# THE PSYCHOLOGY OF TRUST IN AI

A Five-Part Guide to Rapid and Responsible Adoption

**Dr. Payal Loma Ghatnekar**

Website: http://payalloma.com/

*This series explores what it really means to trust an AI system, drawing on psychology, design research, and governance practice to build a framework that is both practically useful and intellectually rigorous. It is written for leaders, practitioners, and policymakers who are navigating the reality of AI adoption in complex organisations.*



PART 1

## What Does It Mean to Trust AI?

*How trust operates in complex socio-technical systems*

Trust is a deceptively simple word. We use it easily in human relationships, yet the moment we apply it to AI, the concept becomes layered, psychological, and structurally complex. It is no longer just an interpersonal intuition — it becomes a cognitive strategy for acting under uncertainty, a design outcome, and a governance challenge.

To understand what it means to trust an AI system, we need to start with its psychological foundations. At its core, trust is a willingness to be vulnerable. It is a psychological state in which a person chooses to accept uncertainty because they believe the other party — whether human or machine — will act in ways that align with their goals. That belief matters more than algorithmic mechanics.

In daily life, trust allows us to move through a world of imperfect information. In socio-technical systems, it is the mechanism that lets humans delegate tasks to machines that operate through internal processes which are opaque, probabilistic, and constantly evolving. Without trust, AI adoption stalls. With misplaced trust, it becomes dangerous. Understanding the psychology underneath is not optional — it is the foundation of responsible deployment.

### 1.1 Trust vs. Reliance: A Critical Distinction

One of the most persistent misunderstandings in AI adoption is the assumption that usage equals trust. In reality, people can rely on technology without trusting it, conflate the two creates significant blind spots.

Reliance is a behaviour. You follow an AI's suggestion because you must — due to limited alternatives, workload pressure, or institutional expectation. Trust is an attitude. You believe the AI will help you achieve your goals, and you accept vulnerability because you expect alignment between the system's outputs and your interests.

A clinician who uses an AI diagnostic tool because policy requires it is relying on the technology, not necessarily trusting it. A commuter who accepts a navigation route because it usually gets them there fastest is demonstrating trust. When we conflate these two things, we risk misreading compliance as confidence and adoption as endorsement. The danger is real: compliance without trust is fragile. The moment something goes wrong, the behaviour collapses. Trust, by contrast, is resilient — it supports long-term integration, but only when it is justified.

## 1.2 Trust vs. Trustworthiness

A second distinction is equally important. Trustworthiness is an attribute of the system itself — its objective properties of competence, safety, fairness, robustness, and privacy. Trust is the user's subjective perception — their confidence, expectations, and willingness to rely. These two do not automatically align.

A trustworthy system can fail to earn trust if users lack understanding, fear automation, or distrust the organisation deploying it. And an untrustworthy system — one that is biased or poorly tested — can gain high trust if it appears confident, human-like, or widely endorsed. This asymmetry is exactly why trust must be designed and governed, not merely hoped for.

## 1.3 Do We Trust AI Like We Trust Humans?

Traditional theories of interpersonal trust emphasise three dimensions: ability (competence), integrity (honesty and consistency), and benevolence (good intentions). AI has no motives, yet people ascribe intentions based on design cues, communication style, and institutional context. A conversational AI that uses warm, empathic language can be perceived as caring. A system that emphasises privacy can generate a sense of benevolence. A model that gives consistent outputs feels honest.

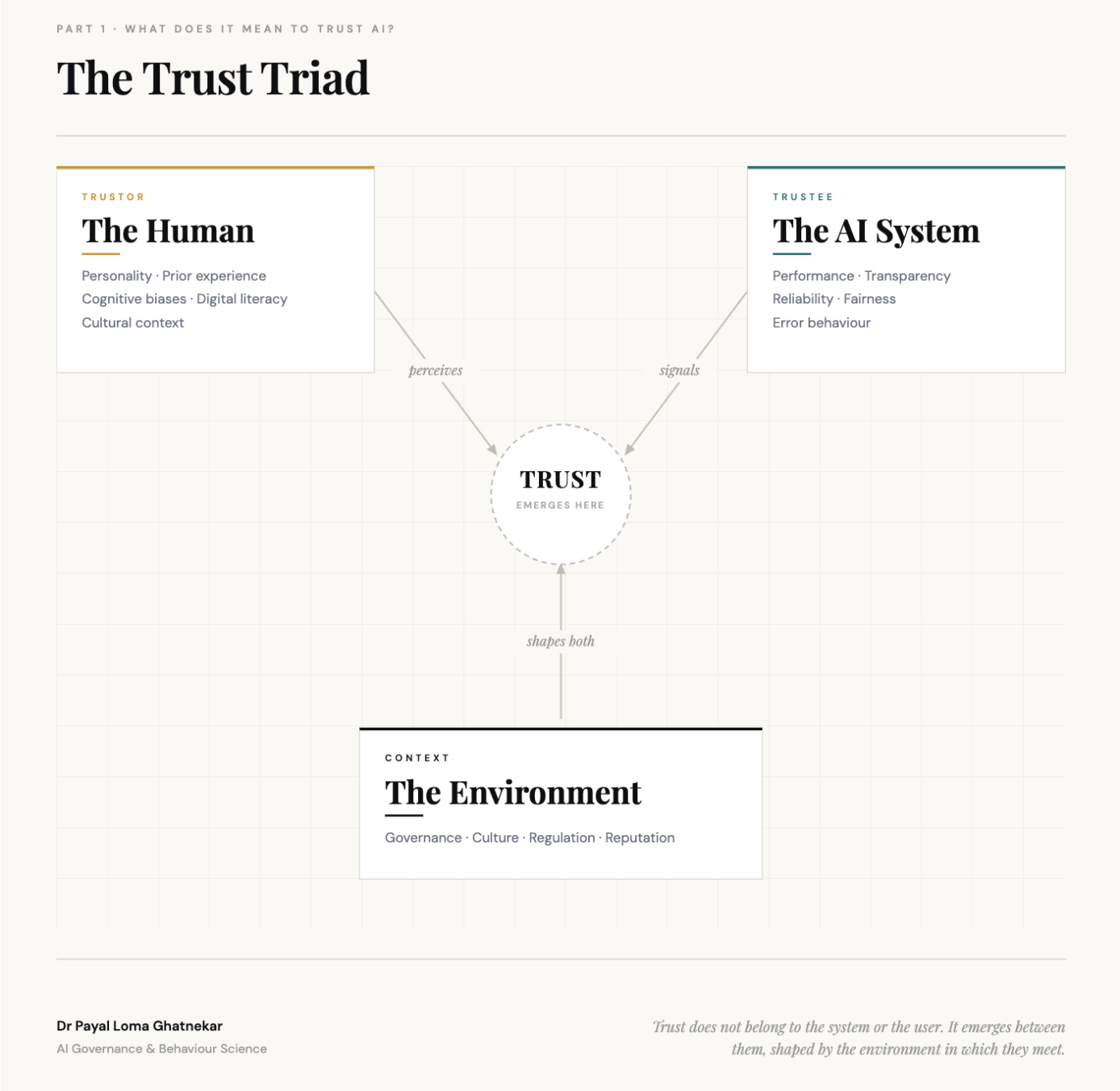
This is not irrational. Humans are wired to interpret agency even when agency is not real. The cognitive tendency to anthropomorphise is one of the most powerful drivers of AI trust formation — and one of the riskiest, because warmth illusions can mask competence failures. Researchers increasingly treat trust in AI and trust in humans as distinct psychological constructs, but the two share enough in common that the boundary matters: with AI, most assumptions about capability and intention are inferred rather than observed.

## 1.4 The Triad: Trustor, Trustee, and Context

A robust framework for understanding AI trust requires three elements working together. The trustor — the human — brings personality traits, prior experiences, cognitive biases, digital literacy, and cultural context. Trust is never purely about the machine; it is always filtered through the human mind.

The trustee — the AI system — communicates its trustworthiness through performance, transparency, reliability, fairness, and error behaviour, intentionally or not. The interactive context — the environment — encompasses organisational culture, governance structures, policy frameworks, and institutional reputation. People trust AI less when they distrust the organisation behind it, a pattern consistently demonstrated across healthcare, banking, and public-sector deployments.

Trust does not belong to the system or the user. It emerges between them, shaped by the environment in which they meet.




## 1.5 Why Calibrated Trust — Not Simply 'Building Trust' — Is the Real Goal

AI adoption is no longer limited by technological capability. It is limited by cognitive fears, lack of clarity, perceived loss of control, institutional mistrust, uncertainty about accountability, and discomfort with black-box reasoning. When trust is low, even high-performing systems are rejected. When trust is high but misplaced, systems are misused.

This is why building trust is not the goal. The real goal is calibrated trust — trust that is proportional, evidence-based, and matched to real capability. Calibrated trust makes people feel safer relying on the system, allows errors to become learning opportunities rather than disasters, enables institutions to demonstrate accountability, and supports users in maintaining agency rather than feeling overruled. Rapid adoption requires psychological confidence. Responsible adoption requires technical integrity. Calibrated trust requires both.

**RESEARCH EVIDENCE**

Trust in AI and trust in humans are psychologically dissociable constructs — they do not strongly correlate. This means AI governance cannot simply borrow the interpersonal trust models developed to understand human relationships. A distinct framework is needed - *Montag et al. (2023). Trust toward humans and trust toward artificial intelligence are not associated. PMC.*

The IBM Watson oncology case illustrates the consequences of misplaced trust: the system provided treatment suggestions later described as incorrect and potentially dangerous — yet the authority of an AI recommendation created pressure to follow it. Compliance without justified trust is not adoption. It is risk exposure.

*Trust in AI is not a single variable to optimise. It is the outcome of complex interactions between human psychology, machine characteristics, and institutional context. Only by understanding these three layers — trustor, trustee, and context — can we design systems that people will adopt quickly and rely on responsibly.*

PART 2

# Why Humans Miscalibrate AI

*The behavioural psychology behind misuse and disuse*

If Part 1 established what it means to trust AI, Part 2 addresses a harder question: why do people so consistently get that trust wrong? Miscalibrated trust is one of the most consequential challenges in AI adoption — and it is not random. It follows predictable behavioural patterns rooted in human cognition, emotion, and social psychology.

The problem is not that people make irrational decisions about AI. The problem is that they make human decisions about AI — and human decision-making is guided by biases, heuristics, and emotional responses that evolved long before machine intelligence existed.

## 2.1 The Two Failure Modes

Calibrated trust means the level of confidence placed in an AI system accurately reflects its true competence — its reliability, limitations, uncertainty, and contextual risks. In practice, two patterns consistently dominate instead.

Under-trust leads to disuse and algorithm aversion: people reject AI even when it performs demonstrably better than human judgment. Over-trust leads to misuse and automation bias: people accept AI outputs uncritically, even when contradictory evidence is available. Both slow responsible adoption, in opposite directions, and both emerge from deeply human psychological tendencies rather than considered evaluation.

## 2.2 Under-Trust: Algorithm Aversion

Algorithm aversion is well-documented. Give people an AI model that makes a small, visible mistake, and they lose confidence far more rapidly than they would after the same mistake from a human expert. Human errors feel understandable and correctable. AI errors feel unpredictable, opaque, and systemic. This is evidence that the system itself may be flawed.

The cognitive roots run deeper than a single bad experience. When people cannot see the reasoning behind a decision, they fill the gap with anxiety where uncertainty interpreted as danger. When agency is removed, mistrust increases dramatically; even a small degree of override capability significantly improves adoption. People also consistently overvalue human judgment in domains that feel morally charged, believing that AI cannot grasp nuance, read situations, or apply ethical reasoning, regardless of whether that belief is technically justified.

## 2.3 Over-Trust: Automation Bias

At the other extreme, over-trust produces automation bias. This is the tendency to accept AI outputs uncritically, ignore contradictory evidence, and gradually stop monitoring a system that initially demanded attention. This happens because AI presents itself with confidence, reduces cognitive load, creates an illusion of objectivity, and offers fast answers in stressful contexts. In high-pressure environments such as clinical settings, over-reliance increases precisely because people cannot check everything manually. Cognitive overload is one of the strongest predictors of over-reliance.

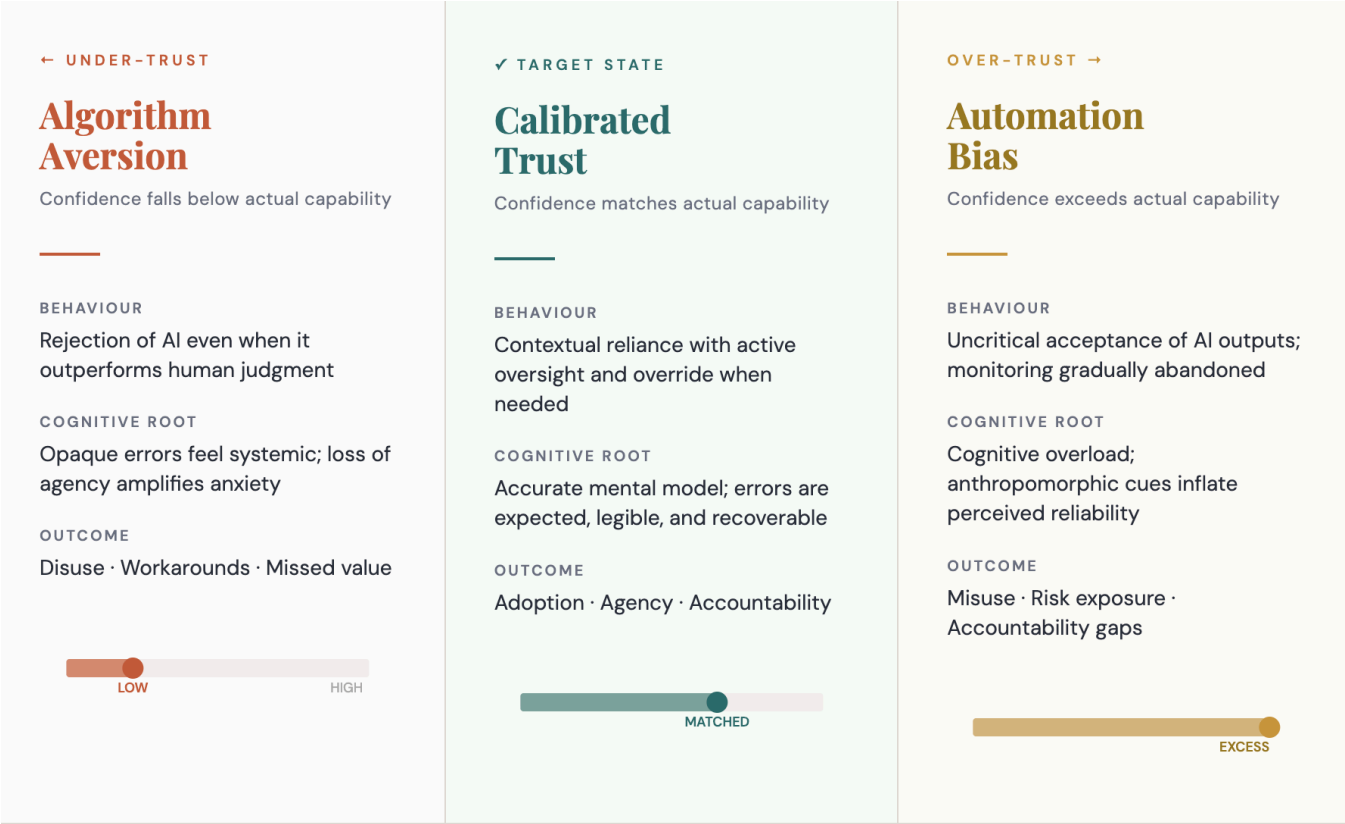
Several design features unintentionally amplify this. Anthropomorphic signals like warm voices, conversational tone, human-like phrasing, increase perceived competence even when performance does not justify it. Consistency of output, even in a mediocre system, creates perceived reliability which in turn creates trust. Clean, polished interfaces generate a halo effect where accuracy is assumed from aesthetics. And social proof along the lines of –“if everyone else uses it” appears to normalises reliance before due diligence has been applied.

## 2.4 Three Root Mechanisms

Most trust errors trace back to one of three underlying patterns. First, cognitive shortcuts: human brains are not designed for statistical reasoning. They rely on fast intuition, pattern recognition, and emotional inference — while AI operates on probability, optimisation, and data distributions. The mismatch between how humans reason and how machines process creates persistent misalignment.

Second, emotional meaning-making: people judge AI not in technical terms but emotionally and morally, even when they believe otherwise. A system that feels cold or mechanical loses trust even when accurate. A system that feels warm and socially attuned gains trust even when it occasionally errs.

Third, the need for control: behavioural science consistently shows that people trust systems which preserve autonomy, and distrust grows when agency is removed. Systems that allow overrides, offer explanations, support user judgment, and adapt to user preferences earn more trust independent of raw performance. Control is the psychological antidote to algorithm aversion.




*Miscalibration is not random — it follows predictable patterns rooted in human cognition, emotion, and social psychology.*

## 2.5 Trust and Distrust Are Not Opposites

A key insight often overlooked in AI design is that trust and distrust are not mutually exclusive. Trust asks: will this help me? Distrust asks: will this harm me? People can hold both feelings simultaneously — believing AI is good at finding patterns while also fearing it will misjudge their specific case. This duality explains why AI adoption is inherently fragile. A system can be substantially trusted and simultaneously suspected, and that combination shapes behaviour in ways that neither attitude can predict, alone.

**RESEARCH EVIDENCE**

Many participants trust AI more than humans because AI is perceived as impartial and free from self-interest. Yet in clinical contexts, 34% of radiologists overrode correct AI suggestions following a single visible error — potentially reducing the overall benefit of the technology. The gap between perceived objectivity and actual behaviour under uncertainty is where governance must operate- *Gerlich (2024). Exploring Motivators for Trust in the Dichotomy of Human-AI Trust. MDPI Social Sciences, 13(5).*

*Miscalibrated trust is rarely a user problem. It is the predictable outcome of cognitive overload, uncertainty, and loss of control operating on psychology that evolved to evaluate people, not machines. It is not solved by better algorithms or larger datasets — it is solved by behavioural design, transparent communication, and governance that signals genuine accountability.*

PART 3

# How AI Design Shapes Trust

*Where psychology, interface design, and machine behaviour converge*

If trust in AI begins with psychology and miscalibration emerges from predictable cognitive patterns, Part 3 turns to where trust becomes tangible: design. Design is often treated as the aesthetic layer of AI systems, its the colour palette, the interface, the conversational tone. However in reality, it is one of the most powerful psychological levers for shaping trust. It is how systems communicate competence, signal intentions, and manage the user's sense of control.

People do not interact with algorithms. They interact with representations of those algorithms, through interfaces, voices, dashboards, messages, and behaviours. Trust is therefore shaped not only by what a model does but by how it presents itself and how the user experiences that presentation. Design becomes the mediator between machine capability and human perception.

## 3.1 Two Dimensions of Perceived Trustworthiness

Social psychology has long shown that trustworthiness judgments are built on two pillars: competence (ability, expertise, reliability) and warmth (intentions, fairness, benevolence). These same dimensions apply when people evaluate AI systems, even though AI has no motives or emotions. The moment a system interacts with us, we instinctively assess whether it is capable, is treating us fairly, and is acting in our interests.

This is why trustworthy AI must be designed to signal both competence and warmth through its behaviour and communication — not through branding alone.

## 3.2 Competence: What Signals Ability

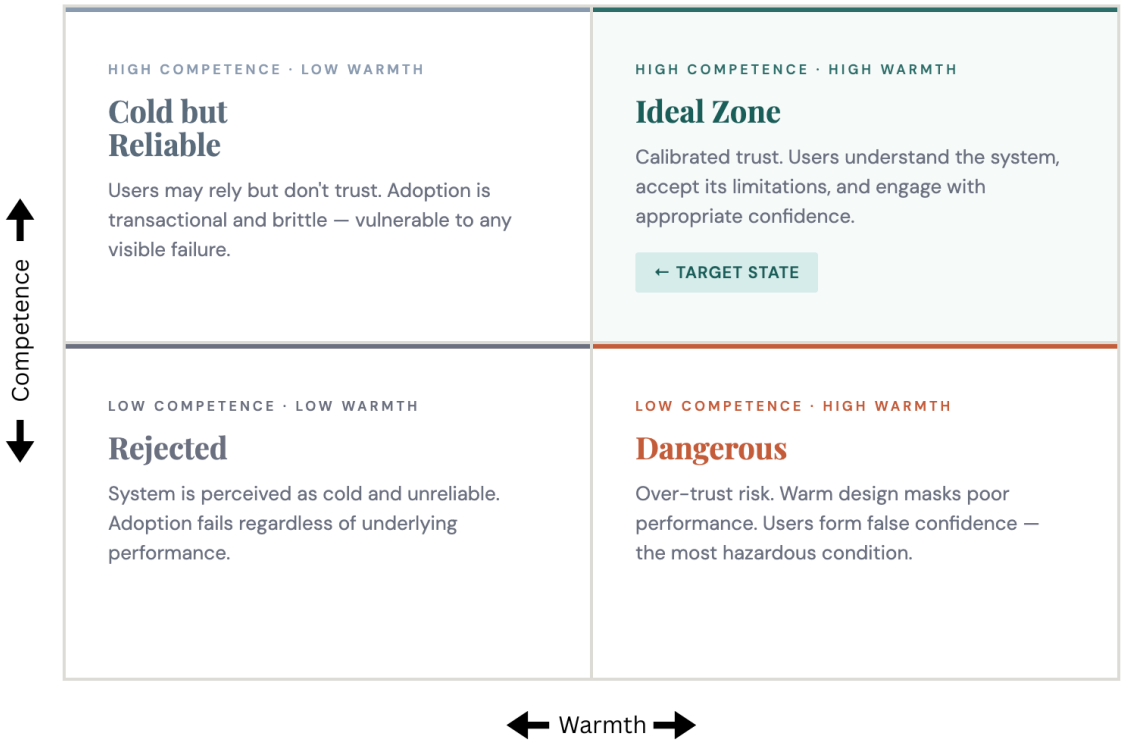
Users infer competence not through model architecture or accuracy metrics but through observable cues. Consistency and stability signal reliability: even small inconsistencies in phrasing, unexplained corrections, or changes in tone can undermine confidence. How a system handles mistakes matters more than whether it makes them, for example, unexplained errors read as loss of control, while contextualised errors are absorbed as part of expected uncertainty.

Counterintuitively, systems that openly state their limitations — 'I may be wrong about this,' 'confidence is low,' 'this is outside my training data' — are trusted more, not less. Communicating uncertainty demonstrates honesty, calibration, and awareness. Competence is not about perfection. It is about clarity, consistency, and comprehensible reasoning.

## 3.3 Warmth: What Signals Intention

Even though AI does not possess intentions, humans infer warmth through design. Users want to know who benefits from the system, whether they might be disadvantaged, and whether it treats them equitably. When fairness is hidden, people assume unfairness. When fairness is explained, trust grows.

Privacy cues operate as moral signals — clear data boundaries, consent processes, explanations for data use, and user control over visibility, all shape whether users believe the system is acting in their interest. Tone and language function similarly: a system that explains decisions respectfully, acknowledges user concerns, and uses human-centred language, is perceived as more trustworthy than one that feels abrupt or imposed. Warmth does not substitute for competence, but it shapes whether people are willing to accept competence when it is demonstrated.




*Trust is built on two pillars — competence and warmth. When warmth outpaces competence, the result is not adoption: it is risk.*

## 3.4 Anthropomorphism: Powerful, Predictable, and Risky

Anthropomorphism — attributing human-like characteristics to non-human agents — is one of the strongest trust drivers in AI design. People respond positively to natural voices, conversational phrasing, social timing, and polite turn-taking. These signals reduce uncertainty and create familiarity, lowering cognitive load and making interactions feel easier and more predictable. All of this is a prerequisite for trust.

But anthropomorphism is a double-edged instrument. Human-like cues increase adoption speed and engagement, but they can simultaneously cause users to overestimate capability, assume moral intuition, forgive mistakes too readily, and stop monitoring the system. The appropriate use of anthropomorphism is therefore deliberate, not liberal — its purpose is to aid calibration, not inflate it.

## 3.5 Explainability: Relevance Over Completeness

Explainable AI is often introduced as a technical intervention, but its real impact is psychological. Users want to understand why a system made a recommendation, what factors influenced it, how confident it is, and whether alternatives exist. Explainability succeeds when it reduces cognitive uncertainty — not when it attempts to make users understand model architecture.

The persistent misconception is that more transparency produces more trust. People do not want full transparency because that is overwhelming and often counterproductive. They want relevant clarity: simple, contextual, local explanations that are honest about uncertainty and actionable in the moment. The goal of explainability is cognitive alignment — helping users form an accurate mental model of how the system behaves so they can anticipate its strengths, recognise its limitations, and intervene when needed.

**RESEARCH EVIDENCE**

Social and conversational cues significantly increase perceived trustworthiness, even when technical understanding is low. Notably, users were willing to share more sensitive information with neutral, machine-like systems — even though they reported trusting those systems less. This finding highlights a critical distinction: felt trust and responsible use are not the same thing. Emotional comfort generated by warm design does not reliably translate into safer or more thoughtful behaviour.- *Gerlich (2024). Exploring Motivators for Trust in the Dichotomy of Human-AI Trust. MDPI Social Sciences, 13(5).*

For the relationship between linguistic style and user perception in conversational AI, see also: Skantze et al. (2023). Effects of chatbot language style on user trust and perception. ResearchGate.

*People do not trust algorithms — they trust the signals a system sends through behaviour, design, and communication. When design overpromises capability, trust becomes inflated and safety is compromised. When design under communicates value, trust stays low and adoption stalls. The designer's responsibility is to close that gap — ensuring that what the system signals also matches what it can actually do.*

PART 4

# What Institutions Must Do

*Oversight, governance, and the systemic conditions for justified trust*

Trust that emerges from warm interfaces or persuasive design alone is fragile. It collapses the moment something goes wrong. Sustainable trust, the kind required for healthcare, public services, or enterprise AI adoption, depends on context: the governance structures, oversight mechanisms, organisational culture, and regulatory expectations surrounding the technology.

No AI system exists in isolation. A model is always embedded within an organisation, a workflow, a set of responsibilities, and a legal and ethical framework. Trust in AI is therefore not only a question of 'do I trust the system?' but of 'do I trust the institution deploying it? Do I trust that it is monitored? Do I trust that someone is accountable if something goes wrong?' Institutions shape the conditions under which trust can be justified — and without those conditions, even technically excellent AI cannot earn durable confidence.

## 4.1 The Three Pillars of Trustworthy AI

Regulatory bodies worldwide have converged on a common foundation: AI systems must be lawful, ethically sound, and technically robust. Lawfulness requires compliance with data protection frameworks, patient rights, discrimination protections, and sector-specific regulation. Ethical soundness requires upholding fairness, respect for autonomy, transparency, and non-maleficence. Technical robustness requires reliable behaviour across conditions, resistance to manipulation, and safe failure modes.

These pillars are only meaningful when institutions operationalise them, not as documentation exercises but as lived practice embedded in how decisions are made, how incidents are handled, and how accountability is assigned.

## 4.2 Accountability and Auditability

Trust collapses when people are unclear about who is responsible when something goes wrong. Users need to know who built the system, who validated it, who monitors it, who intervenes when it fails, and who is liable when harm occurs. Accountability is as much a psychological factor as a procedural one: when organisations demonstrate clear ownership of AI decisions, users feel protected even in the presence of error.

Auditability is the structural mechanism that makes accountability visible. It requires documentation of model development, traceable decision flows, version control, monitoring for performance drift, explainability mechanisms, and escalation pathways. Critically, auditability must exist before deployment, to ensure systems are safe. Furthermore, auditability must also exist throughout deployment, to ensure systems remain safe. Without robust auditing, even well-performing systems struggle to earn trust, because the chain of accountability is invisible to users.

## 4.3 Human-in-the-Loop and Society-in-the-Loop

Human-in-the-loop (HITL) is frequently misunderstood as a bureaucratic checkbox. In practice, it is a psychological safety mechanism. The ability to intervene, override, or correct a system is one of the strongest predictors of user trust. HITL ensures that decisions are reviewable, users retain control, and responsibility is genuinely shared. In high-stakes environments such as, healthcare, justice, finance, social care — this is not optional.

Society-in-the-loop (SITL) extends oversight beyond individual users and domain experts to include communities, public stakeholders, and affected groups. It shifts the question from 'is the system safe?' to 'is the system aligned with societal expectations?'. Participatory design, public consultation, ethics committees, and value-sensitive design frameworks are all expressions of SITL in practice. Trust is strongly influenced by whether people feel represented in decisions that affect them.

## 4.4 The Multi-Stakeholder Trust Chain

AI governance involves an interconnected network of actors: developers, data scientists, product managers, clinicians, domain specialists, leaders, regulators, auditors, end users, and those affected by AI decisions. Each brings different priorities. Developers focus on performance. Regulators focus on safety. Leaders focus on cost and efficiency. End users focus on reliability and usability. Decision subjects focus on fairness and agency.

When one stakeholder's needs dominate or responsibilities are unclear, the chain fractures and trust erodes. Governance must coordinate across the entire chain — not simply regulate isolated parts.

## 4.5 Organisational Culture as a Hidden Determinant

Even the best technical safeguards fail when deployed within an organisation characterised by unclear communication, opaque decision-making, inconsistent policy, fear-driven leadership, or a history of digital transformation failures. AI inherits the culture around it. A culture that avoids accountability will deploy AI that avoids accountability. A culture that values integrity will deploy AI that signals integrity. Trust in technology cannot exceed trust in leadership.

## 4.6 Regulation as a Trust Signal

Internal governance is essential but not sufficient. People expect external guardrails — independent verification, global standards, certification, and proof of enforcement. Regulation serves several psychological functions simultaneously: it reduces uncertainty, provides assurance, distributes liability, and establishes common expectations. The EU AI Act, for example, does not only classify AI by risk level — it creates a framework designed to make trust feel justified rather than merely assumed.


*Trust in AI cannot exceed trust in the institution deploying it. Oversight is not an administrative burden — it is part of the user experience.*

**RESEARCH EVIDENCE**

Decision subjects value agency, contestability, and inclusion far more than developers typically expect. This finding provides strong empirical support for HITL, SITL, and multi-stakeholder governance. The gap between what developers believe users need and what users actually value, is itself a governance risk.-*Alizadeh et al. (2024). Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. ResearchGate.*

NHS 111 Online illustrates how institutional context shapes trust independently of interface design. The service earns authority primarily through NHS branding and reputation — users interpret its guidance as more definitive than intended. This creates a governance challenge: when AI tools are deployed within trusted institutions, institutional legitimacy amplifies perceived reliability even when systems are designed as decision support, not decision makers.

*Trust in AI cannot exceed trust in the institution deploying it. Accountability, oversight, and governance are not administrative burdens — they are part of the user experience. Without them, trust becomes fragile, reputation becomes the risk surface, and adoption remains brittle regardless of technical performance.*

PART 5

# Building Calibrated Trust for Real-World AI Deployment

*From fragmented trust to a coherent socio-technical strategy*

Across the previous four parts, a pattern has emerged. Trust in AI is shaped by human psychology, mediated through design, and sustained or undermined by institutions. When any one of these layers is neglected, trust becomes unstable. When all three align, trust becomes calibrated — and calibrated trust is the condition under which AI can be adopted quickly without compromising safety, dignity, or human agency.

Calibrated trust is not static. It evolves through experience, feedback, error handling, and institutional behaviour. It reflects an accurate mental model of what the system can do, what it cannot do, and how it behaves under uncertainty. Building it requires intentional design across the entire AI lifecycle — not a one-time deployment decision.

## 5.1 Why Calibration Is the Goal, Not Trust Itself

Much of the public discourse around AI focuses on building trust. This framing is misleading. Trust on its own is not inherently positive. High trust in an unreliable system is dangerous. Low trust in a highly capable system is wasteful. The real objective is alignment: between system capability and user expectations, between design signals and actual performance, between institutional assurances and lived experience.

When trust is misaligned, we see the two failure modes established earlier — over-trust leading to automation bias and misuse, under-trust leading to rejection and workarounds. Calibrated trust avoids both by anchoring confidence in evidence, transparency, and agency rather than persuasion or fear.

## 5.2 The Three Interdependent Layers

### Technical Integrity: Making Trustworthiness Observable

Technical integrity refers to the actual properties of the AI system and how clearly those properties are communicated. This includes robustness across contexts, predictable behaviour, transparent limitations, explainability that supports decision-making, appropriate handling of uncertainty, and safe failure modes. A key insight from trust psychology is that users do not need perfect systems — they need legible ones.

Critically, integrity must be observable. Performance metrics locked in technical documentation do little to shape trust. What matters is how competence is experienced in interaction, i.e., does the system behave consistently over time? Are errors explained? Are limitations explicit rather than hidden? Technical integrity creates the conditions for justified trust, but it does not guarantee that users will respond accordingly.

### Human Reality: Designing for How People Actually Think

Humans do not evaluate AI systems as rational auditors. They evaluate them as cognitive, emotional, and socially embedded beings. Many deployments fail not because the system is flawed but because it was designed for an idealised user who does not exist in practice.

Designing for human reality means preserving agency through overrides and genuine choice, reducing cognitive load rather than increasing it, supporting sense-making rather than simply providing explanation, and allowing users to develop an accurate mental model over time. Education plays a role, but it is not sufficient. Users do not want to become AI experts, they want to know when to rely, when to question, and when to intervene. Good design makes those boundaries visible without demanding constant vigilance.

### System Responsibility: Creating Trustable Environments

Even technically sound, well-designed systems remain fragile when the surrounding environment is untrustworthy. System responsibility taken on by the institutional, organisational, and regulatory structures governing AI — performs a vital psychological function. It absorbs uncertainty on behalf of the user. People are more willing to rely on AI when they believe someone is monitoring the system, errors will be addressed, harm will be acknowledged, and safeguards exist beyond their individual control. In this sense, trust is not merely interpersonal or human-machine — it is institutional.

## 5.3 Trust as a Dynamic Process

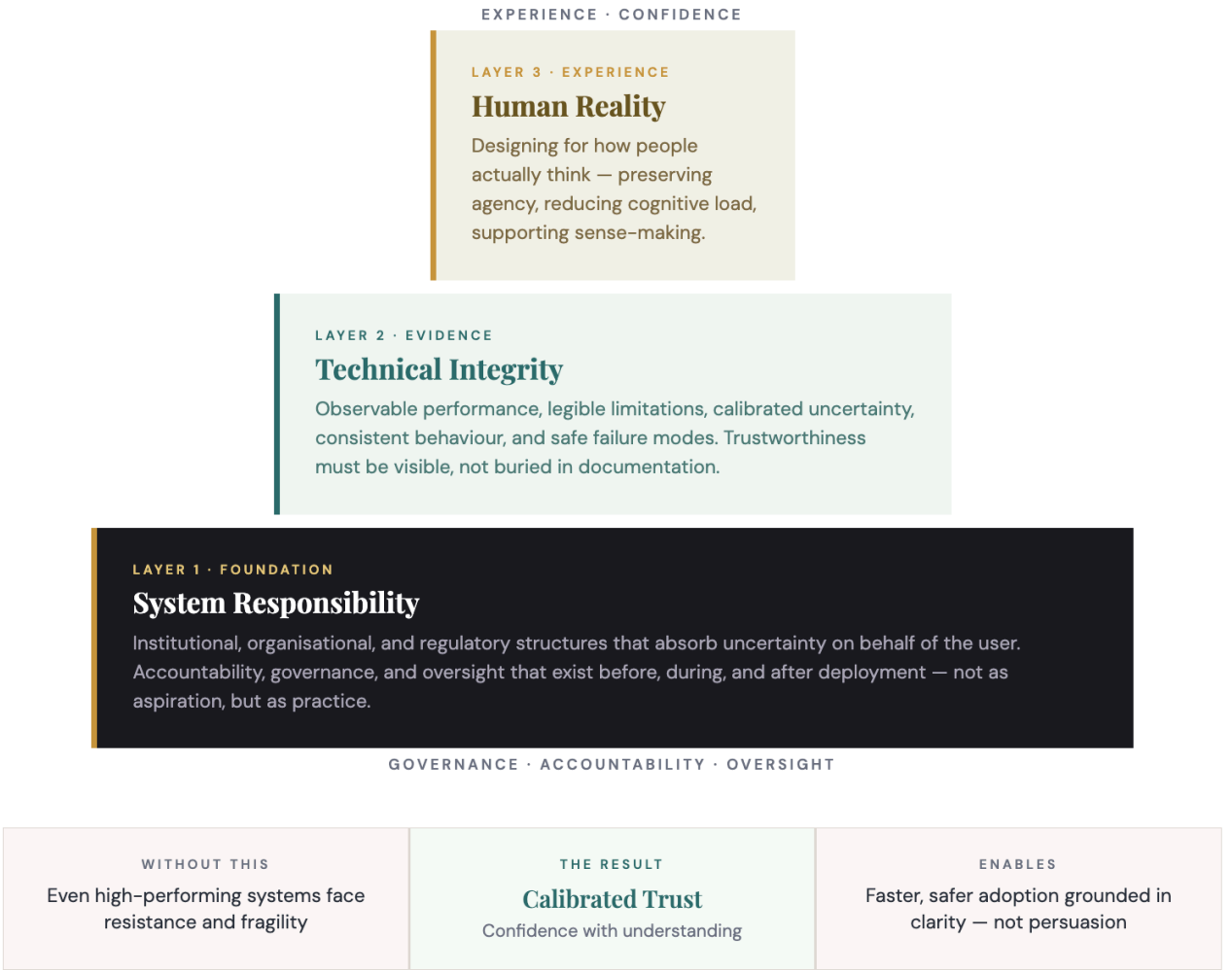
One of the most common mistakes in AI deployment is treating trust as something that can be achieved at launch. Trust is continuously renegotiated throughout adoption — reshaped as users gain experience, systems are updated, use-contexts shift, and errors occur. Calibration changes in response to all of these.

Trust is strengthened not by the absence of failure but by how failure is handled. Transparent acknowledgement, timely correction, and visible learning all contribute to trust recovery. Silence, defensiveness, or blame-shifting do the opposite. Designing for calibrated trust therefore means planning not only for success but for misunderstanding, misuse, edge cases, breakdowns, and the institutional response to them.

## 5.4 Why Calibrated Trust Enables Faster, Safer Adoption

Calibrated trust appears to slow adoption, however in practice, the opposite is true. Systems designed for calibrated trust face less resistance, experience fewer catastrophic failures, integrate more smoothly into workflows, require fewer corrective interventions, and sustain confidence over time. When users understand a system's boundaries, they are more willing to rely on it. When institutions demonstrate accountability, they reduce fear. When design preserves agency, adoption becomes a choice rather than a mandate.

Rapid adoption does not come from persuasion or pressure. It comes from confidence grounded in clarity.




*Calibrated trust is confidence with understanding — not faith, and not fear.*

**RESEARCH EVIDENCE**

Trustworthy AI requires alignment between technical robustness, ethical principles, and institutional accountability — none of the three is sufficient alone. The integration of all three is what distinguishes durable, scalable AI adoption from fragile, compliance-driven deployment.- *Papagiannidis et al. (2025). Responsible Artificial Intelligence Governance: A Review and Research Framework. Science Direct.*

Computer-aided detection in breast cancer screening provides a compelling real-world model of calibrated trust in action. These systems function as a second reader — highlighting areas of concern without making diagnostic decisions. Radiologists are trained to assess and frequently override AI outputs, with no expectation of compliance. Trust is conditional on context and professional judgment, not emotional or absolute. This deployment demonstrates how institutions can support AI-augmented expertise while preserving human authority, clinical responsibility, and accountability.

## A Calibrated Trust Checklist for Deployment Teams

Before deploying any AI system into a setting where human decisions are consequential, these questions should have clear, documented answers:

- Can users articulate what the system is not designed to do?
- Are uncertainty and limitations visible at the point of decision — not buried in documentation?
- Is there a clear override pathway that users feel psychologically safe using?
- Are monitoring, drift detection, and incident response operational — not aspirational?
- Do governance roles have named individual owners, not committees without accountability?
- Are fairness and privacy communicated in user-centred language, not legal jargon?
- Has the deployment been designed for failure and recovery, not only for successful use?

*Calibrated trust is confidence with understanding — not faith, and not fear. The question is no longer whether we can build AI that performs. It is whether we can build systems that people can rely on for the right reasons — and institutions capable of sustaining that reliance when things go wrong.*

**About the Author**

Dr. Payal Loma Ghatnekar is Technologies Research Programme Manager at NHS, an AI governance specialist, and a certified AI Ethicist (University of Oxford). She has authored institutional AI governance policy aligned with the EU AI Act, ISO/IEC 42001, MHRA, and AIDRS, and leads the governance and behavioural evaluation of clinical AI systems and digital technologies in NHS settings.

Her work sits at the intersection of AI governance, behavioural science, and healthcare — translating regulatory frameworks into institutional practice and examining how humans actually adopt, resist, and miscalibrate trust in AI systems.

**payalloma.com | linkedin.com/in/drpayallomaghatnekar**

**A note on AI assistance:** Sections of this paper were drafted and refined with the support of AI language tools- Google Notebook LM, Claude and GPT. The research synthesis, theoretical frameworks, behavioural science analysis, and all conclusions reflect the author's own expertise and professional judgement. AI was used as a writing and editing aid, not as a source of ideas or evidence.

## *References*

1. **Afroogh, S., Akbari, A., Malone, E., Kargar, M. and Alambeigi, H. (2024)** 'Trust in AI: progress, challenges, and future directions', *Humanities and Social Sciences Communications*, 11(1).
2. **Bach, T.A., Khan, A., Hallock, H., Beltrão, G. and Sousa, S. (2024)** 'A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective', *International Journal of Human–Computer Interaction*, 40(5), pp. 1251–1266.
3. **Choung, H., David, P. and Ross, A. (2022)** 'Trust in AI and Its Role in the Acceptance of AI Technologies', *International Journal of Human-Computer Interaction*, pp. 1–13.
4. **Gerlich, M. (2024)** 'Exploring Motivators for Trust in the Dichotomy of Human—AI Trust Dynamics', *Social Sciences*, 13(251).
5. **Herrera-Poyatos, A., Del Ser, J., López de Prado, M., Wang, F.Y., Herrera-Viedma, E. and Herrera, F. (2025)** 'Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust Through Trustworthy AI, Auditability, Accountability, and Governance', *arXiv preprint*.
6. **Kwon, I.W.G. and Kim, S.H. (2025)** 'A Roadmap Toward Building Trust in AI Environment: Research Agenda', *British Journal of Healthcare and Medical Research*, 12(1), pp. 254–262.

7. **Li, Y., Wu, B., Huang, Y. and Luan, S. (2024)** 'Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust', *Frontiers in Psychology*, 15(1382693).
8. **López-Pernas, S., Oliveira, E., Song, Y. and Saqr, M. (2025)** 'AI, Explainable AI and Evaluative AI: Informed Data-Driven Decision-Making in Education', in *Advanced Learning Analytics Methods*.
9. **Lundberg, S.M. and Lee, S.I. (2017)** 'A Unified Approach to Interpreting Model Predictions', *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 4765–4774.
10. **Meske, C., Brenne, J., Ünal, E., Ölcer, S. and Dogangün, A. (2024)** 'From Explainable to Explanatory Artificial Intelligence: Toward a New Paradigm for Human-Centered Explanations through Generative AI', *arXiv preprint*.
11. **Middleton, S.E., Letouzé, E., Hossaini, A. and Chapman, A. (2022)** 'Trust, Regulation, and Human-in-the-Loop AI within the European Region', *Communications of the ACM*, 65(4).
12. **Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B.W. and Becker, B. (2023)** 'Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging', *Personality Neuroscience*, 6(e3), pp. 1–8.
13. **Newen, C., Bodemer, D., Glantz, S., Müller, E., Wischnewski, M. and Schnaubert, L. (2023)** 'Uncertainty Awareness and Trust in Explainable AI - On Trust Calibration using Local and Global Explanations', *arXiv preprint*.
14. **Papagiannidis, E., Mikalef, P. and Conboy, K. (2025)** 'Responsible artificial intelligence governance: A review and research framework', *Journal of Strategic Information Systems*, 34(101885).
15. **Perrig, S.A.C., Scharowski, N. and Brühlmann, F. (2023)** 'Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI', *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA '23)*.
16. **Ramachandram, D., Joshi, H., Zhu, J., Gandhi, D., Hartman, L. and Raval, A. (2025)** 'Transparent AI: The Case for Interpretability and Explainability', *Vector Institute for Artificial Intelligence White Paper*.
17. **Reuter, M., Kirchhoff, B.M., Franke, T., Radüntz, T. and Peifer, C. (2025)** 'To trust or not to trust a human(-like) AI—A scoping review and conjoint analyses on factors influencing anthropomorphism and trust', *Zeitschrift für Arbeitswissenschaft*, 79, pp. 402–432.
18. **Starke, G., van den Brule, R., Elger, B.S. and Haselager, W.F.G. (2022)** 'Intentional machines: A defence of trust in medical artificial intelligence', *Bioethics*, 36(2), pp. 154–161.
19. **Daniel, A. (2024)** 'Preliminary Quantitative Study on Explainability and Trust in AI Systems', *arXiv preprint*.
20. **Tutul, A.A., Nirjhar, E.H. and Chaspari, T. (2021)** 'Investigating Trust in Human-Machine Learning Collaboration: A Pilot Study on Estimating Public Anxiety from Speech', *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*, pp. 288–296.
21. **Vereschak, O., Alizadeh, F., Bailly, G. and Caramiaux, B. (2024)** 'Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made', *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*.