JUST TECH

Interrogative Reasoning and the Problem with the “Human in the Loop”

By: Jameila “Meme” Styles, Zameshia Williams, Jose Teran, and Sylvester Johnson

MOVEMENTS & MOBILIZATION

Styles, Jameila “Meme,” Zameshia Williams, Jose Teran, and Sylvester Johnson. "Interrogative Reasoning and the Problem with the 'Human in the Loop'." Just Tech. Social Science Research Council. June 10, 2026. DOI: https://doi.org/10.35650/JT.3099.d.2026.

The Theory of Interrogative Reasoning is a community-centered framework that reconceptualizes ethical AI as a function of power, belief, and timing rather than solely technical design. This essay represents a collaborative effort among scholars and practitioners working at the intersection of artificial intelligence, community-based research, and ethical governance to explain why the Theory of Interrogative Reasoning is needed now. Together, the contributors reflect a transdisciplinary approach that bridges research, technology, and community practice, advancing interrogative reasoning as both a theoretical framework and an applied methodology for ethical AI. Drawing on real-world application through Measure’s Communities in the Loop initiative with the city of Austin, this essay demonstrates how community-led interrogation can transform lived experience into actionable evidence that informs AI governance, system design, and public accountability.

When Technology Decides Who Is Believed

Jameila “Meme” (mi-mi) Styles

Artificial intelligence (AI) systems increasingly influence and determine how people move through the world, who receives public benefits, who is flagged as a risk, who gets hired, who is surveilled, and who is ignored. The growing demand for AI is increasingly burdened by the social, technological, and environmental consequences of the data centers required to sustain it. AI systems are often described by

its proponents as neutral, efficient, or objective. Yet, again and again, their outcomes mirror structural inequalities, reproducing and amplifying patterns of exclusion and discrimination that long predate the technologies themselves.[1]

Much of today's conversation about ethical AI focuses on technical remedies, such as improving data quality, increasing transparency, or inserting a "human in the loop." While these interventions can be useful, they often do not address the more fundamental question about *who* holds power over the system when harm appears, and *who* bears responsibility for addressing that harm.[2] *Who* is authorized to question the system? *Whose* knowledge counts as credible? *When* is intervention allowed to occur?

The Theory of Interrogative Reasoning (TIR) begins from the premise that ethical failure in AI is not primarily a technical problem. It is a problem of power, belief, and timing. TIR reframes ethical AI not as a checklist or compliance exercise, but as an ongoing, community-led practice of questioning systems before harm becomes normalized.[3] Systems do not change themselves. They are most effectively changed through collective action, especially by those closest to harm.

As founder and president of Measure over the past decade this approach has guided my work in community-led evaluation, data justice, and AI accountability. Across education, healthcare, public safety, and AI-enabled technologies, I have observed the same pattern of harm. Whether the system is institutional (such as schools, healthcare organizations, or criminal justice agencies) or technological (such as algorithms and AI tools), the communities most affected are often the first to recognize problems, yet the last to be heard or believed. They can identify harm clearly and early, often grounding their concerns in lived experience and a desire to prevent further harm to others. Yet their insights are frequently dismissed as anecdotal, emotional, or premature, up to when the harm becomes widespread. TIR emerged as a response to this recurring failure.

At its core, TIR asks three simple questions:

1. *Who* is close enough to the harm to recognize it early and name it clearly?
2. *Who* is believed when evidence of harm is presented?
3. *When* are those people, with the necessary access to shape, pause, or redirect the system, invited into the loop?

From these questions flow five core principles:

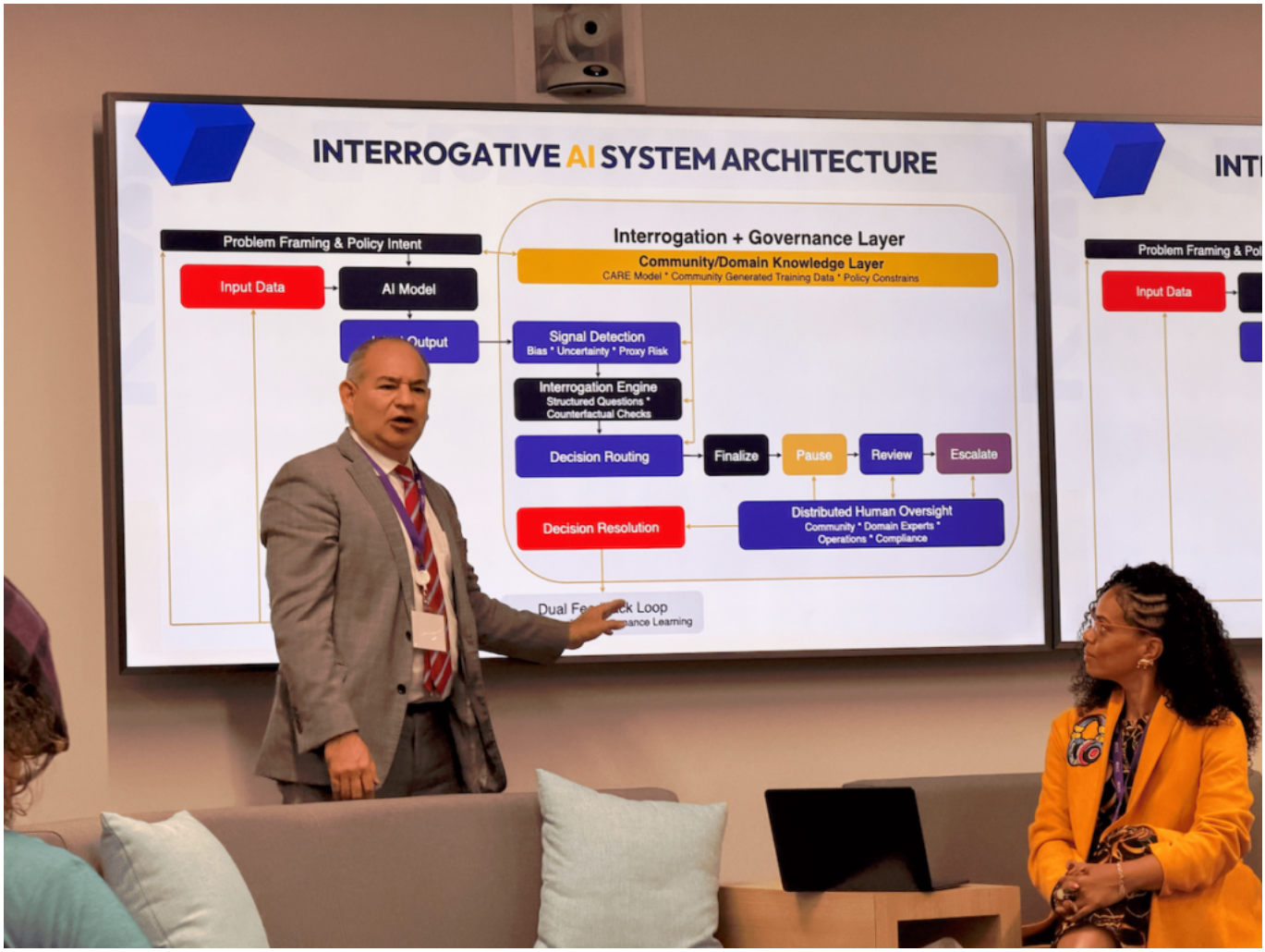
1. Proximity is critical for ethical insight. People closest to harm can first identify problems. Proximity is not anecdotal; it is a form of expertise that ethical systems must take seriously.[4]
2. Ethical accountability requires multiple *humans* in the loop. A single decision-maker, especially one distant from the social context of harm, cannot adequately interrogate complex, lived consequences. Accountability is strengthened through diverse and collective relational interrogation.
3. Interrogation is an ongoing practice, not a one-time corrective fix. Bias does not disappear once identified. It adapts as systems scale, contexts change, and incentives shift. Ethical engagement

must therefore be continuous.

4. Belief is a gatekeeper of influence. Communities can surface harm clearly and responsibly and still be ignored. Whether ethical feedback leads to change depends on whether institutions believe the people raising concerns.
5. Timing determines traction. Even accurate insights are ineffective if delivered too late or met with delayed response or nonaction. TIR treats timing as strategic, shaped by political, institutional, and cultural readiness.

Together, these principles shift ethical AI away from symbolic participation toward shared authority. Interrogative reasoning insists that communities should be co-designers of AI systems with the right to interrupt them before harm scales.

The sections that follow, written by technologists, academics, and pracademics in the field, extend this framework into practice, governance, and research, demonstrating how interrogative reasoning can and are already reshaping how AI systems are built, evaluated, and governed.



Jose Teran presenting the Theory of Interrogative Reasoning at Austin Community College. Photo by Meme Styles.

Interrogative Reasoning in Technical and Organizational AI Systems

Jose Teran

In technical AI environments, ethical failure cannot be explained by biased code alone. Far more potent is the problem of understanding the gap between those who define the system and those who live with its consequences.[5]

Engineering teams are trained to optimize measurable objectives: accuracy, latency, throughput, loss minimization. But many AI deployments operate in social domains, such as hiring, credit risk, education, policing, healthcare, where the core problem is not the accuracy of prediction(s) but the alignment of values. When teams misdefine the social problem, even technically excellent systems reproduce structural harm at scale.

The Theory of Interrogative Reasoning (TIR) addresses this upstream failure by embedding structured questioning into architecture, development, and governance as a design requirement.

At the architectural level, this means formalizing interrogative checkpoints before model training begins. Problem framing documents should require explicit articulation of: Who benefits from this optimization? Who bears disproportionate risk? What alternative definitions of success exist? These are not philosophical prompts; they shape feature selection, label construction, and evaluation metrics. When proximity-based expertise is absent during problem framing, datasets encode institutional assumptions that go unchallenged.

At the model development stage, TIR can be operationalized through participatory system dynamics and structured scenario testing with communities closest to potential harm. Rather than relying on aggregate fairness metrics, development cycles should incorporate qualitative signal channels, early-warning feedback from those directly affected, before deployment. This transforms the “human in the loop” from a reactive override mechanism into a distributed, collective interrogation process.

Within ethical AI, trust and safety, and product workflows, embedding TIR requires authority alignment. Community-informed insights must have defined escalation pathways capable of pausing or redirecting deployment. If interrogation cannot alter system incentives, it becomes symbolic. Governance frameworks should therefore codify trigger thresholds tied to both quantitative anomalies and community-reported harm signals.

Organizationally, this demands cross-functional interrogation forums, engineers, product leads, domain experts, and community representatives sharing the power to define what counts as valid evidence. Ethical review must shift from compliance checklists to structured deliberation among stakeholders about power, belief, and timing. Importantly, timing mechanisms should be built into release cycles, including staged rollouts with mandated interrogation intervals rather than postcrisis audits.

Technical excellence without interrogative reasoning produces scalable efficiency. Interrogative reasoning embedded in system design produces adaptive accountability. The difference is whether questioning occurs before harmful patterns become embedded in how systems operate or after it becomes statistically undeniable.

AI systems do not fail only because they miscalculate. They fail because their builders either fail to ask (or stop asking) the right questions, or they fail to ask them early enough. TIR restores questioning as infrastructure. This transition from technical architecture to social infrastructure highlights that the "right questions" are not just engineering requirements, but fundamental challenges to how authority is exercised and governed in a digital age.

Interrogative Reasoning, Power, and Public Accountability

Sylvester Johnson

The Theory of Interrogative Reasoning (TIR) is, at its foundation, a theory of power.[6] It asks not merely whether AI systems produce accurate outputs, but who holds the authority to define accuracy, who is deemed credible when harm is named, and whose belief is required before institutional action follows. These are not peripheral concerns. They are the central architecture of any serious effort to make AI systems accountable to the people they most affect.

Formal regulation has struggled to keep pace with this reality.[7] Legislative bodies in the United States and elsewhere have pursued frameworks for AI governance, yet the speed of AI deployment consistently outpaces the reach of law.[8] As states have introduced AI governance frameworks to address public concerns, recent federal actions have sought to preempt, challenge, or limit certain state-level regulatory efforts in favor of a more uniform national approach.[9] But even this point misses something more foundational. Regulation typically intervenes after systems are already embedded in consequential domains, such as hiring, benefits adjudication, law enforcement, education—where harm has frequently already scaled. This means the limits of formal regulation are not merely technical or procedural. They are also structural. Rules written at a distance from affected communities easily tend to encode the assumptions of those with institutional power, not those bearing the greatest institutional risk.

This is precisely why governance by other means—community-led accountability, participatory oversight, and what we might call democratic interrogation has become indispensable. Meme Styles has supported nonprofits, educational institutions, and civic organizations and has confirmed a pattern that TIR names with precision: Communities closest to algorithmic harm are often the most capable diagnosticians of that harm. They do not lack insight. They lack the institutional authority to make that insight actionable.

Interrogative reasoning reframes this as a governance imperative. Democratic accountability in the age

of human-machine systems cannot function if only technical experts are authorized to name problems or redirect systems. It requires distributed authority, structures in which proximity to harm confers standing, not suspicion. This is not an argument against technical expertise. It is an argument that technical expertise alone is insufficient to govern systems that are, at their core, social in nature.

What it means to govern human-machine systems in a way that preserves human agency is ultimately a question about whose humanity the system is designed to honor. Technology that concentrates interpretive authority will reproduce the inequalities already present in the institutions that built it. Interrogative reasoning insists that accountability must be practiced collectively, continuously, and with genuine power to interrupt—not merely to observe.



The Measure Team enjoying the HBCU AI CON 2026, which was held March 10–11, 2026, at Huston-Tillotson University.

Interrogative Reasoning as Research Practice and Evidence

Zameshia Williams

The methodological power of the Theory of Interrogative Reasoning (TIR) lies not in its conceptual architecture as an ethical framework, but in its active deployment as an applied research practice within an ongoing national study. Traditional artificial intelligence applications and evaluation models are structurally reactive, often relying on an isolated "human-in-the-loop" optimization framework where a single institutional actor monitors or overrides automated outputs. In practice, this individualized oversight model frequently struggles to catch or prevent systemic algorithmic harms before they are deployed against communities that have been historically pushed to the margins.

A stark example of this systemic vulnerability occurs when generative AI models are utilized to synthesize complex social problems or define community needs. For instance, when tasked with outlining the impact of humanitarian crises on vulnerable populations, generative systems routinely default to deficit framing, stereotyping, and cultural misreads.[10] The AI might generate a narrative that reduces impacted populations strictly to passive victims of violence or displacement. While the model may superficially acknowledge resilience, it simultaneously flattens structural realities and overlooks the deep, assets-based contextual expertise of the community.[11] In a traditional research layout, if a single institutional reviewer lacks the cultural proximity or structural mandate to challenge that output, this deficit framing becomes embedded into the foundational problem statement of a project, normalizing bias before data collection even begins. Bias is not introduced midway through research; it is normalized at the root.

TIR is hypothesized as an active methodological intervention designed to intercept these upfront algorithmic injuries by shifting the governance model from an isolated reviewer to a collaborative communities-in-the-loop framework. Rather than treating community observation as a retrospective narrative gathered after a system fails, TIR moves interrogation upstream, embedding it directly into the week-by-week mechanics of research design.

That testing is grounded in what Measure has already demonstrated at scale. The Communities in the Loop initiative, conducted in partnership with the city of Austin and the Austin AI Alliance, engaged approximately 409 residents across five convenings between 2025 and 2026. Participants were positioned not as survey respondents but as cogovernors whose proximity to harm gave them standing to define what oversight and accountability means. Their insights were synthesized into publicly released governance briefs that directly informed municipal AI policy recommendations, establishing that community-generated evidence can function as authoritative governance data rather than supplemental narrative.

Building on this foundation, Measure is now actively testing TIR as an embedded research methodology through its 2026 national Free Data Support program, funded by the Robert Wood Johnson Foundation. Through this initiative, thirteen community-led organizations were selected to participate in piloting TIR.[12] These organizations, many of which serve historically underresourced and marginalized global majority communities and that could not otherwise access these specialized assets, are being guided

through the twelve-step CARE Model mobilization process at no cost. What distinguishes this cohort from prior participants is the platform on which the work occurs. Measure has transitioned from static facilitation materials to a dynamic, virtual digital system with generative AI assistance integrated at every step. That shift is not incidental; it is the exact mechanism through which TIR is being tested as a living, unproven research practice.[13]

The live experimentation of TIR unfolds sequentially at each stage of the twelve-step CARE Model framework. Week by week, as community teams input and co-construct their project infrastructure, beginning with Step 1: The Problem Statement, the workflow operates through a cyclical, iterative process of generation and active interrogation:

1. **AI-Driven Enhancement:** The executive director, serving as the CARE team lead, and an allocated facilitator vet and refine their community-derived concepts through deep discussion. The embedded AI system then synthesizes the group's dialogue alongside external data to generate an enhanced, comprehensive research asset.
2. **Community-Led Interrogation:** Before the AI-generated asset is validated or moved forward, the CARE team is structurally prompted to interrogate the system's output using the embedded TIR digital tool. The team actively audits the response for specific harms, including stereotyping, deficit framing, cultural misreads, erasure, and hallucination.
3. **Core Dimension Screening:** The team subjects the AI's logic to targeted, platform-driven inquiries centered on the core dimensions of TIR, which include Proximity, asking whose voices should be centered and who might be missing; Belief, examining what assumptions or biases are embedded; and Timing, considering what historical context or evidence must be considered.
4. **Upfront Mitigation:** Based on the collective, proximity-driven feedback of the loop, the AI system corrects, refines, and restructures its output before it can scale into community planning or intervention design.

This layout transforms interrogative reasoning into active research infrastructure. The AI does not simply assist; it becomes the subject of structured community examination. The CARE team's lived experience and proximity to the problem they are addressing becomes the primary lens through which AI-generated content is evaluated, challenged, and corrected before it can shape downstream strategy, data collection, or community narrative. Harm is not identified after it scales; it is surfaced and addressed in real time, at every stage of the process, by the people most likely to recognize it.

By building upon the macrolevel municipal governance frameworks established in Measure's civic pilots, this phase of the study executes a transparent, disciplined research practice to evaluate several important methodological contributions. First, it operationalizes proximity as expertise, structurally positioning community teams as the interrogators of AI output whose corrections carry weight within the system itself. Second, it embeds interrogation continuously rather than as a one-time review, reflecting TIR's insistence that ethical accountability is an ongoing practice, not a static checkpoint. Third, it redistributes interpretive authority, shifting power away from a single institutional actor to a collective community team representing diverse perspectives. In this way, the initiative demonstrates what TIR looks like when it is not merely theorized but practiced; it serves as a structured, repeatable methodology

in which community-rooted interrogation is the mechanism through which AI systems are evaluated, corrected, and made more accountable to the communities they are meant to serve. The results are not yet complete, but the architecture for answering that question has been built, and it is operating now.

From Participation to Shared Authority

Interrogative reasoning, as coined by Meme Styles, challenges the assumption that ethical AI can be achieved through improved code alone. It argues for a redistribution of authority, one that centers proximity, belief, timing, and collective power. This emerging method is not only being coded into algorithmic study at Huston-Tillotson University, but it is intentionally centering communities across the nation through the Measure CARE model. In an era when AI increasingly governs everyday life, TIR insists on a fundamental shift mandating that ethics must be practiced *with* communities, not *on* their behalf.

Footnotes

1 Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity Press, 2019).

2 Andrew D. Selbst et al., "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019): 59–68.

3 Jaakko Hintikka, *Inquiry as Inquiry: A Logic of Scientific Discovery* (Springer, 1999); Arto Mutanen, "The Interrogative Model of Inquiry as a Logic of Scientific Reasoning," *Vestnik RUDN. Philosophy Series*, no. 3 (2011): 22–30, https://journals.rudn.ru/philosophy/article/view/11389.

4 Donna Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies* 14, no. 3 (1988): 575–599, https://doi.org/10.2307/3178066.

5 Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (The MIT Press, 2020).

6 Benjamin, *Race After Technology*; Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need* (The MIT Press, 2020).

7 Esmat Zaidan and Imad Antoine Ibrahim, "AI Governance in a Complex and Rapidly Changing World: A Global Perspective," *Humanities and Social Sciences Communications* 11 (2024), https://doi.org/10.1057/s41599-024-03560-x.

8 S. Matthew Liao et al., "Navigating the Complexities of AI and Digital Governance: The 5W1H Framework," *Journal of Responsible Technology* 23 (September 2025): 100127, https://doi.org/10.1016/j.jrt.2025.100127.

9 "National Policy Framework for Artificial Intelligence: Legislative Recommendations," The White House, Executive Office of the President, March 20, 2026, https://www.whitehouse.gov/wp-content/uploads/2026/03/03.20.26-National-Policy-Framework-for-Artificial-Intelligence-Legislative-Recommendations.pdf.

10 Benjamin, *Race After Technology*; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018).

11 Costanza-Chock, *Design Justice*.

12 Peter Reason and Hilary Bradbury, eds., *The SAGE Handbook of Action Research*, 2nd ed. (Sage, 2008).

13 Michael Quinn Patton, *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use* (Guilford Press, 2010).