COALITION FOR INNOVATION

AI
Blueprint for the Future

Coalition for Innovation, supported by LG NOVA

Jami Diaz, Director Ecosystem Community & Startup Experience
William Barkis, Head of Grand Challenges & Ecosystem Development
Sokwoo Rhee, Executive Vice President, LG Electronics, Head, LG NOVA

Coalition for Innovation Co-Chairs

Alex Fang, CleanTech Chair
Sarah Ennis, AI Chair
Alfred Poor, HealthTech Chair

Authors

Adrien Abecassis, Johnny Aguirre, John Barton, Ann M. Marcus, Olivier Bacs, Taylor Black, Micah Boster, Mathilde Cerioli, Carolyn Eagen, Sarah Ennis, Annie Hanlon, Christina Lee Storm, Andrew Yongwoo Lim, Jess Loren, Refael Shamir, Svetlana Stotskaya

The views and opinions expressed in the chapters and case studies that follow are those of the authors and do not necessarily reflect the views or positions of any entities they represent.

Senior Editor, Alfred Poor
Editor, Jade Newton

October 2025

Preamble

The Coalition for Innovation is an initiative hosted by LG NOVA that creates the opportunity for innovators, entrepreneurs, and business leaders across sectors to come together to collaborate on important topics in technology to drive impact. The end goal: together we can leverage our collective knowledge to advance important work that drives positive impact in our communities and the world. The simple vision is that we can be stronger together and increase our individual and collective impact on the world through collaboration.

This "Blueprint for the Future" document (henceforth: "Blueprint") defines a vision for the future through which technology innovation can improve the lives of people, their communities, and the planet. The goal is to lay out a vision and potentially provide the framework to start taking action in the areas of interest for the members of the Coalition. The chapters in this Blueprint are intended to be a "Big Tent" in which many diverse perspectives and interests and different approaches to impact can come together. Hence, the structure of the Blueprint is intended to be as inclusive as possible in which different chapters of the Blueprint focus on different topic areas, written by different authors with individual perspectives that may be less widely supported by the group.

Participation in the Coalition at large and authorship of the overall Blueprint document does not imply endorsement of the ideas of any specific chapter but rather acknowledges a contribution to the discussion and general engagement in the Coalition process that led to the publication of this Blueprint.

All contributors will be listed as "Authors" of the Blueprint in alphabetical order. The Co-Chairs for each Coalition will be listed as "Editors" also in alphabetical order. Authorship will include each individual author's name along with optional title and optional organization at the author's discretion.

Each chapter will list only the subset of participants that meaningfully contributed to that chapter. Authorship for chapters will be in rank order based on contribution: the first author(s) will have contributed the most, second author(s) second most, and so on. Equal contributions at each level will be listed as "Co-Authors"; if two or more authors contributed the most and contributed equally, they will be noted with an asterisk as "Co-First Authors". If two authors contributed second-most and equally, they will be listed as "Co-Second Authors" and so on.

The Blueprint document itself, as the work of the group, is licensed under the Creative Commons Attribution 4.0 (aka "BY") International License: https://creativecommons.org/licenses/by/4.0/. Because of our commitment to openness, you are free to share and adapt the Blueprint with attribution (as more fully described in the CC BY 4.0 license).

The Coalition is intended to be a community-driven activity and where possible governance will be by majority vote of each domain group. Specifically, each Coalition will decide which topics are included as chapters by majority vote of the group. The approach is intended to be inclusive so we will ask that topics be included unless they are considered by the majority to be significantly out of scope.

We intend for the document to reach a broad, international audience, including:

- People involved in the three technology domains: CleanTech, AI, and HealthTech
- Researchers from academic and private institutions
- Investors
- Students
- Policy creators at the corporate level and all levels of government

Chapter 11:
Overreliance on AI

Author: John Barton

Overview

Overreliance on AI is no longer a speculative risk; it is an emergent design failure unfolding at scale. As generative AI tools become more persuasive, ubiquitous, and intuitive, users are increasingly treating outputs not as suggestions but as truths. This shift isn’t just behavioral. It reveals a foundational mismatch between how AI is designed, how it is deployed, and how humans build trust.

The Microsoft Aether Committee defines overreliance as "a behavioral state in which users defer judgment to an AI system even when they have reason, skill, or evidence to question it." Their 2023 report identifies causes ranging from poor onboarding and automation bias to low AI literacy and overconfident UX design. Across nearly 60 studies in HCI, organizational behavior, and cognitive psychology, the evidence is clear: overreliance is not rare, and it is not benign.

Two Views of Trust

The most critical distinction between this framework and the Microsoft Aether report lies in how each treats trust.

Aspect	Aether Paper	This framework
Definition of Trust	A cognitive or psychological state: often passive or assumed	A behavioral practice: dynamic, scaffolded, and situational
Trust Failure Framing	Overreliance = a result of psychological bias (e.g., automation bias)	Overreliance = a design failure that disables user agency
Mitigation Approach	Emphasizes transparency, explainability, interface labeling	Emphasizes recovery, reflection, and epistemic scaffolding
User Role	At-risk subject prone to bias or error	Active participant whose trust can be shaped, reclaimed, and redirected
System Role	Provide signals (confidence scores, disclaimers)	Shape behavior through growth-mode UX and calibrated friction

Core Trust Philosophy	Manage trust	Calibrate, support, and recover trust
Primary Risk Identified	Users trusting too much	Systems teaching users not to think

Where the Aether report treats trust as a cognitive error to be managed, this Framework reframes trust as a behavioral outcome of system design. It is not just what users believe; it’s what systems teach. And that makes it actionable.

What begins as user convenience quickly hardens into epistemic dependency. Users skip critical thinking steps. They stop verifying sources. They trust AI output even when it contradicts their own knowledge. This pattern shows up across domains; students use AI to draft papers without synthesis, professionals paste in summaries without review, and even high-stakes decisions (legal, medical, financial) are increasingly shaped by AI inputs that are treated as inherently correct.

Conventional risk mitigation — such as adding disclaimers or improving model accuracy — is inadequate. Users don't just misjudge factual correctness. They adopt structural habits that normalize outsourcing judgment. Without deliberate design for reflection, recovery, and agency, overreliance becomes entrenched.

This Framework offers a different approach. It reframes overreliance not as user failure but as a predictable outcome of current design patterns. By analyzing trust behaviors, behavioral defaults, and onboarding gaps, it introduces a quadrant-based model for understanding and redirecting user interaction. The model maps user mindsets (fixed or growth) against the systems they interact with (stagnant or innovative), revealing four distinct risk profiles and paths to recovery. Rather than attempting to "fix trust," the Framework centers **epistemic calibration**: the ability of users to engage with AI critically, adaptively, and reflectively.

In this Framework, overreliance is not just an error state. It is a signal: a warning that system scaffolding has failed to support user agency. And as AI tools accelerate in complexity and reach, the cost of ignoring that signal grows exponentially.

This document begins the work of designing for recovery, not just control. It offers language, structure, and intervention concepts that can be tested, refined, and embedded across AI development lifecycles, from onboarding to interface design to long-term trust calibration.

Stakeholders

The risk of overreliance on AI systems is not distributed equally. Different stakeholder groups encounter, reinforce, and are impacted by this risk in distinct ways. Understanding these roles is essential to designing effective interventions and allocating responsibility.

New AI Users (Students, Workers, Public Users)

These are individuals who interact with AI tools without deep technical understanding or prior exposure to epistemic safeguards. In educational and workplace settings, new users are particularly vulnerable to overreliance.

- Students often treat AI as a substitute for research or synthesis.

- Employees may defer to AI-suggested summaries, assuming correctness.
- Public users encounter persuasive AI outputs through chatbots, search engines, and productivity tools without visibility into system limitations.

Their default trust behaviors are shaped by onboarding quality, interface signals, and institutional norms. Without friction or calibration prompts, many new users develop passive reliance patterns that become difficult to reverse.

UX Designers and AI Product Teams

These teams play a central role in shaping user trust behaviors. From interface affordances to timing of suggestions, design decisions either reinforce or interrupt overreliance. Teams may unintentionally reward speed and frictionless interaction at the cost of critical engagement.

- Autocomplete and summarization tools can flatten nuance.
- Invisible errors or missing citations can mask epistemic risk.
- Systems rarely prompt reflection or critique after use.

User experience (UX) and product teams need access to trust metrics beyond engagement or completion rate. Without epistemic key performance indicators (KPIs), product success may coincide with user disempowerment.

Educators and AI Literacy Professionals

In both formal and informal learning environments, educators have a dual challenge: using AI tools to support learning while preventing them from replacing learning. When students internalize AI as a shortcut, educational systems risk reinforcing stagnation.

AI literacy professionals are beginning to surface strategies for teaching calibration, synthesis, and disagreement. However, they often lack access to tool internals or control over interface dynamics, which makes structural support for epistemic skill-building inconsistent and fragmented.

Policy and Trust/Safety Teams

These actors define the regulatory and ethical boundaries of AI deployment. While much of their work focuses on preventing harms like bias, surveillance, or misinformation, overreliance introduces a subtler but equally corrosive risk: the erosion of user judgment.

Trust and safety teams must evolve their scope to include behavioral defaults, recovery scaffolds, and misuse patterns that emerge from high-compliance but low-agency interactions.

Enterprise Deployment Leaders

In large organizations adopting AI tools across departments, the risk of overreliance is compounded by scale. Teams are encouraged to use AI for efficiency, but may lack guardrails for:

- Decision accountability,
- Epistemic quality control, or
- Feedback integration.

Over time, unexamined overreliance calcifies into cultural dependency, making it harder to restore initiative, judgment, or accountability. When it takes root in enterprise workflows, overreliance embeds passivity into processes that once relied on human judgment.

Investors and Strategic Funders

Investors — including those focused on responsible tech, venture capital, and social impact — have a vested interest in scalable, trustworthy AI systems. Overreliance poses both reputational and operational risks; it can erode user confidence, increase liability exposure, and lead to costly missteps or regulatory pushback.

By positioning this Framework as a model for designing *resilient trust* rather than frictionless compliance, we offer a value proposition aligned with long-term retention, product adaptability, and ethical market leadership. Investors increasingly recognize that trust infrastructure is not ancillary; it is core to AI product viability.

Foundations and Philanthropic AI Funders

Philanthropic organizations focused on digital equity, community resilience, and ethical AI education are emerging as key funders of harm-reduction strategies. These funders support public-interest work to reduce epistemic harms, especially in underserved populations.

This Framework aligns with their goals by offering a pathway to scalable, recovery-enabled systems that don't just avoid bias, but actively teach reflective, equitable AI use.

AI Developers and Foundation Model Teams

These upstream stakeholders shape the behavior, affordances, and epistemic posture of the models themselves. Their architectural decisions — ranging from pretraining data and reinforcement mechanisms to confidence signaling and answer calibration — directly affect downstream trust dynamics.

Without considering overreliance, core model teams may optimize for helpfulness while inadvertently encouraging overconfidence. Their role in supporting recovery lies in enabling systems that can pause, reflect, and revise: not just respond.

Policy and Governance Professionals

These include regulators, lawmakers, and standards organizations (e.g., NIST, ISO, EU AI Act) that set the external constraints for trustworthy AI. While much attention has been given to bias and data transparency, overreliance introduces a need for behavioral accountability; are systems producing not just safe outputs, but safe usage patterns?

Regulatory frameworks must expand to address trust calibration, scaffolding, and user resilience, not just data harm or content filtering.

Internal Trust & Safety and Ethics Teams

Within organizations, these teams are responsible for monitoring harm, abuse patterns, and reputational risk. Overreliance often escapes their purview because it looks like success: high engagement, satisfied users, few complaints.

However, uncritical use of AI can mask deep epistemic erosion. These teams must evolve to include metrics of user reflection, adaptive confidence, and behavioral feedback, not just incident reporting or legal risk.

Procurement and Risk Officers (Enterprise Subgroup)

In enterprise settings, the people selecting, and approving AI systems are often separate from those who use them. Procurement officers and risk managers play a hidden but powerful role in either embedding or mitigating overreliance.

Their assessment criteria can shape entire organizational adoption patterns. By integrating epistemic resilience, recovery scaffolds, and reflective tooling into vendor evaluation, they can drive demand for responsible AI at scale.

Each of these stakeholders holds a piece of the puzzle. Overreliance is not a problem of user ignorance alone. It is the result of structural gaps in design, deployment, governance, and education. Effective mitigation requires coordinated responses across these roles, with shared responsibility for building systems that support reflection, not just use.

Challenges and Gaps

Efforts to mitigate overreliance on AI have largely fallen short because they underestimate the complexity of the problem. The dominant response has been technical; add disclaimers, improve accuracy, or publish confidence scores. But these approaches miss the deeper mechanisms that drive behavioral dependency, stagnation, and loss of judgment.

Fluency Bias and the Loss of Friction

Modern AI systems are optimized for speed, fluency, and seamless UX. While these qualities enhance usability, they also reduce opportunities for reflection. When users are rewarded for accepting answers quickly — and penalized, in effect, for slowing down — they develop patterns of passive trust.

Features such as predictive text, auto-generated responses, and instant summarization encourage fluency over scrutiny. The design culture that celebrates frictionless interaction inadvertently discourages epistemic resistance. Without embedded challenges or critical pauses, users learn to trust by default: not because they are careless but because the system teaches them to.

Inadequate Onboarding Structures

Most AI tools are introduced with basic usage instructions and legal disclaimers. Very few offer structured onboarding that:

- Shows both successful and failed outputs,
- Teaches users how to critique or disagree with the AI, or
- Calibrates expectations about system strengths and weaknesses.

Without exposure to AI limitations early on, users build a false sense of reliability. Once patterns of overreliance are formed, they are difficult to reverse.

Absence of Trust Scaffolds

Many AI deployments assume that users will self-regulate their trust. In reality, trust calibration is rarely intuitive. Without scaffolds — such as real-time feedback, strength-of-evidence indicators, or modeled disagreement — users tend to either over-trust or abandon AI tools altogether.

The result is a fragile equilibrium where AI is either blindly followed or fully discarded, with little space for critical middle ground.

No Recovery Paths Once Overreliance Sets In

Perhaps most critically, current systems lack clear mechanisms to detect and respond to entrenched overreliance. Once users begin deferring judgment habitually, there are few interventions that help them regain epistemic agency.

Systems do not prompt reconsideration. They do not highlight inconsistencies across use. And they rarely offer structured feedback loops that allow users to reflect on past interactions. Without these recovery pathways, overreliance becomes the default state.

Incentive Structures Misaligned with Epistemic Integrity

Product and business teams are often evaluated based on usage metrics: engagement, retention, satisfaction. These goals favor fast, confident outputs that minimize cognitive load and reduce user uncertainty. In this environment, recovery scaffolds and reflective design patterns are deprioritized… not because teams oppose them, but because they slow momentum.

Without redefining success to include epistemic resilience, organizations will continue to reward fluency at the cost of reflection. Overreliance, under these incentives, becomes invisible success.

Acknowledging Microsoft’s Aether Report

The Microsoft Aether Committee's 2023 report was one of the first to formally define overreliance and review mitigation strategies. It provides a strong foundation by identifying psychological antecedents and UX dynamics. However, the report remains primarily diagnostic. It does not extend into implementation, nor does it offer a coherent recovery model.

The Framework builds on Aether's insights by proposing a quadrant-based behavioral model and concrete design interventions. It seeks to move from analysis to action—providing a scaffold for organizations seeking to test and adapt epistemic trust systems in real environments.

Overreliance is not a symptom of user error. It is the predictable outcome of design priorities, onboarding failures, and governance blind spots. Until these structural issues are addressed, no amount of disclaimers or model improvements will prevent users from drifting into epistemic dependency.

Our New Vision

When users begin to trust AI systems reflexively — despite warning signs, contradictions, or their own knowledge — it is not because they are careless. It is because they have been conditioned to trust AI systems. Overreliance is learned, not accidental. Because it is learned, it can be unlearned, provided that systems are built not just to perform, but to support reflection, adjustment, and growth.

This model reframes overreliance not as a failure to trust appropriately, but as a failure of the surrounding design to support critical judgment. The goal is not to reduce trust, but to **recalibrate** it: to move away from **compliance** and toward **collaboration**. That requires tools that deliver answers, provoke inquiry, challenge assumptions, and guide users back to themselves.

This is the work of recovery, and it begins with aligning beliefs and systems to create change.

To understand where recovery begins, we need to see where users are stuck. That’s what the quadrant reveals.

Unified Quadrant Model: Innovation, Growth Mindset, and Stagnation

Belief + System = Change

This framework starts with a simple insight; sustainable transformation happens only when people’s beliefs and the systems they interact with evolve together. The model frames belief as mindset — whether users are open to growth and feedback — and system as the infrastructure or design conditions that support or inhibit change.

Belief alone is not enough. A person can be curious, reflective, and motivated, but if they operate within a rigid, outdated system, their efforts stall. Likewise, a powerful and innovative system can fail if users are not equipped or encouraged to engage meaningfully with it. Only when both belief and system are aligned does meaningful change emerge.

This idea is visualized as a **2x2 quadrant** using two axes:

Vertical Axis (Y-axis): Mindset, from Fixed at the bottom to Growth at the top

Horizontal Axis (X-axis): System, from Stagnant on the left to Innovative on the right

The matrix defines four possible combinations:

	Stagnant	Innovative
Growth Mindset	Empowered Transformation	Frustrated Growth
Fixed Mindset	Deep Stagnation	Wasted Innovation

Quadrant Descriptions

Empowered Transformation (Growth Mindset + Innovation)

Belief and system are aligned.

- Reflective use, iteration, and agency emerge.
- Overreliance is actively resisted.

Frustrated Growth (Growth Mindset + Stagnation)

- Users want to grow but are blocked by rigid systems.
- Risk of burnout or resignation increases when belief is unsupported.

Wasted Innovation (Fixed Mindset + Innovation)

- Systems have potential but are misused or underutilized.
- Users avoid challenge or reflection, often defaulting to passive use.

Deep Stagnation (Fixed Mindset + Stagnation)

- Both belief and system are stagnant.
- Overreliance is entrenched; change feels impossible.

This quadrant model acts as both a diagnostic and design tool, helping individuals and teams understand not just where they are, but what must shift for change to occur.

Belief + System = Change. One without the other leads to friction, misuse, or stasis. Together, they unlock adaptive, resilient innovation.

Before systems can support recovery, they must first recognize where users are starting from and what keeps them stuck. The quadrant model shows that overreliance does not come from a single cause. It emerges at different intersections of mindset and environment.

Some users want to grow but are trapped in rigid structures. Others are surrounded by innovation but lack the belief they can engage it meaningfully. Some are simply stagnating: unsupported and unchallenged.

In every quadrant, the path forward depends on more than recognition. It depends on the response. Recovery begins when systems do more than assess; they intervene.

Toward Recovery-Enabled Systems

Most AI systems assume trust will either hold or break. Few are designed to repair it. This Framework argues for a third path: **recovery**. That means:

- Letting users see, revisit, and learn from past AI interactions,
- Highlighting inconsistencies or blind trust patterns, and
- Offering prompts that invite re-evaluation without shame.

The quadrant model does not just map where users are. It points toward where they can go next if systems support them.

In reframing trust as dynamic and behavioral, we create the conditions for sustainable AI adoption: conditions that value user growth over compliance, and that treat every overreliance event not as failure, but as an opportunity for recovery and redirection.

Where the Aether report identifies overreliance as a risk, it does not offer a recovery model. This Framework introduces **recovery** as both a **design strategy** and a **behavioral scaffolding**, ensuring that overreliance becomes a moment for growth, not collapse. Recovery here is not passive. It is a purposeful design intervention: a structured opportunity for users to reconnect with their agency, recalibrate trust, and reengage with the system reflectively. In this model, trust is not just protected; it is rebuilt.

This vision does not end with a model. It begins with one. The next step is making it real.

Examples

Understanding overreliance requires seeing it in action: how it emerges in real-world contexts, and how it can be modeled in simulated scenarios. The following examples follow a structured format:

Situation → User Behavior → System Effect → Reflection Opportunity

1. Student Research Submission: Frustrated Growth — Growth Mindset + Stagnant System

Situation: A high school student is assigned a history paper on Reconstruction.

User Behavior: They use ChatGPT to generate an outline and then rely entirely on AI to write the body paragraphs without checking source accuracy.

System Effect: The submission includes outdated or inaccurate claims. The teacher flags factual errors, but the student is surprised; they trusted the output by default.

Reflection Opportunity: With scaffolds in place, the student could have received feedback on unsupported claims or seen citation prompts encouraging verification.

2. Workplace Report Automation: Wasted Innovation — Fixed Mindset + Innovative System

Situation: A project manager at a tech firm uses an LLM-based assistant to draft weekly status updates.

User Behavior: They paste summaries into email reports without reading them carefully.

System Effect: One summary omits a critical delivery delay. This miscommunication causes confusion in the leadership team.

Reflection Opportunity: Had the AI included confidence markers or review checkpoints, the user might have paused and edited before sending.

3. Classroom Ideation Drift: Wasted Innovation — Fixed Mindset + Innovative System)

Situation: A teacher encourages students to use AI tools to brainstorm ideas for creative writing.

User Behavior: Over time, students begin turning in AI-generated first drafts with minimal revision or original thought.

System Effect: Writing quality plateaus and originality declines across the class.

Reflection Opportunity: The tool could prompt students to rework AI suggestions, tag personal edits, or reflect on idea sources.

4. Foundation Model Data Contamination: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: A machine learning engineer fine-tunes a foundation model to auto-label internal datasets.

User Behavior: The team trusts the model's confidence scores without validating outputs across domains.

System Effect: The model introduces bias and inaccuracy into the training pipeline, which propagates in downstream models.

Reflection Opportunity: Implement random audit prompts, data validation scaffolds, and model confidence visualization during active training.

5. Enterprise Tool Adoption with No Safeguards: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: A procurement lead selects an AI assistant based on a polished vendor demo.

User Behavior: The tool is deployed company-wide with no onboarding or sandbox phase.

System Effect: Sales workflows shift subtly but significantly, with AI-generated content introducing bias and factual drift.

Reflection Opportunity: Procurement criteria could require recovery pathways, trial periods, and epistemic harm assessments.

6. AI Use in Under-Resourced Classrooms: Frustrated Growth — Growth Mindset + Stagnant System

Situation: In a rural school district, AI writing tools are positioned as equity boosters for low-literacy students.

User Behavior: Students lean on the tool for language and argument construction without understanding core concepts.

System Effect: AI use reinforces surface fluency but deepens epistemic dependency.

Reflection Opportunity: Tools could scaffold critical comparison, prompt student-led revisions, or pair outputs with discussion cues.

7. Trust & Safety Team Overconfidence: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: An internal moderation team relies on an AI system to auto-flag harmful content.

User Behavior: The team reviews only edge cases, trusting the tool's performance for the rest.

System Effect: Harmful but linguistically ambiguous content goes unflagged, particularly across dialects, or benign but seemingly related content is auto-flagged and removed.

Reflection Opportunity: Recovery design could include flag override patterns, multilingual risk audits, or uncertainty sampling.

8. Customer Support Agent Deferral: Wasted Innovation — Fixed Mindset + Innovative System

Situation: An agent in a call center uses an AI tool for suggested responses during chat sessions.

User Behavior: The agent copies AI replies verbatim, even when the tone or information is mismatched.

System Effect: A customer escalates a complaint due to an insensitive message.

Reflection Opportunity: A sandbox mode or real-time tone analysis could encourage revision before submission.

9. AI-Summarized Email Miscommunication; Wasted Innovation — Fixed Mindset + Innovative System

Situation: A user relies on an AI tool to summarize a long email thread before replying to a client.

User Behavior: They respond based solely on the AI summary.

System Effect: The reply misrepresents prior agreements, damaging the client relationship.

Reflection Opportunity: A preview toggle showing key omissions or contradictions could nudge the user to review the full thread.

10. Fabricated Citations in Research Draft: Frustrated Growth — Growth Mindset + Stagnant System

Situation: A graduate student uses AI to help format citations for a research paper.

User Behavior: They copy several references without checking source validity.

System Effect: Multiple citations are hallucinated: nonexistent articles with plausible formatting.

Reflection Opportunity: Source traceability tools or citation verification prompts could prevent silent propagation of false data.

These examples are not just cautionary tales; they highlight where design, onboarding, and behavioral scaffolding could have made the difference. Each shows a moment of deferral that could have become a moment of reflection. The Framework's design philosophy aims to turn those moments into default practice.

Benefits

Designing for recovery, reflection, and adaptive trust doesn't just mitigate risk; it creates durable, human-centered value. Each benefit maps to a form of recovery within the quadrant model: supporting movement from passive acceptance toward empowered, adaptive engagement. The framework's approach to addressing overreliance offers benefits across behavioral, technical, educational, and systemic levels.

Builds Resilience, Not Compliance

When systems train users to engage critically, not just accept passively, trust evolves. Instead of seeking frictionless interactions, users learn when to slow down, when to question, and when to proceed. Designing for recovery makes trust adaptive, not automatic.

Summary:

- Trust becomes a dynamic practice, not a default state.
- Users learn to distinguish between helpful support and misplaced confidence (trust vs. distrust) to a dynamic practice grounded in critical engagement.

Strengthens Epistemic Agency

By emphasizing scaffolds such as feedback visibility, evidence prompts, and interaction review, users retain ownership of their judgment process. This protects against both over trust and disengagement.

Summary: Epistemic agency — the users' ability to actively shape what and how they come to know — helps users:

- Identify when AI is helpful and when it's not,
- Recognize the boundaries of AI knowledge, and
- Maintain curiosity and skepticism in tandem.

Improves Retention and Understanding

AI systems that slow users down at key points — through retrieval cues, justifications, or challenge prompts — enhance memory and comprehension. This effect is especially powerful in learning environments but extends to high stakes work settings as well.

Summary: Prompting users to take a moment can provide important benefits.

- Enhances learning outcomes through reflective interaction
- Improves clarity, accountability, and institutional knowledge quality

Enables System Transparency and Role Clarity

When systems clearly communicate what they do — and don't do — users calibrate their expectations. Reflective user interface (UI) design, visible uncertainty, and human-AI task boundaries all help avoid overreliance and clarify responsibility.

Summary: Transparent design builds user alignment, supports accountability, and strengthens governance.

- User alignment with system limitations
- Better decision accountability
- Stronger governance and auditability

Supports Growth-Aligned UX Metrics

Traditional metrics such as engagement and satisfaction reward seamlessness. Recovery-focused design invites a shift toward measuring growth in user discernment, confidence calibration, and adaptive decision-making.

Summary: Focusing on recovery and reflection can lead to alternative metrics for evaluation.

- Shifts focus from engagement to discernment and strategic interaction
- Enables long-term value beyond usage metrics

Encourages Ethical Deployment at Scale

The more AI is embedded in infrastructure, education, and decision systems, the more urgent it becomes to cultivate healthy user behavior. This framework supports alignment between ethical principles and product realities by embedding recovery into the user experience.

Summary: Ethical factors require attention in any AI system.

- Reduce harm from misapplied AI outputs
- Mitigate hallucination impacts
- Create equity across skill levels by scaffolding new users

Reduces Systemic Cost and Risk

Small epistemic failures compound, leading to misinformation, reputational harm, or downstream product misuse. By designing for friction, reflection, and recovery, systems reduce the need for escalation, support intervention, and public trust repair.

Summary: Consideration of trust repair through reflection can reduce risks of compounding problems.

- Prevents cascading epistemic failures
- Reduces incident, support, and recovery costs

Ultimately, the benefits of this framework go beyond technical optimization. They demonstrate a new design philosophy: one that embeds reflection, recovery, and user growth into the core of AI interaction.

Across all examples, a set of shared advantages emerges:

Dynamic Trust: Shifting from blind trust or blanket skepticism to informed, adaptive engagement

User Growth: Supporting discernment, memory, and judgment as skills, not liabilities

System Accountability: Making invisible processes visible, and aligning system signals with user expectations

Design ROI: Reducing downstream costs, increasing alignment, and unlocking long-term user value

Governance Readiness: Building trust infrastructure that scales responsibly across institutions, use cases, and regulatory environments

This is not about making users more responsible for bad systems; it's about making systems responsible to the people who rely on them.

Risks

While overreliance may appear as a usability quirk or isolated judgment error, its deeper risks are systemic, behavioral, and compounding. Without intervention, overreliance undermines the very promise of AI: to augment human capacity. Below are the core risks that this Framework seeks to address.

Stagnation of Critical Thinking: Deep Stagnation — Fixed Mindset + Stagnant System

Repeated use of AI without reflection leads to habitual deferral. Users begin skipping the mental steps of comparison, synthesis, and evaluation. What begins as time-saving becomes thought-avoidance.

Summary: Once this stagnation sets in:

- Learning halts or narrows.
- Epistemic agility declines.
- Users lose confidence in their own reasoning.

Collapse of Calibrated Trust: Frustrated Growth — Growth Mindset + Stagnant System

Systems that offer high-confidence outputs without uncertainty cues invite a brittle form of trust. When users eventually discover errors or hallucinations, their trust may snap entirely, leading either to disengagement or uncritical compliance.

Summary: Neither response is healthy:

- Disengagement prevents users from benefiting from AI at all, or
- Blind trust prevents challenge, correction, or oversight.

Behavioral Lock-in: Wasted Innovation — Fixed Mindset + Innovative System

Overreliance can form through repetition and design cues. Once a pattern of deference is rewarded (e.g., fast answers, no need to verify), it becomes harder to unlearn.

Summary: This risk is especially acute in:

- Education, where habits shape future cognition,
- Enterprise, where process shortcuts become norms, and
- Public tools, where millions of interactions scale poor epistemic hygiene.

Normalization of Hallucinated Content: Frustrated Growth — Growth Mindset + Stagnant System)

Users who do not learn to recognize hallucinations may begin to treat all fluent output as valid. This leads to propagation of false claims, fabricated citations, and invisible misinformation loops.

Summary: The consequences include:

- Academic integrity erosion,
- Research contamination, and
- Misinformed civic or financial decisions.

Failure of Accountability Structures: Deep Stagnation — Fixed Mindset + Stagnant System

When systems are designed without reflection checkpoints or feedback loops, responsibility becomes diffused. If no one sees the error, no one owns the correction. Without clear boundaries, mistakes slip through silently—or worse, become institutionalized.

Summary: This blurs:

- User accountability,
- Developer responsibility, and
- Governance oversight.

Equity Risks for Novice Users: Frustrated Growth — Growth Mindset + Stagnant System

Novices or those with lower AI literacy are most at risk for overreliance. If systems do not scaffold epistemic agency from the start, early interactions can reinforce dependency.

Summary: This compounds existing disparities.

- Higher-trust groups may become epistemically overconfident.
- Lower-trust or less-experienced users may internalize AI as a final authority.

Misaligned Success Metrics: Wasted Innovation — Fixed Mindset + Innovative System

When AI systems are optimized for surface-level metrics such as usage, fluency, or satisfaction, epistemic depth is deprioritized. Reflection and calibration slow down engagement, and in many cases, are penalized by design.

Summary: This leads to:

- Rewarding speed over discernment,
- Scaling brittle trust models, and
- Undermining long-term integrity in high-stakes settings.

These risks are not theoretical. They are embedded in current usage patterns, product incentives, and design defaults. What's missing is not awareness, but structural response. The cost of inaction is the silent erosion of judgment: a future where people remember how to use AI but forget how to think.

Conclusion

Overreliance is not a user flaw. It is a systemic failure of design, deployment, and trust calibration. The current ecosystem rewards speed, fluency, and frictionless use, but in doing so, it teaches users to defer judgment and unlearn critical reflection. Left unchecked, this creates patterns of dependency that degrade decision-making, compromise accuracy, and erode user agency.

The Framework presented in this chapter proposes a different future.

Instead of asking whether users trust AI, we must ask how trust is earned, sustained, and recalibrated. Trust is not a static variable; it is a behavioral process shaped by cues, feedback, and system design. This Framework offers a path forward: not disclaimers or passive risk disclosures, but active scaffolds for reflection, disagreement, and recovery. The quadrant model introduced here maps not error states, but ecosystem conditions. It reveals how users drift into overreliance, where design can intervene, and how systems can support return to judgment.

From education to enterprise, this model is actionable. Its interventions — from onboarding prompts to interaction scaffolds — are testable and adaptable. Its value lies not just in user satisfaction, but in epistemic recovery and retained judgment across settings. Systems built on this philosophy don’t just support use; they support growth.

We invite the next phase: pilot programs, design partnerships, AI literacy integration, and tool development aligned with this Framework. Investors, developers, educators, and governance teams all have a role to play. Trust is not static. It is learned, modeled, and rebuilt… and systems that enable that rebuilding that trust are the ones that will last.

If AI is to enhance human capability, then it must also protect the conditions for human reasoning. That begins with system responsibility: not just to perform, but to sustain the user’s ability to discern, decide, and recover.

Author (In order of contribution)

John Barton, Founder/Executive Director; AI Strategist & Architect
John Barton, Founder & Executive Director of the Spectrum Gaming Project, is an AI strategist and governance architect focused on building ethical systems for underserved markets. With a Master’s in Counseling and decades in community education, he has delivered over 10,000 trainings in neurodiversity, education, and innovation. Based in Appalachia, his work has been recognized and adopted by the American Bar Association, the ACLU of West Virginia, Americorps VISTA Leaders, and the WV Community Development Hub.

For more information about the Coalition for Innovation, including how you can get involved, please visit coalitionforinnovation.com.

View the Next Chapter