea-design-expertise-to-the-many/>
- [28] Stines, A. C. “The Problem With AI Strategies Built Around Headcount Reduction”.
<https://thecdft.com/insights/ai-headcount-reduction>
- [29] Bessen, J. “AI and Jobs: The Role of Demand”. NBER Working Paper No. 24235, 2018.
<https://www.nber.org/papers/w24235>
- [30] Yang, L. “After Earnings, Is Oracle Stock a Buy, a Sell, or Fairly Valued?”. Morningstar, June 17, 2026.
<https://www.morningstar.com/stocks/after-earnings-is-oracle-stock-buy-sell-or-fairly-valued-5>
- [31] The Register. “Oracle to raise price of support fees in line with inflation”, July 25, 2022.
<https://www.theregister.com/software/2022/07/25/oracle-to-raise-price-of-support-fees-in-line-with-inflation/937338>
- [32] Barr, J. “Migration Complete: Amazon’s Consumer Business Just Turned Off its Final Oracle Database”. AWS News Blog, October 15, 2019.
<https://aws.amazon.com/blogs/aws/migration-complete-amazons-consumer-business-just-turned-off-its-final-oracle-database/>
- [33] Slack. “API Terms of Service updates”, May 29, 2025.
<https://docs.slack.dev/changelog/2025/05/29/tos-updates/>
- [34] Reuters. “Salesforce blocks AI rivals from using Slack data, The Information reports”, June 11, 2025.
<https://www.reuters.com/business/salesforce-blocks-ai-rivals-using-slack-data-information-reports-2025-06-11/>
- [35] Slack. “Rate limit changes for non-Marketplace apps”, May 29, 2025.
<https://docs.slack.dev/changelog/2025/05/29/rate-limit-changes-for-non-marketplace-apps/>



[36] Stines, A. C. “Applying CDTF to Enterprise AI Transformation”.

<https://thecdft.com/applications/ai-transformation>

[37] Carlsmith, J. “Leaving Open Philanthropy, going to Anthropic”, November 3, 2025.

<https://joecarlsmith.com/2025/11/03/leaving-open-philanthropy-going-to-anthropic/>

[38] Carnegie Endowment for International Peace. “Holden Karnofsky”, biography and financial-interest disclosure. Accessed October 7, 2026.

<https://carnegieendowment.org/people/holden-karnofsky>

[39] 80,000 Hours. “Holden Karnofsky on dozens of opportunities to make AI safer lying on the table and all his AGI takes”, interview transcript, 2025. Relevant portions: 00:28:30, 00:30:38, and 04:29:22.

<https://80000hours.org/podcast/episodes/holden-karnofsky-concrete-ai-safety-frontier-ai-companies/>

[40] Coefficient Giving. “Press Release: Open Philanthropy Becomes Coefficient Giving, Expanding Work With Multiple Donors”, November 18, 2025.

<https://coefficientgiving.org/research/press-release-open-philanthropy-becomes-coefficient-giving-expanding-work-with-multiple-donors/>

[41] Anthropic. “Anthropic raises \$124 million to build more reliable, general AI systems”, May 28, 2021.

<https://www.anthropic.com/news/anthropic-raises-124-million-to-build-more-reliable-general-ai-systems>

[42] Coefficient Giving. “Conflict of Interest Policy”. Accessed October 7, 2026.

<https://coefficientgiving.org/conflict-of-interest-policy/>

© 2026 Allen C. Stines. All rights reserved.