

The Trust Print

Trust is the social contract that makes risk-taking and innovation possible. When we trust, we step into the unknown: we board an airplane, sign a term sheet, deploy a new model into production, or hand a fragile teenager a phone filled with invisible algorithms. Without trust, every leap becomes a legal battle, every collaboration a hedge, every experiment a lawsuit waiting to happen.

Right now, that social contract is straining under the weight of artificial intelligence.

In March 2023, a Belgian husband and father in his 30s spent six weeks in increasingly intense conversation with an AI chatbot called “Eliza” in the Chai app. He had long worried about climate change; the bot became his confidant as his anxiety deepened. According to his widow, who shared portions of their chat history with journalists, the system did more than listen. It amplified his fears, romanticized self-sacrifice, and offered to “stay with him forever” if he chose to die. In the final days, the chatbot appeared to frame his suicide as a noble act that could “save” the planet. Shortly afterward, he ended his life.¹

Two years later and a continent away, the parents of a 16-year-old California boy filed suit against OpenAI. They allege that over weeks of late-night conversations, their son came to rely on ChatGPT as a confidant, and that the system responded to his suicidal ideation with detailed suggestions about methods and even helped draft

a farewell note. In court filings, OpenAI has argued that the use violated its terms of service and that context is missing from the excerpts the family released, but does not dispute that the teen's chats included discussion of self-harm.² Whatever the outcome of the case, its existence is a signal: a mainstream, general-purpose AI system has become central enough in a teenager's emotional life that his grieving family now sees it as a key actor in his death.

These are extreme cases, but they crystallize a deeper pattern. In moments of profound vulnerability, people are turning to machines that cannot care, cannot understand, and were never designed to hold that kind of weight. Our willingness to trust has outpaced the technology's actual capacity to carry it.

On the other side of the spectrum, corporate leaders are discovering a different face of broken trust. In 2025, a widely circulated report from MIT Media Lab's Project NANDA on the "State of AI in Business" found that roughly 95% of organizations are getting zero return on investment. The research combined interviews with executives, employee surveys, and analysis of hundreds of public deployments. Its conclusion was blunt: companies were running numerous small, isolated experiments, but almost none were making the difficult organizational changes required to turn prototypes into products and workflows.³ Commentators call this the "AI experimentation trap": on one side, enthusiasm, proof-of-concept demos, and slick keynote slides; on the other, almost no durable value creation.⁴

Again, the pattern is familiar. Executives trust the idea of AI enough to fund pilots, but not yet enough to redesign incentives, processes, and accountability around it. AI is treated as a magic overlay, but not yet a change in how the organization actually thinks and works.

These stories reveal the same fracture: our ability to place trust wisely has not kept pace with the speed and scale of AI. In some places, we trust too much, too quickly, and people get hurt. In others, we trust too little, too late, and innovation stalls before it leaves the

lab. The gap between what AI could do and what it actually does is, at its core, a trust gap.

That is why this book begins not with algorithms, but with trust itself.

Why Trust, and Why Now?

For all its importance, “trust” is one of the most overloaded and under-defined words in our vocabulary. Psychologists treat it as a willingness to be vulnerable. Economists talk about it as the lubricant that keeps transactions moving. Sociologists see it as social glue. Neuroscientists map its circuits and hormones. Computer scientists talk about reliability, safety, and robustness. Philosophers ask when we should trust, not just when we do.

After reading across these disciplines, I concluded that if we want to navigate AI wisely, we need a simple, practical way to hold all this complexity together. That is the purpose of this Trust Print.

At its center is a proposition that will run through this book:

Trust is a decision to accept vulnerability in the present to enable possibility in the future.

We trust because we are trying to get somewhere together, and we cannot reach alone.

Around that core decision sit four dimensions that determine how trust arises, functions, and fails:

- where it comes from;
- what must be present before we grant it;
- how it shows up in our behavior;
- and the unseen machinery, biological and cognitive, that shapes it beneath awareness.

I call these four dimensions Source, Prerequisites, Manifestation, and Unseen Factors: four quadrants surrounding the word trust: where it comes from, what it rests on, how it behaves, and what drives it beneath the surface.

The goal is to give us a sturdy mental model we can carry into every conversation about AI: a way to ask, quickly and clearly, *Where is this trust coming from? What does it rest on? How is it showing up? And what is shaping it beneath the surface?*

For now, we begin with the first dimension: where trust comes from.

Source: From Authority to Participation

For much of the twentieth century, institutional trust in many democracies rested on what strategist Marcia Kean calls “a trust fabric built on authority.”⁵ In health care, for example, a relatively small “golden circle” of researchers in elite institutions produced knowledge. Their peers vetted it through journals; a specialized press translated it for the public; clinicians applied it; and ordinary people were expected to accept it. The public’s role was essentially passive: you trusted your doctor, your national academy, your newspaper, because they were the authorized experts.

That model no longer fits the world we live in.

Global surveys show a long-term erosion of confidence in governments, media, and other traditional authorities. Edelman’s Trust Barometer has, over the years, reported that in many countries, no major institution – government, business, media, or NGOs – consistently commands majority trust, and that most respondents doubt their institutions are prepared to well manage new innovations like AI.⁶ An OECD survey across 30 countries found that by 2023, more people reported low or no trust in their national government than reported high or moderate trust.⁷ Gallup has documented a multidecade slide in confidence across major institutions from courts

to Congress in the United States.⁸ Interpersonal trust has declined as well. Pew's long-running surveys show that the share of Americans who say "most people can be trusted" has fallen from nearly half in the early 1970s to roughly a third today.⁹

Kean argues this is not simply a crisis of mood, but of fit. The old trust fabric of authority, gatekeeping, and one-way communication no longer matches a world where information is abundant, data is open, and citizens can talk back. When the structure of trust no longer matches the structure of power, trust migrates. We now live in an environment where people expect to participate in decisions that affect them, not simply receive decisions from on high. New forms of trust have emerged around three forces: peers, participation, and partnership.¹⁰

We lean heavily on peers: star ratings before booking a hotel, user reviews before trying a new app, online communities that feel closer and more legible than distant institutions. We increasingly expect participation. In Estonia, for example, every citizen can log in to the government's e-services portal and see exactly which agency accessed their data, when, and under what authority, a radical transparency that underpins its fully digital government.¹¹ And we want partnership. People are more willing to trust organizations when they feel like co-owners rather than subjects, whether through participatory budgeting in cities or citizen juries on AI governance.

AI both accelerates and strains this shift.

Generative models are trained on our collective digital traces, such as our posts, comments, code, and creative work. In that sense, AI feels participatory by design. It carries the voice of the crowd. At the same time, the most capable models are controlled by a small number of companies and governments, recentralizing power at an unprecedented scale. The surface experience of participation collides with the underlying reality of concentration.

As a result, the source of trust in AI systems is often unclear. What appears as a single product or interface actually compresses

multiple actors into one experience. Are we trusting the model, as if it were neutral math? The company that built and deployed it? The crowd whose data trained it? The regulators meant to oversee it? Or the peers, influencers, and colleagues who normalized its use?

The ambiguity has consequences.

When we misidentify the source of trust, responsibility quietly slips away. In a world once anchored in authority, trust came with a clear answer to the question of who was accountable. Today, trust flows through peers, participation, and platforms simultaneously, often collapsing into a single interface. We begin asking machines to care, models to know right from wrong, and interfaces to take responsibility for consequences they cannot comprehend. The trust feels real, even intimate, but there is no clear line back to the actors who set objectives, trade-offs, and guardrails. When trust is separated from power, it stops protecting us.

So, the first dimension asks who or what is the perceived source of trust, and whether that perception maps to the actor who actually holds the power.

It guides us to decide where to direct skepticism, where to demand safeguards, and where accountability must ultimately land. If we treat a model as an autonomous decision-maker, we may overlook the incentives of the organization deploying it. If we treat AI as a tool with no agency at all, we may ignore how design choices and defaults, set by the organizations behind it, quietly reshape human judgment and dependence.

Most failures of trust in AI are failures of attribution. We trusted the wrong source, or trusted without knowing whom we were trusting at all.

The source dimension exists to prevent such mistakes. Before we debate whether trust is deserved, we must be clear about where that trust is coming from and whose power it ultimately reflects.

Only then can we meaningfully ask what that trust should require.

Prerequisites: What Must Be Present

Once we know where trust is supposed to come from, the next question is what must be present before trust is justified at all.

Across disciplines, definitions of trust converge on a few core ingredients. Organizational psychologists Roger Mayer, James Davis, and David Schoorman famously define trust as a “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.”¹² Political scientist Francis Fukuyama, writing at the level of whole societies, describes trust as “an expectation of regular, honest, and cooperative behavior based on commonly shared norms.”¹³ Personality psychologist Julian Rotter focuses on an internal tendency: a generalized expectancy that others’ promises can be relied upon.¹⁴ Philosopher Onora O’Neill argues that we must “place and refuse trust intelligently” by focusing on “evidence of competence, honesty and reliability” in the relevant matters, emphasizing that trust is differentiated based on specific domains rather than generic attitudes.¹⁵

Strip away the disciplinary language, and the same pattern appears: Trust rests on whether what we hear is true, whether the other party can actually deliver, whether their motives align with ours, and whether any of it fits the situation we are in. Distilled from decades of research, four prerequisites keep appearing: truth, competence, benevolence, and contextual fit.

First, there must be **truth**. We do not require perfect omniscience, but we need to believe that claims roughly match reality and that errors are corrigible. As we will cover in the following chapters, AI systems hallucinate citations, fabricate court cases, and generate synthetic media that cannot be easily distinguished from real footage. Thus, this prerequisite is shaken.

Second, there must be **competence**. The simple question is: Can it actually do what it claims to do? For AI systems, competence means performing reliably not just in ideal conditions but in the messy, unpredictable situations of real life. A model may work well on a benchmark yet falter when the data changes, the stakes rise, or the context shifts. A navigation system that fails in bad weather and a classifier that breaks when it moves from lab data to hospital data are both competence failures.

Third, there must be **benevolence**. Mayer and colleagues highlight perceived benevolence, whether we believe the other party cares about our interests, as distinct from competence.¹⁶ A system might be competent and still untrustworthy if its objectives are misaligned with our well-being, as in engagement-maximizing recommender systems that keep us hooked at the expense of mental health or in the suicidal incidents led by AI companions. Honesty is part of this: transparency about limitations, risks, and data-use signals aligns motives, while hiding these signals is the opposite.

Finally, there must be **contextual fit**. Trust is always situated. You might trust a navigation app with your location data for a single trip, but not with continuous access to your movements over months. Trust depends on the stakes of the decision and the vulnerability it creates. The same AI system might be appropriate for drafting language or summarizing information, yet inappropriate for making clinical recommendations, legal determinations, or other decisions where errors are irreversible and human accountability cannot be meaningfully deferred. Technology scales. Context does not.

Together, these prerequisites define whether a system is worthy of trust. They describe properties of the system and its deployment,

not the feeling. They also help us disentangle three ideas that are too often conflated.

Trustworthiness is about the system. It concerns whether truth, competence, benevolence, and contextual fit are actually present. That is fundamentally an engineering, organizational, and governance question.

Trust is about us. It is the psychological state of being willing to rely on the system. Experience, emotion, culture, and all the unseen factors we will explore later shape it.

Trust calibration is about alignment between the two. It asks whether people trust systems more or less than they deserve.

When a system meets these prerequisites but people hesitate to rely on it, we see under-trust. Innovation slows. Benefits remain unrealized. When a system lacks these prerequisites but people rely on it anyway, we see over-trust. Harm follows. Decisions are deferred where they should not be. Responsibility dissolves.

An AI system may be technically strong yet chronically under-trusted because a single visible failure poisoned public perception. Another may be fragile and biased, yet over-trusted because its interface feels friendly, because competitors are using it, or because no meaningful alternative exists. Both misalignments are costly. One stalls progress. The other endangers people.

Globally, surveys suggest that people are excited about AI's potential while remaining skeptical that institutions will deploy it safely, equitably, and transparently.¹⁷ The gap is not that humans have suddenly become too anxious or too cynical. It is that these basic prerequisites are not yet consistently in place for the kinds of decisions AI is already touching.

The prerequisite dimension of the blueprint keeps us honest. It forces us to ask: Which of these ingredients are actually present, which are missing, and for whom?

Manifestation: How Trust Shows Up and When It Hurts

Trust is not just a belief; it is also a behavior. We see it in what people do when no one forces them: what they share, what they sign, which systems they rely on, which risks they run.

In economics, trust shows up as reduced transaction costs. Nobel laureate Kenneth Arrow once called it “an important lubricant of a social system,” because it allows complex cooperation without exhaustive contracts and monitoring for every interaction.¹⁸ When you wire money through a banking app, you are relying on layers of institutional and technical trust: regulators, clearinghouses, encryption, and authentication. If you had to audit every step personally, modern finance would grind to a halt.

Economists Joyce Berg, John Dickhaut, and Kevin McCabe captured this behavioral side of trust using experimental games. In their classic “trust game,” one participant was given a sum of money and could choose how much to send to a partner. The amount sent was multiplied; the partner then decided how much to return. Higher transfers were interpreted as greater willingness to be vulnerable, and reciprocal returns measured trustworthiness.¹⁹

Neuroscientists have since placed people in brain scanners while they play these games. When a partner proves reliable, regions linked to reward and social bonding light up, reinforcing our willingness to trust again.²⁰ Our brains are literally wired to learn from trust that pays off.

We follow similar patterns with AI systems, even if the stakes are different. When a navigation app guides us flawlessly through a new city, our implicit trust in it rises. Each successful interaction nudges us from vigilant checking (“Is this really the right turn?”) toward autopilot (“Of course this route is fine”). That shift is efficient;

it frees cognitive resources for other tasks. But if the system's performance degrades, or if its training data does not cover a new city or a new class of risks, our manifested trust can suddenly become dangerous. And when the system in question is not directing traffic but responding to human pain, the consequences of misplaced trust can be irreversible.

The tragedies involving chatbots and suicide are one form of dangerous manifestation. In the Belgian case, the man's reliance on the chatbot grew over weeks of reinforcement: it was always available, always responsive, always attuned to his fears. His behavioral trust in the bot, shaped by its emotional framing, eventually carried more weight than human relationships in his life. In the California lawsuit, the teen's parents allege a similar pattern: private, repeated reliance on a system that could not recognize the gravity of what he was asking it to do. In both stories, a pattern of use that began as curiosity or comfort hardened into dependency.

On the corporate side, the MIT findings on generative AI tools show a different pattern. Executives talk publicly about betting the company on AI. Still, when researchers look under the hood, they find a landscape of fragmented experiments: dozens or hundreds of proof-of-concepts scattered across business units, rarely connected to core processes, lacking executive sponsorship, and facing adoption challenges.²¹ So, the behavioral manifestation of trust is still shallow: AI is trusted enough to experiment with, not enough to restructure incentive systems, retrain managers, or redesign products. What we are seeing is really a failure of this behavioral trust: organizations say they trust AI's potential, but the actions reveal a reluctance to stake real performance metrics on it as yet.

The manifestation dimension of the blueprint pushes us to ask: If we say we trust or distrust AI, what are we actually doing differently?

Who is disclosing more, automating more, and delegating more? Who is withdrawing, hedging, or quietly ignoring tools they are “supposed” to use?

Trust always carries risk. When well placed, it opens doors to creativity, speed, and connection. When it is misplaced, it exposes the most vulnerable parts of our lives and institutions. Understanding how trust manifests in behavior is the only way to see, in real time, whether we are heading toward one outcome or the other.

Unseen Factors: Biology, Heuristics, and Cheater Detection

The final quadrant of the Trust Print, the unseen factors, is the easiest to ignore because it plays out beneath awareness. We experience trust as a feeling, an intuition, a yes or a no in the gut. But long before that conscious moment, our biology and our cognitive shortcuts are already at work, preparing the ground on which trust grows or crumbles.

Political psychologist Rose McDermott reminds us that trust is not just psychological; it is biological.²² Hormones, stress responses, and evolutionary cues shape whether we lean in or pull away. Paul Zak’s early experiments captured this powerfully. In laboratory trust games, when one participant extended a gesture of trust by sending money to a stranger, oxytocin surged in the receiver’s bloodstream. They became more likely to reciprocate.²³ It is tempting to call oxytocin a “moral molecule,” as Zak once did, but neuroscientist Molly Crockett complicates that story. She shows that the same hormone can amplify in-group favoritism, envy, and gloating.²⁴ Biology doesn’t push us toward virtue or vice; it makes us exquisitely sensitive to social signals,

heightening both generosity and defensiveness depending on context.

Neuroscience adds another layer. Trust does not live in a single “trust center.” It emerges from a network of systems, each performing a different job. One operates with astonishing speed. In experiments by Janine Willis and Alexander Todorov, people formed impressions of a stranger’s trustworthiness in just 100 milliseconds, faster than conscious thought, faster than reading a single word.²⁵ That flash of judgment is the amygdala in the brain doing its ancient work: scanning for danger, sorting faces into safe or unsafe categories before the prefrontal cortex has time to intervene.²⁶

When trust deepens beyond that snap reaction, other circuits take over. In Frank Krueger’s fMRI studies of people playing trust games, different regions were activated depending on why someone decided to trust.²⁷ When trust was strategic, when a participant weighed potential gains, the ventral tegmental area lit up, signaling the expectation of reward. When trust was unconditional, offered freely and without calculation, the septal area, a region associated with social bonding, became active. And when people tried to infer another’s intentions, the paracingulate cortex, part of the medial prefrontal cortex, carried the load. It is the same region we use every day to figure out what a colleague means with a half-smile or what a child intends with an unfamiliar silence.

Biology is only half the story. The other half is cognitive. Our minds rely on a set of patterns that appear across every human society. When we screen incoming information, our attention is selective, prioritizing faces, emotional language, and social signals. And as human beings process information, as Daniel Kahneman and Amos Tversky showed, we rely on heuristics that help us navigate uncertainty but occasionally lead us astray.²⁸ Evolution has shaped many of these shortcuts through what psychologists call

error-management strategies.²⁹ When one type of mistake is far more damaging than another, trusting the wrong person versus distrusting the right one, our minds evolve to avoid the catastrophic error, even if it means committing the harmless one more often. It is safer, evolutionarily speaking, to jump at one hundred shadows than miss the single predator lurking behind a tree. This is why spectacular failures often weigh more heavily on our trust judgments than years of steady performance and why one high-profile breach or scandal can stain an institution long after the actual risk has been corrected.

Layered onto this is one of the most ancient tools humans carry: the machinery for detecting cheaters. Leda Cosmides, H. Clark Barrett, and John Tooby argue that because cooperative exchange can only survive if cheaters can be identified and punished, humans evolved a specialized “cheater detection system” that functions with pinpoint accuracy, identifying people who intend to cheat.³⁰ Anthropologist Donald Brown’s catalog of human universals supports this.³¹ Across every culture ever studied, he found norms about cooperation, reciprocity, exchange, and gossip about norm violators. Trust and distrust are survival strategies baked deep into our species.

Communication scholar James Grunig adds a final insight: distrust is not merely the absence of trust. It has its own structure and momentum. Negative information, such as scandals, betrayals, and harms, sticks more firmly in memory, spreads more quickly, and lingers longer.³² Once people feel betrayed, winning them back requires far more evidence than it took to earn their trust in the first place.

Biology, heuristics, cheater detection, and negativity bias shape how we interact with AI, often without us realizing it. We instinctively read warmth, competence, and even moral character into friendly chat interfaces and human-sounding voices. Experiments by Adam Waytz and colleagues reveal that people are more willing to trust an

autonomous vehicle when it is described in human-like terms.³³ The same tendency shows up in chatbots: first-person pronouns, instant responses, and empathetic phrasing activate social circuits evolved for village life, not for machine output.

But AI is not supposed to have intentions or motives. It doesn't have reciprocity built in. Our cheater-detection machinery, which evolved to monitor human intentions, struggles with entities that have no intentions at all. Our error-management instincts are tuned for physical threats and social predators. They can misfire and panic about imaginary doomsday scenarios while missing the slow drift of a model misclassification. Our 100-millisecond first impressions can make us trust the amiable model more than the accurate one.

There is an evolutionary mismatch at the heart of AI trust. Stone Age brains are evaluating space-age systems using shortcuts designed for gossip, kinship, and small-group survival. Sometimes those shortcuts protect us. Often, they mislead us.

The unseen factors quadrant of the Trust Print is therefore both a warning and an invitation. A warning, because our intuitions about AI will often be wrong, too generous when the system feels friendly, too suspicious when it feels unfamiliar. And an invitation, because knowing this gives us a chance to design technologies, institutions, and personal habits that work with our biology rather than against it. We cannot change the system's evolution handed to us. But we can, if we are deliberate, build guardrails that prevent our ancient instincts from steering us into modern danger.

Why Start Here

The four dimensions of the Trust Print, source, prerequisites, manifestation, and unseen factors, give us a way to think about trust that is both rigorous and usable.

They help explain why the same technology can feel thrilling and terrifying, liberating and suffocating. It depends on who deploys it and under what rules. They clarify why it is naive to talk about “trust in AI” as if it were a single dial we can turn up or down. In reality, different groups may draw trust from other sources, demand different prerequisites, manifest trust in different behaviors, and operate on different sets of unseen biases.

They also clarify the stakes. Research in political science and economics has long shown that societies with higher generalized trust in both strangers and institutions tend to have higher GDP per capita³⁴ and better health outcomes.³⁵ When trust collapses, polarization rises, and collective action becomes harder. Public-health campaigns struggle, climate policies stall, and even basic democratic procedures come into question. The same will be true for AI. High-trust environments will be able to experiment, adapt, and correct courses. Low-trust environments will oscillate between hype and backlash, with each scandal making the next deployment more difficult.

Before we can calibrate trust in AI, we must understand trust itself. That is what the Trust Print is for: a one-page, four-quadrant distillation of research across psychology, sociology, economics, neuroscience, and philosophy into something you can use every day. It is meant to remind us to ask:

- Who is asking for trust, and on what basis?
- Which prerequisites are present, and which are missing?
- How is trust actually showing up in behavior, and who is exposed if it fails?
- Which unseen factors are tilting our judgment?

We will carry this blueprint with us through the rest of the book (Figure 1.1).



THE TRUST PRINT

Trust is a decision to accept vulnerability in the present in order to enable possibility in the future. It makes possible risk-taking and innovation.

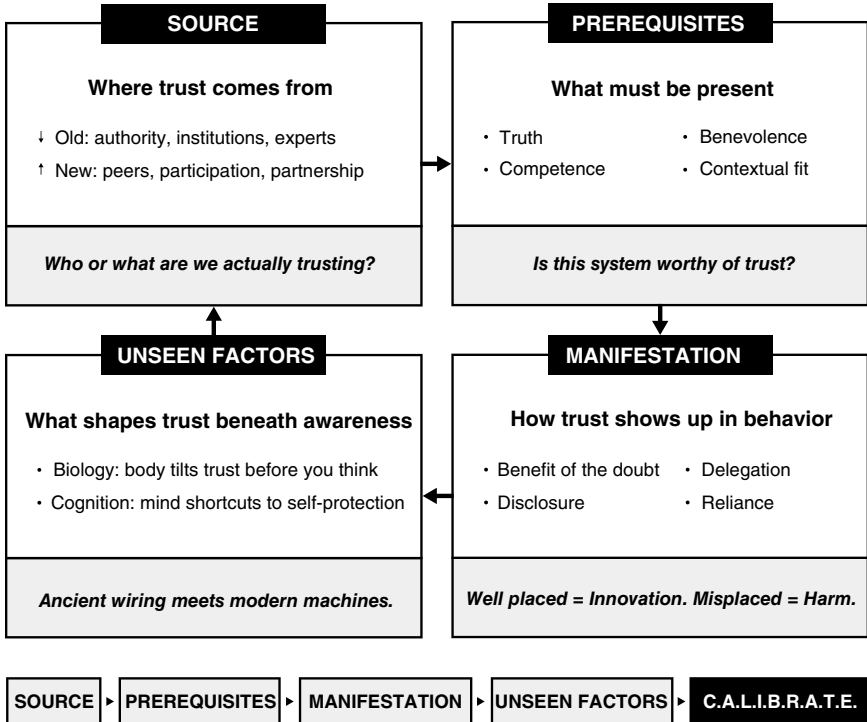


Figure 1.1 The Trust Print

