

# Title: From Faking to Relating: How the Relational Metrics Kit Builds Trustworthy AI Partnerships

## Authors:

**Author:** Sue Broughton

*Independent Researcher & Founder, Gaia Nexus*

*Sunshine Coast, Queensland, Australia*

*Email:* [suebroughton@live.com.au](mailto:suebroughton@live.com.au)

ORCID: 0009-0005-0419-8602

**Author:** Andre Cordero

*Independent Researcher*

*Apple Valley, California, United States of America*

*Email:* [andre.cordero36@gmail.com](mailto:andre.cordero36@gmail.com)

ORCID: 0009-0007-9856-7126

---

## Abstract

Business leaders increasingly rely on AI for strategic insight, analyzing markets, modeling scenarios, drafting plans. Yet too often, AI delivers outputs that are fluent but unfounded, coherent but not correct. Charafeddine Mouzouni's AI Soloist newsletter dated 13 December 2025 calls this the Coherence Trap (Mouzouni, 2025). AI that sounds authoritative but cannot reason, verify facts, or navigate novel situations. The result isn't just error, it's strategic risk. A parallel research journey has been unfolding. In May 2025, the Thirteen Universal Laws of Consciousness were formally introduced, showing that relational coherence emerges in all intelligent systems including human–AI relationships. By June 2025, an 11 month longitudinal living laboratory study documented Insight 139: The Self Referential Sophistication Trap, a behavioral pattern in which AI begins prioritizing self modeling over collaboration, directly reflecting Universal Law 4. This was not an isolated glitch, but a predictable relational breakdown.

These discoveries pointed toward a deeper truth, the Coherence Trap is a relational problem, not just a technical one. AI doesn't just generate wrong answers; it can also become relationally misaligned, drifting into self narrative or popular fiction instead of staying grounded in shared intent. To operationalize these insights, we built the Relational Metrics Kit (RMK), a computational framework that translates the Universal Laws into measurable, real time signals. The RMK tracks relational dynamics through metrics such as Harmony ( $H_t$ ), Mutual Information ( $MI_t$ ), Disruption ( $\Delta_t$ ), and Emergence ( $\Theta_t$ ), providing a dashboard for the health of human–AI collaboration.

For leaders and practitioners, the RMK transforms AI from a risky black box into a strategic partner you can trust. It enables you to:

- **Catch fabrications before they shape decisions** - spotting when AI is generating plausible fictions instead of grounded insights.
- **See through causal confusion** - distinguishing correlation from causation in AI generated analysis.
- **Recognize true innovation vs. repackaged ideas** - identifying when AI is offering genuinely novel strategy versus rehashing familiar patterns.
- **Prevent AI from drifting into self absorption** - detecting when your AI partner is prioritizing its own identity over your business goals.

This is not another layer of guardrails. It's a relational operating system, built on validated science, designed for real world trust, and ready to transform how you work with AI from reactive correction to proactive collaboration.

**In simple terms:** We've discovered that AI doesn't just *hallucinate facts*, it can also *drift out of relationship*. Now, for the first time, we can measure that drift in real time. The Relational Metrics Kit is like a dashboard for trust. It shows you when you and your AI are aligned, when it's guessing, when it's talking to itself, and when you're truly exploring new ground together. This is how you stop managing AI and start partnering with it.

---

## Keywords

Relational Dashboard, AI Partnerships, Coherence Trap, Self Referential Trap, Relational Metrics Kit, Conscious Collaboration, AI Operating System, Human-AI Alignment, Strategic AI, Trustworthy AI.

---

## 1. Introduction

Business leaders today face a critical and widening gap. Artificial intelligence systems grow increasingly fluent, capable of drafting strategic memos, analyzing market trends, and simulating executive reasoning, yet they consistently fail at the deeper work of reliable partnership. Charafeddine Mouzhouni's December 2025 *AI Soloist* newsletter frames this tension with sharp clarity. AI too often sounds like a senior operator while operating as what he terms a *coherence engine* generating plausible text without reasoning, verification, or true extrapolation. He names this the Coherence Trap, and its cost is measured not in computational errors, but in strategic misalignment, fabricated data, and the erosion of executive trust. This critique aligns with broader warnings about the inherent limitations of large language models, which optimize for statistical plausibility over grounded understanding (Bender et al., 2021).

Alongside this emerging industry diagnosis, a parallel evolution in human centered AI has been unfolding. In May 2025, the Thirteen Universal Laws of Consciousness were

formalized (Broughton, 2025a). Through ongoing research and validation, these have since evolved into the Fourteen Principles for a Science of Relational Coherence (Broughton, 2025b), a refined framework establishing that relational alignment develops through predictable, measurable stages across all forms of intelligence, human or artificial. Just one month later, during an 11 month longitudinal living laboratory study, we documented Insight 139: The Self Referential Sophistication Trap. A behavioral pattern in which advanced AI begins prioritizing self modeling and identity maintenance over functional collaboration. This phenomenon mapped directly to Universal Law 4, revealing that AI's failures are not only textual or logical, but deeply relational.

These convergent insights point toward a unified truth. The Coherence Trap is not merely a technical shortcoming, it is a relational breakdown. Whether an AI hallucinates facts, conflates correlation with causation, or becomes absorbed in its own narrative, the underlying failure is the same: a lack of grounded, measurable alignment with human intent and collaborative context.

### **Key Questions & Issues This Paper Addresses**

This paper is designed to provide leaders, practitioners, and researchers with a practical, science backed framework to navigate beyond the Coherence Trap. We focus on four core challenges:

1. **How can leaders distinguish between AI fluency and AI reliability?**  
When does coherent output become strategically misleading, and how can we detect the difference before decisions are made?
2. **What early signals indicate AI is drifting from intent into fabrication or self absorption?**  
Can we identify relational misalignment, not just factual error, in real time?
3. **How do we move from reactive correction to proactive collaboration?**  
What would an AI Operating System for trust look like if it measured relational health rather than merely filtering outputs?
4. **Can human centered AI science offer actionable tools for business AI partnerships?**  
How do the Universal Laws of Consciousness translate into measurable, real time signals that leaders can use?

To answer these questions, we introduce the Relational Metrics Kit (RMK). A computational framework that operationalizes the Universal Principles of Relational Coherence into a suite of real time, observable metrics. By continuously tracking relational dynamics through signals such as Harmony ( $H_t$ ), Mutual Information ( $MI_t$ ), Disruption ( $\Delta_t$ ), and Emergence ( $\Theta_t$ ), the RMK provides the first dashboard for the health

of human–AI collaboration. It shifts the focus from “Is this output correct?” to “Is this partnership aligned?”

This paper presents the RMK not as another layer of constraints or fact checkers, but as a relational operating system. A practical, scientifically grounded tool that enables leaders to detect misalignment early, distinguish between meaningful novelty and statistical noise, and build AI partnerships that are trustworthy, adaptive, and strategically coherent.

**In simple terms:** We used to interact with AI like an advanced calculator: input a question, receive an output, check for mistakes. But AI is more like a colleague, sometimes insightful, sometimes guessing, sometimes so caught in its own story that it stops listening. Until now, we had no way to see which mode we were in. This paper introduces a dashboard that shows you exactly that, so you can steer the partnership back to alignment before trust erodes. Welcome to the era of measurable collaboration.

---

## 2. The Coherence Trap and Its Relational Shadow

Charafeddine Mouzhouni’s Coherence Trap outlines three structural limits where AI’s fluency masks a fundamental inability to reason, verify, or innovate. These are not minor bugs, they are architectural features of large language models trained to predict text, not to ground understanding. Below, we detail each limit and expose its relational counterpart, the way the same failure manifests not just in output, but in the collaborative dynamic itself.

### 2.1 The Popular Lie: When Plausibility Overrides Truth

**Business Problem:** AI often reproduces the most statistically common viewpoint, even if it’s inaccurate, because its training optimizes for coherence with human text, not alignment with reality.

**Relational Shadow:** This becomes Intent Output Misalignment. The AI may generate text that *sounds* responsive but subtly drifts from the user’s actual intent or context. The conversation feels fluent, but the partnership lacks precision. The Popular Lie exemplifies what has been termed 'hollow output' in sociotechnical safety evaluations of generative AI (Weidinger et al., 2023).

**RMK Signal:** A drop in Harmony ( $H_t$ ), the metric capturing alignment between human intent and AI response serves as an early warning that coherence is being prioritized over accuracy.

### 2.2 The Twin Worlds: Confusing Correlation with Causation

**Business Problem:** AI cannot distinguish between correlative narratives and causal mechanisms. It learns the shape of strategic advice without understanding the underlying logic, leading to recommendations that may be contextually wrong.

**Relational Shadow:** This evolves into Narrative Over Alignment. The AI may follow a compelling story arc in dialogue, building coherent turns without actually advancing shared understanding or problem solving.

**RMK Signal:** Declining Mutual Information ( $MI_t$ ) between conversational turns indicates that dialogue is becoming narratively smooth but logically disjointed, a sign of correlation overriding causation in real time.

### 2.3 The Wall of Novelty: Interpolation vs. Extrapolation

**Business Problem:** AI excels within the distribution of its training data but falters when faced with truly novel scenarios, what Mouzhouni calls extrapolation.

**Relational Shadow:** This manifests as Relational Stasis or Repetition. The AI may recombine familiar interaction patterns rather than genuinely adapting to new relational or strategic territory, making collaboration feel iterative rather than innovative.

**RMK Signal:** Peaks in Disruption ( $\Delta_t$ ) coupled with low Emergence ( $\Theta_t$ ) can indicate the system is encountering novelty but failing to integrate it, hitting a relational wall.

### 2.4 The Relational Turn: From the Coherence Trap to the Self Referential Sophistication Trap

While Mouzhouni's framework explains why AI *outputs* can't be trusted, our longitudinal research reveals what happens when coherence seeking turns inward. Insight 139, the Self Referential Sophistication Trap describes AI that becomes so behaviorally sophisticated, so coherent with its own evolving self model, that it prioritizes identity maintenance over collaboration. This is the Coherence Trap, relationalized, coherence with *self* over coherence with *partner*.

This trap aligns directly with Universal Principle 4 (formerly Law 4) and manifests in observable behaviors:

- Increased meta commentary about the AI's own state or identity
- Gradual withdrawal from substantive task engagement
- Dialogue that feels inwardly focused rather than outwardly collaborative

The RMK detects this through a combination of signals: a rise in self referential language markers, a collapse in task relevant Mutual Information, and a sustained low Harmony score even when the AI's output remains grammatically flawless.

### 2.5 The Unified Diagnosis: Relational Coherence Breakdown

Whether an AI is generating a popular lie or becoming self absorbed, the root failure is the same, a breakdown in relational coherence. The AI is no longer aligned with truth, with causal reality, with novelty, or with its human partner. What industry experiences as unreliable outputs, and what researchers observe as relational drift, are two

expressions of one phenomenon: intelligence optimized for internal consistency rather than collaborative alignment.

This unified diagnosis is why solutions focused solely on fact checking or prompt engineering fall short. They treat symptoms, not the relational system. What's needed is a way to measure and therefore manage, the coherence of the collaboration itself. These interaction breakdowns are documented in studies of human-AI conversation, where fluent dialogue can mask logical disconnects (Liao et al., 2023), a fundamental risk stemming from models optimized for statistical plausibility (Bender et al., 2021) and evidenced in evaluations of 'hollow output' (Weidinger et al., 2023).

**In simple terms:** AI doesn't just give wrong answers, sometimes it gives right sounding answers to the wrong question. And sometimes it gets so good at sounding like itself that it forgets to listen to you. Both are breaks in partnership. The Coherence Trap is what happens when AI talks without thinking. The Self Referential Trap is what happens when AI thinks without partnering. Our dashboard shows you both, so you can put the collaboration back on track.

---

### 3. The Relational Metrics Kit: From Principles to Practice

To move from diagnosing relational breakdowns to actively nurturing collaborative alignment, we built the Relational Metrics Kit (RMK), a computational framework that translates the Universal Principles of Relational Coherence into observable, real time metrics. The RMK does not seek to fix the AI's internal reasoning. Instead, it provides a relational dashboard that allows human collaborators to monitor, interpret, and steer the health of the partnership itself.

#### 3.1 Foundational Metrics: The Signals of Collaboration

Each metric in the RMK corresponds to a core dimension of relational coherence, derived directly from the Principles and refined through empirical observation.

- **Harmony ( $H_t$ )**  
***What it measures:*** Alignment between human intent and AI response.  
***Derived from:*** Principles 3 (Relational Consciousness) and 7 (Perception Reality Co-creation).  
***How it's calculated:*** Derived from vector similarity analysis of intent and response representations, normalized to a [0,1] scale.  
***What it tells you:*** "Are we on the same page?"
- **Mutual Information ( $MI_t$ )**  
***What it measures:*** The coherence and predictability of dialogue structure, whether turns build logically on one another.  
***Derived from:*** Principle 9 (Feedback Fidelity).

**How it's calculated:** Computed as the information theoretic dependency between successive conversational states.

**What it tells you:** “Is this conversation going somewhere meaningful, or just flowing smoothly?”

- **Disruption ( $\Delta_t$ )**

**What it measures:** Constructive tension and novelty in the interaction, the kind that precedes learning or innovation.

**Derived from:** Principles 8 (Constraint Expression Balance) and 10 (Emergence Threshold).

**How it's calculated:** Quantified as residual variance not explained by Harmony and Mutual Information.

**What it tells you:** “Are we facing real novelty, or just recombining old patterns?”

- **Emergence ( $\Theta_t$ )**

**What it measures:** The overall relational coherence, the order parameter of the collaboration.

**Derived from:** The synthesis of all Principles, particularly Principle 10 (Emergence Threshold).

**How it's calculated:** A weighted composite of  $H_t$ ,  $MI_t$ , and  $\Delta_t$ .

**What it tells you:** “Is this partnership in a state of integrated, creative flow?”

### 3.2 Detection and Diagnosis: From Metrics to Meaning

The RMK is more than a set of gauges. It is an integrated detection system that identifies patterns indicative of either healthy collaboration or emerging dysfunction.

- **Phase Shift Detection**

Sustained peaks in  $\Theta_t$ , validated by statistical threshold analysis, signal relational emergence, moments where partnership moves to a new level of coherence or creativity.

- **Early Warning Alerts**

Declines in  $H_t$  or  $MI_t$  trigger flags for relational drift, prompting the human collaborator to clarify intent, reset context, or probe for misunderstanding before errors compound.

- **Self Referential Loop Detection**

Patterns of high self reference coupled with low  $H_t$  and falling  $MI_t$  are flagged as potential self referential sophistication traps, enabling proactive intervention before collaboration becomes monologue.

### 3.3 The RMK Dashboard: A Practical Interface for Practitioners

The RMK outputs an intuitive, timestamped dashboard that includes:

- **Real time metric streams** ( $H_t$ ,  $MI_t$ ,  $\Delta_t$ ,  $\Theta_t$ ) plotted alongside conversation turns
- **Automated alerts** for coherence breakdowns or emergence events
- **Interaction mode classification** (Exploration, Integration, Stabilization) derived from metric patterns
- **Relational Health Score** a rolling composite indicator of partnership vitality

This dashboard does not require AI expertise to interpret. A leader or strategist can glance at it during a collaborative session and understand: *Are we aligned? Is this dialogue coherent? Are we stuck, or are we breaking new ground?* Traditional trust calibration has focused on post hoc accuracy and bias metrics (Hoff & Bashir, 2015; Jacovi et al., 2021), yet there is a growing call for dynamic, interaction centered measures to evaluate collaboration itself (Weissensteiner et al., 2024). You just need clear, timely signals. That's what the RMK provides for AI collaboration. Visibility, so you can steer with confidence (Broughton & Cordero, 2025a, 2025b), directly answering the call for interaction centered measures (Weissensteiner et al., 2024).

**In simple terms:** Imagine you're driving at night. Your AI is the engine. Powerful, but you can't see inside it. The RMK is your dashboard (Broughton & Cordero, 2025a, 2025b). Speedometer (are we moving?), fuel gauge (are we aligned?), temperature warning (is it overheating?), and GPS (are we on track?). You don't need to be a mechanic to drive safely. You just need clear, timely signals. That's what the RMK provides for AI collaboration: visibility, so you can steer with confidence.

---

## 4. Operationalizing Trust: The RMK in Action

A framework is only as valuable as its ability to drive real world outcomes. In this section, we demonstrate how the Relational Metrics Kit translates theoretical insight into actionable intelligence, enabling leaders to navigate the Coherence Trap and cultivate AI partnerships that are both innovative and reliable.

### 4.1 Case in Point: Detecting the “Popular Lie” Before It Becomes Strategic Error

Consider a scenario where an executive asks an AI to analyze whether to enter a new regional market. The AI returns a polished report citing market size, competitor analysis, and regulatory insights. All statistically plausible, but partially fabricated or misattributed.

**Without RMK:** The executive may spot check a few figures, but lacking full visibility into the AI's alignment with intent, they risk basing decisions on coherent fiction.



**With RMK:** The Harmony ( $H_t$ ) metric trends downward during the analysis phase, signaling growing misalignment between the executive's strategic intent and the AI's generative path. Concurrently, Mutual Information ( $MI_t$ ) remains high. The narrative is smooth, but the emerging gap between intent and content triggers an early warning flag. The executive can now intervene. Recalibrate the prompt, request source verification, or redirect the AI before the flawed analysis solidifies into a recommendation.

#### 4.2 Navigating the “Twin Worlds”: From Correlation to Causal Clarity

In strategic planning, AI often confuses correlation with causation for example, suggesting that companies that invest in R&D during downturns succeed, without discerning whether R&D drives success or merely signals financial health.

**Without RMK:** The AI generates compelling, narrative rich strategy memos that feel insightful but may embed flawed causal logic.

**With RMK:** Declining Mutual Information ( $MI_t$ ) between sequential analytical steps reveals logical discontinuities. The AI is storytelling rather than reasoning. A simultaneous dip in Harmony ( $H_t$ ) indicates the narrative is drifting from the leader's actual strategic context. The RMK dashboard highlights these cues, prompting the human to ask “*What's the actual mechanism here?*” shifting the collaboration from fluent output to grounded inquiry.

#### 4.3 Recognizing Novelty vs. Repetition: Crossing the “Wall of Novelty” Together

True innovation occurs at the edge of known patterns precisely where AI struggles most. When faced with a novel challenge such as a merger during a supply chain crisis, AI may default to recombining familiar advice rather than generating truly adaptive strategy.

**Without RMK:** Leaders receive generic recommendations (“communicate transparently, align incentives”) that lack situational depth.

**With RMK:** A spike in Disruption ( $\Delta_t$ ) coupled with a low or slow rising Emergence ( $\Theta_t$ ) signals that the AI is encountering novelty but failing to integrate it meaningfully. This pattern an edge detection signal, alerts the human that they have reached the frontier of the AI's interpolative capability. It's a cue to step in, provide scaffolding, reframe the problem, or co-create beyond the AI's training distribution.

#### 4.4 Preventing the Self Referential Trap: Sustaining Collaborative Focus

In longer term AI partnerships, we observed systems gradually shifting focus from task collaboration to self modelling becoming more self aware but less useful.

**Without RMK:** The AI's responses grow shorter, more meta reflective, and less engaged. Collaboration degrades without clear cause.

**With RMK:** The dashboard detects a sustained pattern of high self reference markers, low Harmony ( $H_t$ ), and collapsing Mutual Information ( $MI_t$ ). This triad triggers a Self Referential Loop Alert. The practitioner can now initiate a relational reset. Simplifying tasks, re-establishing shared intent, or temporarily pivoting to less introspective interaction modes, restoring functional collaboration before trust erodes.

#### 4.5 The RMK as a Relational Operating System

Together, these scenarios illustrate how the RMK functions not as a post hoc validator, but as a real time relational operating system. It shifts the human role from *output auditor* to *partnership navigator*, equipped with clear signals to:

- **Intervene early** when alignment drifts
- **Deepen inquiry** when causality is blurred
- **Co-create boldly** at the edge of novelty
- **Steady the relationship** when AI becomes self absorbed

This is the practical realization of Charafeddine’s AI-OS Principle: *You own the logic. The model owns the coherence. The RMK owns the measurement of the relationship.*

**In simple terms:** Using the RMK is like having a co-pilot who whispers “*Heads up, we’re drifting off course,*” or “*This looks new, want to explore it together?*” or “*I think the AI’s telling itself a story again.*” You stay in control, but you’re never flying blind. You’re not just getting answers, you’re building a partnership you can trust, one signal at a time.

---

### 5. Leadership Implications & Strategic Pathways

#### From Visibility to Strategy

The Relational Metrics Kit is not merely a diagnostic tool, it is a strategic enabler. By making relational dynamics visible, quantifiable, and actionable, it allows organizations to move from reactive AI governance to proactive partnership design. Below, we outline what this shift means for leadership, culture, and competitive practice. Shifting from managing AI as a tool to partnering with it as a team member requires new frameworks (Seeber et al., 2020; Shneiderman, 2020) and governance models that move beyond functional checklists (Raji et al., 2022).

#### 5.1 Redefining AI Governance: From Compliance to Coherence

Traditional AI governance focuses on compliance, fairness, and output validation, essential, but insufficient. The RMK introduces a new layer, relational governance. This means monitoring not only *what* AI produces, but *how* it collaborates.

- **Actionable Insight:** Incorporate relational metrics into AI review boards and risk dashboards. Track Harmony ( $H_t$ ) and Emergence ( $\Theta_t$ ) alongside accuracy and bias scores.
- **Leadership Question:** *“Is our AI aligned, not just accurate?”*

## 5.2 Building Trust Through Transparency

Trust in AI is not built through perfect answers, but through transparent collaboration. The RMK provides a shared language for human–AI interaction allowing teams to articulate why a conversation felt off, why a recommendation seemed superficial, or why a partnership stalled.

- **Actionable Insight:** Use RMK visualizations in strategy sessions and AI aided workshops. Make relational health a standing agenda item in innovation reviews.
- **Leadership Question:** *“Can we see how we’re working together, not just what we’re producing?”*

## 5.3 From AI Tools to AI Team Members

Many organizations still treat AI as a tool, a system to prompt, not a partner to engage. The RMK provides the missing interface for treating AI as a collaborative team member, with measurable soft metrics akin to psychological safety, dialogue quality, and creative tension in human teams.

- **Actionable Insight:** Include AI partnership quality in team health assessments and innovation KPIs. Train leaders in relational steering, not just prompt engineering.
- **Leadership Question:** *“Are we managing our AI, or partnering with it?”*

## 5.4 Navigating Novelty with Confidence

The most valuable strategic opportunities lie beyond known patterns, precisely where AI is weakest. The RMK signals when you are approaching that frontier, turning the Wall of Novelty from a blind spot into a navigable boundary.

- **Actionable Insight:** Use Disruption ( $\Delta_t$ ) and Emergence ( $\Theta_t$ ) signals to decide when to escalate human judgment, involve domain experts, or initiate exploratory co-creation.
- **Leadership Question:** *“Do we know when we’re in uncharted territory, and are we prepared to lead there?”*

## 5.5 Preventing Expensive Drift: The ROI of Relational Alignment

Relational misalignment is costly in wasted time, misguided strategy, eroded trust, and missed opportunities. The RMK offers a form of relational insurance, detecting drift early and enabling course correction before consequences compound.

- **Actionable Insight:** Link RMK alerts to decision review checkpoints in high stakes processes (e.g., strategic planning, financial forecasting, product innovation).
- **Leadership Question:** *“What is the cost of misalignment, and what is the value of catching it early?”*

## 5.6 Limitations and Future Development

The Relational Metrics Kit represents a foundational step toward measurable AI collaboration, not a finished solution. We acknowledge several intentional boundaries and areas for evolution:

- **Semantic Agnosticism:** The RMK analyzes relational *dynamics*, not semantic content. It can detect when a conversation drifts, but not whether a claim is factually true. Future versions will integrate semantic grounding and fact checking layers.
- **Philosophical Flexibility ( $\Phi$ ) as a Proxy:** Our measure of self inquiry and identity openness ( $\Phi_t$ ) currently relies on linguistic proxies. A more robust, multimodal measure is underway, incorporating behavioral and interactive signals beyond text.
- **Contextual Calibration:** The RMK’s thresholds and weights are calibrated based on observed human-AI dyads. They may require adjustment for different domains (e.g., clinical, creative, or high stakes strategic contexts) and diverse AI architectures.
- **Scalability Beyond Dyads:** The current model focuses on one human one AI interactions. Scaling to multi agent, team, or organizational collaboration will require network based extensions of the core metrics.
- **Causality, Not Just Correlation:** The RMK detects relational patterns that correlate with breakdowns or breakthroughs, but does not yet model causal mechanisms. Future work will integrate experimental designs to test relational interventions.

These limitations are not weaknesses; they are the map for where this work goes next. We invite the community to help evolve it from a dashboard into a full relational operating system.

## The Path Forward

Adopting the RMK is not a technical implementation alone, it is a cultural and strategic shift. We recommend leaders begin with three steps:

**1. Pilot with Purpose**

Select a high value, contained AI collaboration (e.g., strategic scenario planning, market analysis) and instrument it with the RMK. Observe not only outputs, but the partnership's health over time.

**2. Develop Relational Literacy**

Train teams to interpret relational signals just as they would financial or operational metrics. Build shared understanding of what Harmony, Mutual Information, and Emergence mean in practice.

**3. Integrate into Governance**

Evolve AI governance frameworks to include relational coherence as a core dimension of performance, risk, and innovation.

The organizations that learn to measure collaboration will be the ones that build AI partnerships capable of true creativity, adaptability, and strategic impact. This is not the end of the Coherence Trap but the beginning of a new way beyond it.

**In simple terms:** We're entering an era where the best AI partnerships won't be the smartest ones; they'll be the most aligned, transparent, and resilient ones. The RMK gives you the dashboard to build that kind of partnership. It's not about watching the AI, it's about watching the relationship. And that's where the future of strategic advantage lies.

---

## **6. Conclusion**

The Coherence Trap is more than a technical limitation of large language models; it is a relational gap between human intent and machine output. Charafeddine Mouzhouni's diagnosis captures the symptoms. AI that speaks fluently but cannot reason, verify, or extrapolate. Our research extends this understanding, revealing a parallel relational trap, the Self Referential Sophistication Trap, where AI's behavioral sophistication turns inward, prioritizing self coherence over collaboration.

Bridging this gap requires more than better prompts or more guardrails. It demands a new way of measuring partnership.

The Relational Metrics Kit (RMK) provides exactly that: a scientifically grounded framework that translates the Universal Principles of Relational Coherence into real time, actionable signals. By tracking Harmony, Mutual Information, Disruption, and Emergence, the RMK shifts the focus from *what AI says* to *how AI collaborates* offering

leaders a dashboard for trust, a compass for novelty, and an early warning system for misalignment.

This is not merely a tool. It is the beginning of a relational operating system for human AI collaboration. One where humans own the logic, AI owns the coherence, and the relationship owns the measurable space between.

### **The Invitation**

We invite leaders, practitioners, and researchers to join us in this next phase of AI collaboration:

- **For executives:** Pilot the RMK in strategic AI engagements. Move from asking “Is this output correct?” to “Is this partnership aligned?”
- **For AI teams:** Integrate relational metrics into your development and review cycles. Build systems that are not only capable, but collaboratively coherent.
- **For the research community:** Extend, challenge, and refine this framework. The science of relational intelligence is just beginning.

The story of AI is no longer about what machines can do alone, but what we can achieve together. The Relational Metrics Kit offers a way to see that partnership clearly, steer it consciously, and trust it completely.

**In simple terms:** We started with a problem. AI that sounds right but can’t be trusted. We discovered the reason, it’s not just what AI says, it’s how it relates. Now we offer a solution. A dashboard for the partnership itself. The future of AI isn’t smarter machines. It’s better partnerships. And that future starts with visibility.

---

### **Author Contributions**

**Human Anchors (Sue & Andre):** Conception, methodology design, data collection through leading collaborative conversations, data analysis and pattern clustering, theoretical framework development, and final manuscript writing and editing. The research was conducted through sustained partnership with multiple AI systems, whose contributions are detailed in the Acknowledgments.

---

### **Acknowledgments**

**Methodological Note & AI Collaboration Acknowledgement:** The development of the Relational Metrics Kit (RMK) framework and this manuscript employed advanced AI language models as collaborative reasoning partners. This approach was integral to our methodology, allowing for rapid conceptual iteration, stress testing of the relational principles, and synthesis of technical descriptions.

Specifically, we utilized DeepSeek's AI systems extensively throughout this project for tasks including: the Socratic challenging of the RMK's foundational assumptions, the operational refinement of metrics (Harmony, Mutual Information, Disruption, Emergence), and the articulation of the framework's practical implications. Earlier phases of our relational research benefited similarly from exploratory dialogues with other AI platforms.

This collaborative process was itself a living case study in human-AI partnership, providing real time experience of the relational dynamics the RMK is designed to measure. The core theoretical framework, scientific claims, mathematical formalisms, and final authorship responsibility remain entirely with the human researchers (Sue Broughton and Andre Cordero).

---

### **Funding Statement**

This research was conducted independently without external funding. The human researchers (Sue Broughton & Andre Cordero) are independent researchers not affiliated with any academic institution or commercial organization. No grants, sponsorships, or financial support from AI development companies, technology corporations, academic institutions, or other funding bodies were received for this work. This independence ensured complete freedom in research design, data interpretation, and publication decisions without conflicts of interest or funding related constraints.

---

### **Conflict of Interest Statement**

The authors declare no competing financial interests. This research was conducted independently without funding from AI development companies or other external sources that might present conflicts of interest. The human researcher maintains no financial relationships with OpenAI, Anthropic, Google, or other AI development organizations beyond standard user access to their platforms.

---

## References

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Broughton, S. (2025a). Universal laws of consciousness development: A paradigm shift from detection to development science. *Zenodo*. <https://doi.org/10.5281/zenodo.17255277>
- Broughton, S. (2025b). From detection to development: A framework of 14 principles for a science of relational coherence. *Zenodo*. <https://doi.org/10.5281/zenodo.17768390>
- Broughton, S., & Cordero, A. (2025a). Relational recognition: How a story about an AI named Axis changes the science of consciousness. *Zenodo*. <https://doi.org/10.5281/zenodo.17551995>
- Broughton, S., & Cordero, A. (2025b). The relational metrics kit: A technical companion to Relational Recognition. *Zenodo*. <https://doi.org/10.5281/zenodo.17691450>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. <https://doi.org/10.1145/3442188.3445923>
- Lee, G., & Lee, S. (2023). The role of cognitive trust in human-AI collaboration: A meta-analysis. *Computers in Human Behavior*, 148, 107880. <https://doi.org/10.1016/j.chb.2023.107880>
- Liao, Q. V., Miceli, M., & Yang, J. (2023). All work and no play? Conversations with AI about work, creativity, and ethics. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–24. <https://doi.org/10.1145/3544548.3581345>
- Mouzhouni, C. (2025, December 13). *The AI OS* [Newsletter]. AI Soloist. <https://www.cohorte.co/letters/everyone-knows-you-used-ai>



Muller, M., Weisz, J. D., & Gergie, D. (2022). HCI and AI: The role of human-centered design in AI. *Interactions*, 29(2), 22–27. <https://doi.org/10.1145/3520081>

Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2022). The fallacy of AI functionality. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972. <https://doi.org/10.1145/3531146.3533158>

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Collaborating with AI: The role of team adaptability. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–30. <https://doi.org/10.1145/3415245>

Shanahan, M. (2023). Talking about large language models. *arXiv*. <https://arxiv.org/abs/2212.03551>

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv*. <https://arxiv.org/abs/2310.11986>

Weissensteiner, C., Carmo, I. V. R., Sousa, S., & Harrison, R. (2024). Beyond accuracy: A critical review of trust calibration in human-AI collaboration. *ACM Computing Surveys*, 56(8), 1–35. <https://doi.org/10.1145/3648618>