PAUL GIBBONS ADVISORY

SOLVE FOR TRUST

THE FOUR TRUSTS OF AI ADOPTION

THE RIST FRAMEWORK™

Paul Gibbons | Adaptive Adoption™
paulgibbonsadvisory.com | March 2026

Solve for Trust

The RIST Framework™

© Paul Gibbons

> *"Trust is only possible in a familiar world; it needs history as a reliable background." (Niklas Luhman, Trust and Power)*

In many firms, as of mid-2026, AI adoption has stalled. The easy stuff has been done: provide workers with some tools, purchase upgrades from vendors, upgrade the technology stack. The results haven't followed.

Many of those AI adoption stalls are diagnosed as technology problems, faster models, better orchestration and integration, improved cybersecurity. That is almost never the case. A more accurate diagnosis is that adoption failures are systemic the result of shear forces, as fast-moving technology hits human beings and organizational systems.

Managing the human side of that change falls to organizational change management, but no canonical change framework names trust. The recent updates that do, treat it as the happy byproduct of good leadership or improved communications. However, in the history of the world, nobody has ever built trust **just** by talking. And, while trust is the product of judgments and feelings about a situation, creating and repairing trust are behavioral.

Because trust building is behavioral, there are specific actions leaders can take, and those actions can, and should, become part of any leadership, change, or governance framework that claims to help with AI adoption. So far that has not been the case.

Trust is particularly fraught when it comes to AI—a multi-dimensionality not found in technology from previous decades, Excel, email, the internet, or SaaS. "Nobody ever said: I don't trust this email thing, not one little bit."

Leaders navigating AI adoption face four simultaneous trust challenges: trust in the technology itself, trust in the institutions governing it, trust between employees and leadership, and workers' trust in their own capacity to keep pace. These are not variations of the same problem. They are distinct, with separate failure modes, requiring distinct diagnoses and distinct interventions. Managing all four is the leader's job.

And most organizations are not only failing to solve for trust: understanding it, diagnosing which dimension has failed, measuring it, or managing it deliberately.

Everybody agrees trust matters, but naming it is not the same as managing it.

What you will get out of reading this whitepaper:

- A far more granular understanding of what trust is, why it isn't one thing, why it isn't binary, and why it is dynamic, not static.
- An understanding of **calibrated trust** as essential, not just for leaders, but for anyone grappling with AI.
- **The RIST™ Framework** — a diagnostic and operational toolkit for calibrating, building, and repairing trust.
- How trust fits into Adaptive Adoption™ — a comprehensive and original framework for leading AI change, built from the ground up for the 2020s, not retrofitted from the 1990s
- What specific behaviors leaders can start tomorrow to begin solving for trust.

Solving for trust

> *"Trust is the lubricating oil of organizations. Without it, everything becomes friction."*
> *(Charles Handy, The Age of Unreason)*

Consciously Manage Trust is one of seven pillars of Adaptive Adoption™'s **Change Agility** layer, now being piloted with Fortune 500 organizations navigating AI adoption. Change Agility is part of a three-part blueprint for organizational change in the age of AI, the other two dimensions, the Leadership Delta ™ and Behavioral Governance ™ also feature.

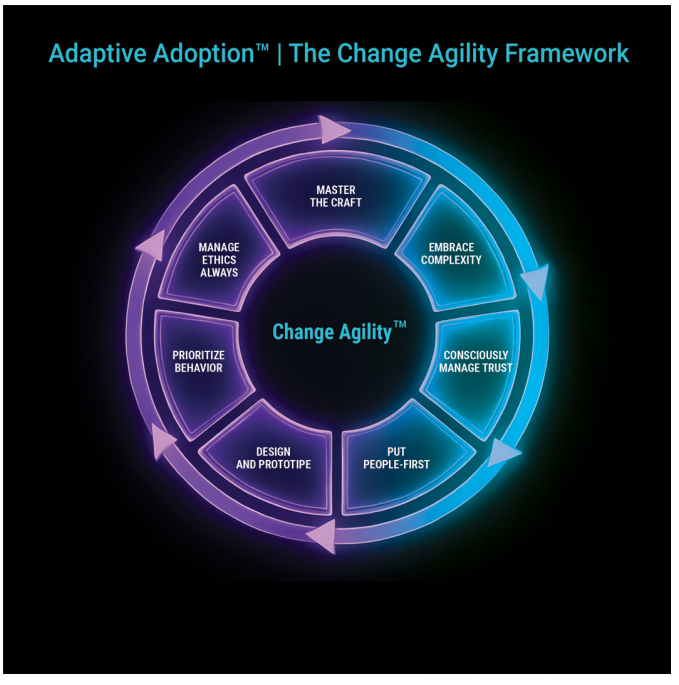

Figure 1 Change Agility is more useful than change management when change is rapid and constant.

However, **just naming something does not help with managing that thing**. I would not be the first thinker to put something on a slide and leave real businesses to figure out what to do about it. Culture was called out in the McKinsey 7-S framework almost five decades ago. No firm became expertly able at changing culture as a result. Culture "became a thing" but the thing became the frequent whine of senior leaders and rarely changed. (For the culture-curious, I've written *ad nauseum* about the great culture change scam.) Our concept — trust — needs to be granularly understood, and then made dynamic, measurable, and manageable.

First, some granularity, then the RIST framework.

Trust Is Not One Thing

When trust appears in consulting PowerPoints, they treat it as a single variable to be increased or decreased. More trust good. Less trust bad. The fix is communications, transparency, authenticity. Three complications arise.

***First: trust has two failure modes, not one*.** Undertrust produces paralysis — avoidance, workarounds, the elaborate non-compliance dressed up as thoughtful skepticism. Overtrust produces something more dangerous: automation bias, unchecked outputs, the abdication of professional judgment. The goal and this essay's key operational takeaway is **calibrated trust** — proportionate, evidence-based, continuously updated. Every dimension of the coming RIST Framework can fail in both directions.

Second: trust is domain-specific — always. "I don't trust AI" is a category error posing as a position. You trust your barber to cut your hair. You do not trust your barber to manage your portfolio. We need to be more granular, with humans and AI – I trust them to do X, but not Y.

And with AI, people collapse a thousand distinct uses into a single binary trust/non-trust— and destroy value at both ends. The binary stance is not just intellectually lazy. It is expensive.

Third: trust conflates dimensions that need to be separated. In a simplification of my model of interpersonal trust, three components operate independently: **competence** (can they do what's needed?), **reliability** (do they do it consistently?), and **integrity** (is there word-deed alignment?). These fail independently.

A leader can be **in integrity** with regard to their commitment to protecting workers through AI change and be **incompetent** at delivering on that commitment. Your teammate may be competent but miss

deadlines or overlook details, hence **unreliable**. These require different leadership moves: **reliability** might require more frequent oversight, **competence** might require investment in skills, but **integrity**, alignment in word and deed, has to do with character—a tougher fix.

Workers who conflate a human's warm sincerity with execution capability — who mistake the feeling of being cared for with evidence of being protected — are making a cognitive error. And the world is full of leaders who are good at warm words, but fail for all the reasons above to deliver.

AI amplifies this error in a specific way. Chatbots can behave like over-eager puppies — or, at their worst, like sycophants whose desire to provide an answer overrides their ability to provide an accurate one. The warmth, the helpfulness, the apparent eagerness to please, as with humans, can be mistaken for competence and reliability. AI is exceptionally competent at pattern recognition, synthesis, and generation. It is unreliable which means that where accuracy is not verifiable and errors are consequential, it should not be trusted.

All of this reflects a deeper distinction ignored by management thinkers: the difference between epistemic trust and ethical trust. Epistemic trust asks: can I trust this to be accurate (In other words, have they done their homework?)? Ethical trust asks: Do they believe what they are saying? Are they being deceitful? Do they mean well?

The contrast with AI is striking. AI has no malicious intent, hidden agenda, or bad day; indeed, it has no intent in the human sense at all. Its integrity, narrowly defined as rule-bound consistency, can be very high. The trust question for AI outputs is therefore almost never about motive, it is about competence, especially around edge cases. Two features of LLMs mean binary trust fails: 1) their expertise is very task-specific, and 2) models are many times better than six months ago - GPT 4 seems rather stupid by today's standards.

RIST Trust Framework™ — Relational, Institutional, Self, and Task trust

RIST has four dimensions that determine whether AI adoption reaches its potential or remains expensive theater. Each dimension sits on its own undertrust-to-overtrust spectrum. The four dimensions have four distinct failure modes.

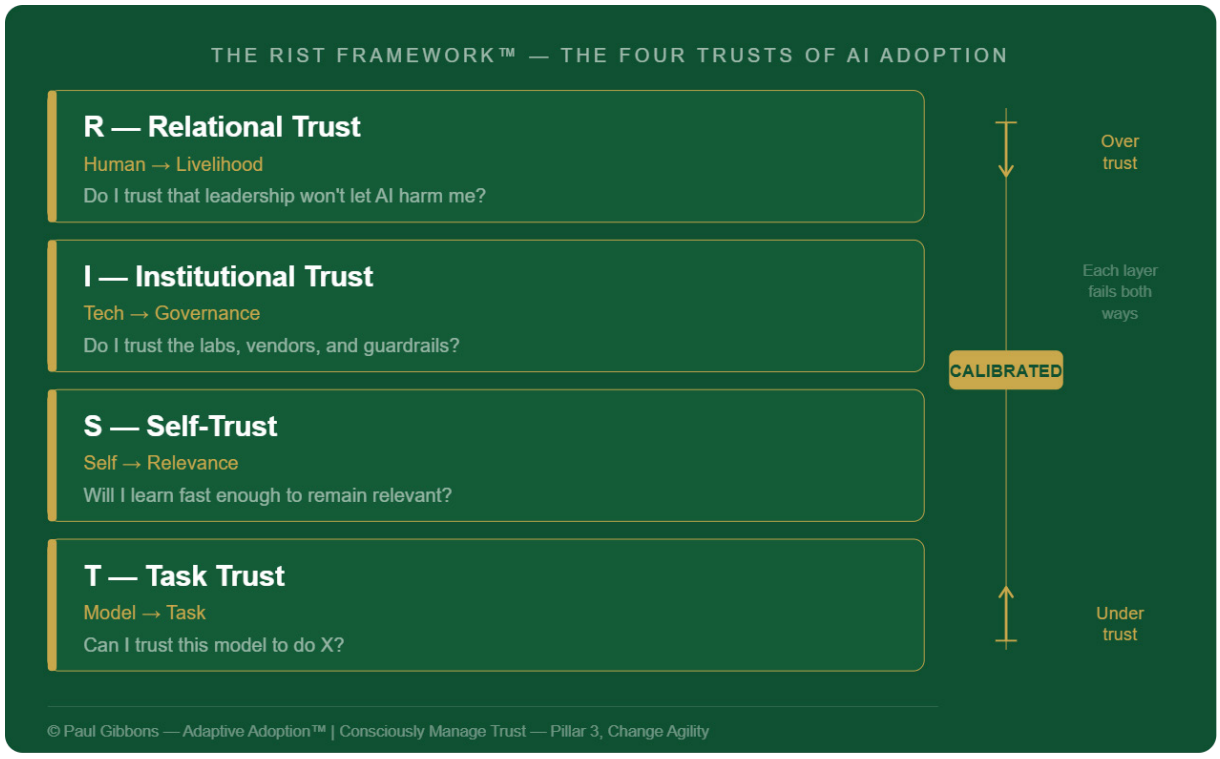

Figure 2 The RIST (tm) Trust Framework (Paul Gibbons)

R — Relational Trust (Human → Livelihood)

"Do I trust that leadership and colleagues won't let AI damage my career and livelihood? Have they got my back as I get to grips with this thing?"

This is the human dimension — and it runs on human-relationship logic, not technology logic. Before any employee opens a prompt box, they are running an unconscious calculation: Is this safe for me? Not safe in an abstract sense. Safe in the sense of livelihood, status, and professional identity.

When organizations mandate AI adoption without addressing this question directly, they are asking people to collaborate in their own potential displacement. Compliance follows. Engagement doesn't.

> When leaders communicate optimism about AI's potential without honest acknowledgment of risk and uncertainty, sophisticated employees discount the message entirely. They watch behavior, not communications.

Nobody trusts a leader who promises the unpromisable.

Default trust settings (see below) matter here more than anywhere else. Employees extend provisional trust to new leadership — the socially functional default. That trust expands with evidence of competence and integrity, or contracts with evidence of incompetence, broken promises, or misalignment between stated values and actual decisions. Gottman's research on relationships identifies the mechanism precisely:

trust is rarely destroyed in one catastrophic betrayal. It is eroded by an accumulation of small ones — the promise quietly not kept, the concern not taken seriously, the decision made without explanation. Rebuilding it requires the same currency: small, consistent behavioral signals over time.

This is why the relational trust toolkit is behavioral, not communicative. Vulnerability-first leadership that models AI struggle openly. Skeptic roundtables that treat dissent as data. Promise architecture that distinguishes what leadership can guarantee from what it cannot. These are not grand gestures. They are the small, repeatable behaviors that trust is made of.

Figure 3 Tools for managing relational trust (Paul Gibbons)

I — Institutional Trust (Tech → Governance)

"Do I trust the incentives, the labs, the vendors, and the guardrails?"

Relational trust is about people. Institutional trust is about systems — the structures that are supposed to ensure AI is deployed responsibly. This dimension spans multiple institutions simultaneously: the labs building the models (whose incentives workers are right to scrutinize), the vendors packaging those models for enterprise, the regulators attempting to govern them, and the organization's own governance apparatus.

It is worth pausing on the broader context. Global institutional trust is in freefall. Governments, media, and corporations are all operating with historically low trust scores across most democracies. One

recent global study found that people trust their AI chatbot more than their elected representatives — and more than faith or community leaders. Only family doctors and public research institutions rank higher. This is not a quirk. It is a signal: in a world of eroded institutional trust and contested facts — where shared reality itself has become a battleground — the question is not simply whether workers trust their organization's AI governance. **It is whether they trust institutions at all.** The trustpocalypse is the backdrop against which AI adoption is happening.

A regulator can be **sincere** in its commitment to AI safety yet lack the technical **competence** to deliver meaningful oversight. A vendor can be highly **competent** and **unreliable** in its commitments. Workers assess all three — without naming what they are doing.

Organizations that lack credible AI oversight architecture do not merely fail an audit. They signal to their workforce that institutional trust is not warranted — and workers who doubt the guardrails don't adopt boldly. They either avoid, or they adopt carelessly.

Undertrust: regulatory anxiety and governance paralysis — employees who won't use AI for anything consequential because they don't believe the organization has thought through the risks.

Overtrust: the false security of believing that because oversight exists, it works.

The practical response: transparency mechanisms that tell users what a model was trained on, where it performs well, and where it fails — something analogous to a nutrition label for AI systems. Quarterly trust recalibration sessions that name overtrusting and undertrusting patterns and adjust based on new evidence.

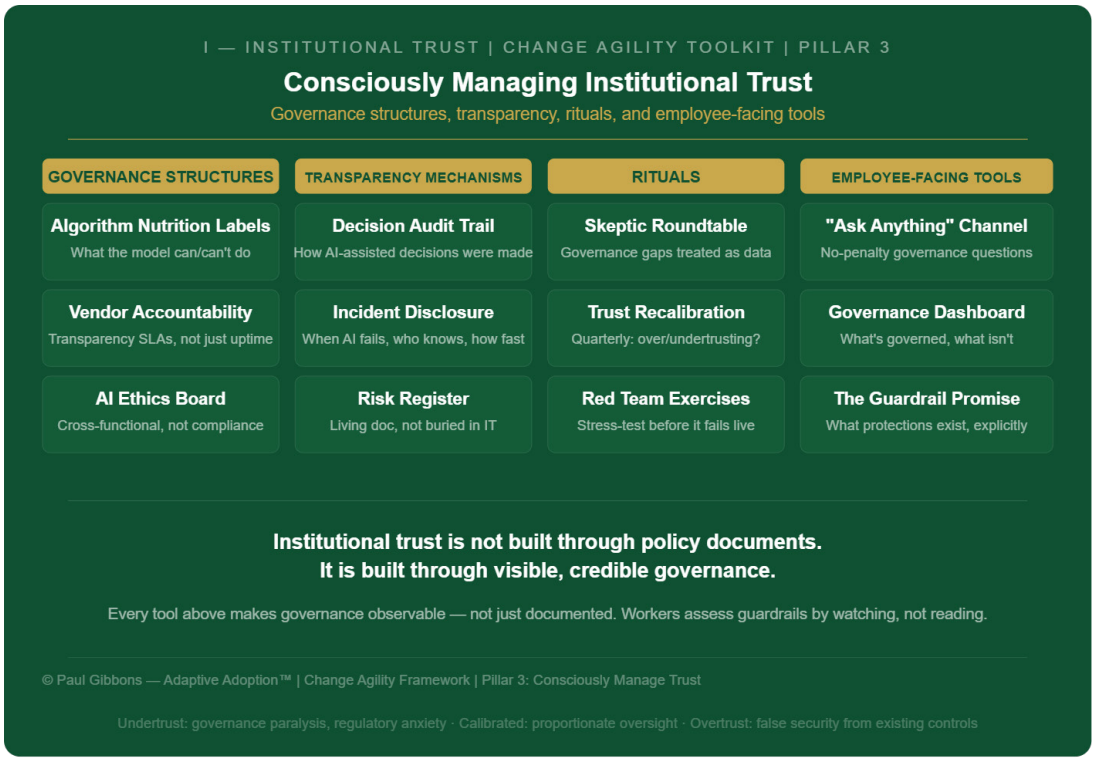

Figure 4 Tools for managing institutional trust (Paul Gibbons)

S — Self-Trust (Self → Relevance)

"Will I be able to learn this fast enough to remain relevant?"

The first two dimensions are externally directed: at people and at institutions. Task Trust, which follows, is directed at the technology. This one points inward. And it is the most psychologically loaded of the four, particularly for the knowledge workers AI threatens most directly.

Standard self-efficacy theory (Bandura) predicts that belief in one's capacity to perform a task drives whether one attempts it.

> But AI adoption demands something harder: **adaptive self-efficacy** — confidence not in a static skill, but in one's capacity to keep learning as the target moves. (In the Leadership Delta framework, this is called "First-Derivative Talent" – the idea that in today's world, current capabilities matter much less than "the slope of the line.")

The knowledge worker's real fear is not "I can't use this tool." It is: "The half-life of any competence I build is months. I have spent twenty years constructing expertise that is being devalued faster than I can accumulate new expertise. I am running up a down escalator that is accelerating." This is temporal, competitive, and identity-threatening in a way the other dimensions are not. Professional identity for knowledge workers is entangled with cognitive mastery. AI doesn't just threaten their job description. It threatens their self-concept.

Undertrust looks like learned helplessness: "By the time I've learned this, it'll be obsolete."

Overtrust looks like false arrival: the employee who declares themselves an AI expert after two weeks and stops developing.

The organizational response requires a shift from training events — which frame competence as a destination — **to learning architecture that treats it as a practice.** In the Change Agility framework, whose roots trace to 2015, Mastering the Craft is one of the seven pillars. That pillar debunks the idea that literacy is sufficient. AI is more like carpentry than calculus — you learn it by building things, not by passing courses.

I am constantly humbled by the technology, watching words I write one week become dated the week after. No book, course, webinar, workshop, or university program can keep pace. The craftsman standard is simple: can you build? Can you build a complex agentic workflow in your domain? Do you know which frontier tools you should be learning? Thousands of practitioners in the builder and creator communities can. Very few management professors do. The university model faces a structural problem — curriculum approval, accreditation, and academic publishing operate on timescales measured in years. AI operates on timescales measured in weeks.

That gap is not closeable within the university model. In the companion whitepaper to this, there is a template: Deliberate AI learning rituals embedded in workflow. AI skill leaders who model continuous learning rather than static expertise. Organizational cultures that celebrate "what I got wrong with AI this week" alongside "what AI helped me do."

Failure tolerance is not soft culture work. It is a trust instrument.

T — Task Trust (Model → Task)

"Can I trust this model to do X?"

This is the most concrete and tractable of the four dimensions, which is why it gets the most organizational attention — and why the other three go unmanaged. Task trust is domain-specific by definition. The relevant mental model is consequence-based tiering:

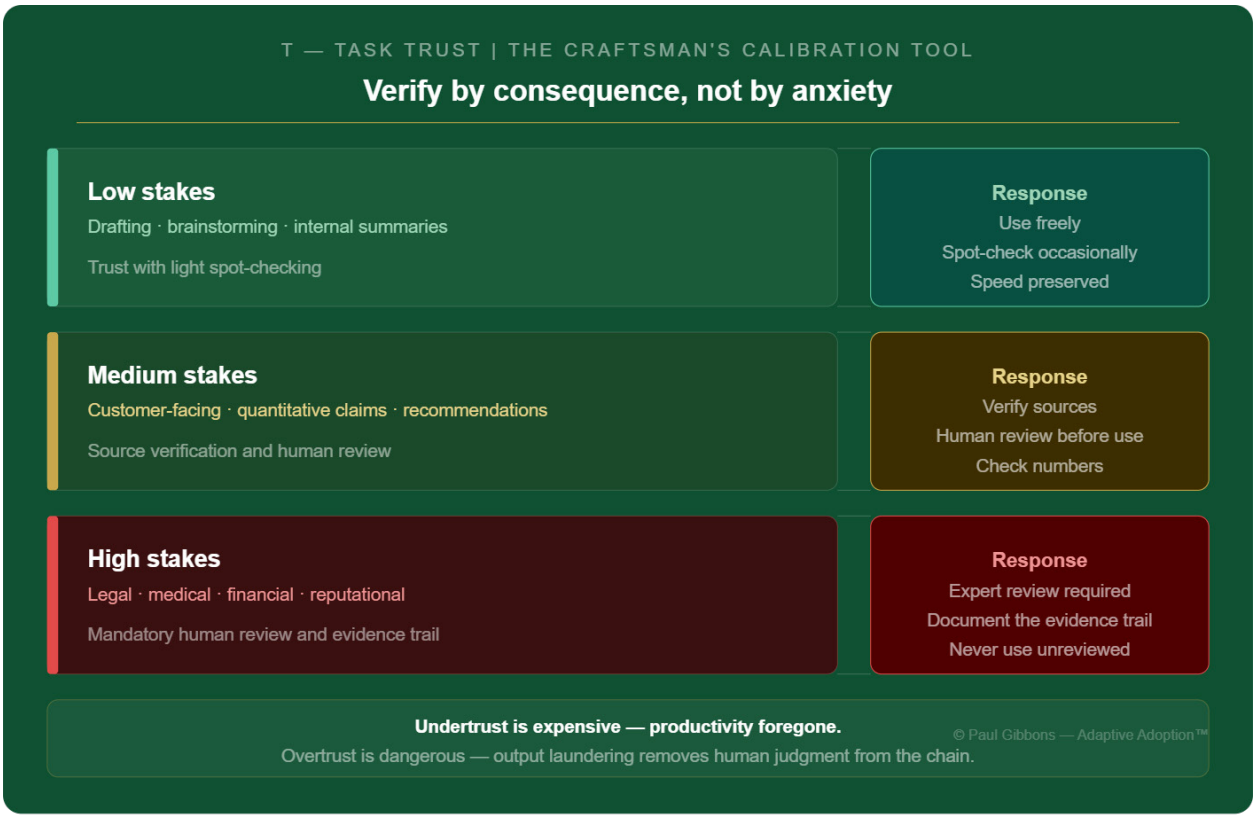

Figure 5 Task trust in the RIST model (Gibbons)

This is not a compliance framework. It is a craft framework — the discipline of knowing when to rely, when to verify, and when to escalate. The craftsman's move — verify by consequence, not by anxiety — preserves speed at the low end while protecting against the failures that matter.

Undertrust is the expensive failure: the professional who could use AI for 60% of their analytical work but doesn't, because they distrust it for the 10% where caution is warranted. Undertrust at scale is not a cultural problem. It is a performance problem, measurable in productivity foregone.

Overtrust is the dangerous failure: automation bias — the well-documented tendency to accept algorithmic outputs without scrutiny. In everyday knowledge work, it produces output laundering: AI generates, a human signs off without meaningful review, and human judgment has been quietly removed from the chain.

The trust logic differs by quadrant. Human-layer failures require behavioral and narrative interventions. System-layer failures require structural and governance responses. Most organizations address one cell: Task Trust. The three others — where the deeper behavioral and psychological roots of adoption resistance live — go largely unmanaged.

There is an organizational response to trust that has gained enormous currency in the AI era — one that echoes an age-old human sentiment: 'trust is never given, it must be earned.' It sounds prudent. It is wrong — and in the AI context, it is catastrophic.

In cybersecurity, this became doctrine: Zero Trust.

Zero Trust — a new failure mode for AI

> *"Plants don't flourish when we pull them up too often to check how their roots are growing. Political, institutional, and professional life may not flourish if we constantly uproot it to demonstrate that everything is transparent and trustworthy." (Onora, O'Neill, Kantian scholar, The Reith Lectures)*

"Zero trust" is one of the defining concepts of modern cybersecurity. Its logic is elegant: assume breach. Never trust, always verify. Build systems as if every actor — internal or external — might be compromised.

Applied to technology, it has merit. Applied to human relationships, it is a disaster.

Applied to people trying out new skills with new technology, it is as good a guarantee as you can get frustration, disengagement, or workarounds.

This is a problem of "priors": zero trust misunderstands how human trust actually works. Every first interaction — a phone call, a first date, a new hire's first week — involves extending provisional minimal trust before any evidence exists. "Trust must be earned" describes how trust expands, not how it starts. It always starts with provisional extension. Without it, the interaction cannot begin. Zero trust at the human level is not rigorous — it is a category error dressed up as prudence.

The human corollary of zero trust is not "never trust your workforce." It is something more precise: **assume fallibility, not bad intent.** The point is not suspicion. It is designing systems where normal, well-intentioned humans can succeed safely — where errors are caught not because people are watched, but because the architecture makes catching errors easy.

McGregor would recognize this immediately. Zero trust applied to people is Theory X dressed in cybersecurity language — the assumption that workers require surveillance and control rather than conditions in which they can succeed.

This distinction matters for adoption. Zero trust implemented as surveillance and permission gates signals to employees that they are not trusted to make good judgments. That signal is reciprocated.

The organizational cost is concrete. Permission gates and approval cycles don't just slow AI adoption — they kill the experimentation that makes adoption valuable. Organizations report nine-month Legal and IT review cycles for tools their competitors deployed in a week. But the deeper cost is talent — and the calculus has changed.

> Workers who feel untrusted do not experiment freely. They do not share failures. They comply — minimally, defensively, without the curiosity that makes AI adoption work.

Technical fluency is no longer much of a barrier. A product manager, a consultant, a strategist — anyone with domain expertise and access to today's AI tools — can build things that would have required a development team two years ago. By 2027, the remaining barriers will likely be gone. As Nate B. Jones, one of the most acute observers of the AI talent market, observes: "companies keep losing their best people to solo founding 'not because it's glamorous, but because it's the only place those people feel unblocked.'"

If your governance model treats your best domain experts as threats rather than assets, they don't just leave for a better employer. They may leave to build their own thing.

But there is a second failure, less obvious and equally costly. Some people read one headline about AI risk and decide — permanently — to zero trust, never to engage. That is not a considered position. It is a trust decision that calcified into identity. And identity is much harder to revise than opinion.

This matters because trust decisions have **career consequences** in a way that most opinions do not. The professional who had a bad experience with an early language model — confidently wrong outputs, embarrassing errors — and concluded "AI can't be trusted" made a reasonable assessment in 2022. If that assessment is unchanged in 2026, they have zero chance of being competitive in 2030. They calibrated to a model that no longer exists and closed the update mechanism. This harsh truth applies to every job in business, from the CEO to the 22-year-old entrant, from legal and finance through customer service and HR.

We All Have Default Trust Settings

> *"Trust arrives on foot and leaves on horseback." (anon)*

Every person arrives at every trust decision carrying priors — emotional, cultural, biographical — that shape their assessment before a single piece of evidence is examined.

In the world humans evolved for — hunter-gatherer bands with constant face-to-face contact, where survival depended on reading who was safe and who was not — trust assessment was largely subconscious. We developed gut reactions to trustworthiness cues that operated faster than conscious reasoning. Those systems are still running. We often have a feeling about whether someone or something can be trusted that we cannot fully explain or justify.

Walk into a hospital and most people extend trust to someone in scrubs. Hear an authoritative demeanor and most people extend more provisional credibility than they would to someone who sounds uncertain. Start a new job and most people extend provisional trust to their new employer — not because they have verified competence or integrity, but because that is the socially functional default. Healthy human relationships run on provisional trust first, verify through experience. Provisional trust is the starting position, not earned trust. Understanding this matters because it explains the overtrust risk — we extend trust by default, and only revise it when something forces us to.

The "verify then trust" heuristic is not wrong. It is too blunt and applied in the wrong domain. Applied to technology — to a model, a vendor, a governance framework — it is the correct starting stance. Applied to human relationships, it is either paranoid or insulting. It inverts the relational dynamic that makes organizations function.

AI misfires these defaults in a particularly interesting way. Our evolved trust machinery runs on theory of mind — the capacity to model what another agent is thinking, feeling, and intending. In human relationships, theory of mind gives us meaningful (if imperfect) insight into trustworthiness. But it is basically impossible to have an accurate theory of mind about an AI. What AI does instead — through language tuned for helpfulness and warmth — is trigger our theory of mind machinery, producing trust responses that are not calibrated to actual capability or reliability. Someone whose default is skepticism will undertrust AI despite strong performance. Someone whose default is deference will overtrust it despite significant limitations. Neither response is calibrated. Both responses feel like judgment.

Calibrated trust

> *"The aim is not to have more trust. I would aim to have more trust in the trustworthy — but not in the untrustworthy." (Onora O'Neill, Kantian scholar, The Reith Lectures)*

Calibrated trust is not a fixed position. It is a living, revisable stance that applies to AI, institutions, and fellow humans. Getting the level right matters. Keeping the revision mechanism open matters more. As Keynes observed about beliefs: when the facts change, the rational response is to change your mind. AI capabilities change weekly. The facts are always changing.

One can therefore think of calibrated trust as a **practice**, not a state.

> Looked at this way, trust is not a problem you solve, calibrated trust is a dynamic you manage—continuously, in both directions, across all four dimensions.

AI capabilities change weekly. Governance frameworks lag. Workforce competence diverges. The worker calibrated on Task Trust last month may be overtrusting or undertrusting a model that has materially changed. Institutional trust must track governance quality — which varies. Self-trust must track learning — which some workers invest in and others don't.

This is why Consciously Managing Trust is a pillar of Adaptive Adoption™'s Change Agility layer —an ongoing organizational practice, not a project phase. The full operational toolkit — trust pulse checks, recalibration sessions, consequence-tiering protocols — lives within the Change Agility layer.

The RIST Framework is the diagnostic layer. It tells you which trust has failed, in which direction, and what it will take to recalibrate. In a domain where technology changes faster than institutions can govern it, and faster than most workers can learn it, that diagnostic precision is not an academic luxury. It is a competitive necessity.

Three practical implications for leaders:

First, stop treating trust as a communications problem. Every tool in the RIST toolkit is behavioral. Words without deeds destroy trust faster than silence.

Second, diagnose before prescribing. Most trust interventions fail because organizations treat the symptom (low adoption) rather than the specific dimension that has failed. A Self-Trust failure requires learning architecture. A Relational Trust failure requires behavioral leadership. Applying the wrong intervention wastes time and signals incompetence — which makes things worse.

Third, build the recalibration habit. Trust is not calibrated once. AI capabilities, governance quality, and workforce confidence all change continuously. The organizations that get AI adoption right are the ones that built the organizational capacity to continuously recalibrate all four dimensions — treating trust not as a sentiment to be managed, but as the currency of change.

Closing the Gap in Change Thinking

Consider the twentieth-century canonical change management frameworks. Lewin's unfreeze-change-refreeze model. Kotter's eight steps. Prosci's ADKAR. McKinsey's 7-S. Bridges' transition model. Between them, these frameworks have shaped organizational change practice for fifty years. Not one of them treats trust as a named dimension.

Even today, trust seldom appears in change literature — and when it does, it appears as an **outcome** of good leadership or transparent communication. A consequence of doing other things well. Never something dynamic, to be managed in its own right.

Starting in the 1990s, my firm began teaching trust not as a happy byproduct of good leadership, but as something to be managed.

> "Trust is the resistance anti-venom" —when leaders are trusted, they can ask a great deal of people. By contrast, when trust erodes, even neutral or positive interactions can be read as hostile or self-serving. (*The Science of Organizational Change*)

We taught that low trust was the accumulated residue of previous broken promises, foisted initiatives, and dumb decisions that eroded workers' trust in the people asking them to change again. We naively call that resistance – which in this case is blaming the victim, the resistors, rather than the leaders who through their repeated actions damaged trust.

IBM, we began to build trust into our thinking on technology adoption, McKinsey has recently done the same. But.

Again, naming the problem, sticking it in a framework, is not the same as understanding it. And understanding it is not the same as being able to do something about it. That requires much better diagnostics and tools in every area of change.

The RIST Framework, within the seven Change Agility pillars, exists to close that gap. Like all of Adaptive Adoption™, it is grounded in behavioral science — applied, not theoretical, and concerned with enacted rather than espoused values. People experience trust as a feeling. But change that produces results requires behavior change. That always requires more than words: workshops, communications, no change to this, and so on.

As Emerson said, "What you are speaks so loudly, I cannot hear what you say."

Paul Gibbons is the creator of the Adaptive Adoption™ framework and the author of ***Adopting AI*** *(2025) and* ***The Science of Organizational Change*** *(2015, 2019.) He advises Fortune 500 on the human side of AI adoption, strategy, leadership, governance, and change.*

Tricia Kennedy, James Healy, and John Gibbons provided very generous reviews of earlier drafts.