THE
kyndryl
institute

Trend topic: **Artificial Integrity**

By

Hamilton Mann

Thinkers50 award winner, author of the award-winning book Artificial Integrity and the forthcoming Trias Algorithmica

Hamilton Mann is an AI researcher, Executive Chairman and President of the Artificial Integrity Institute, a nonprofit organization dedicated to advancing AI systems that prioritize integrity over intelligence. He is the best-selling author of Artificial Integrity: The Paths to Leading AI Toward a Human-Centered Future, named a Must-Read of 2024 and one of the Top 10 Essential AI Books of 2025 by the Next Big Idea Club, and selected as one of the Best New Management Books of 2025 by Thinkers50. He lectures at INSEAD and HEC Paris, serves as a mentor at the MIT Priscilla King Gray (PKG) Center and conducts doctoral research at École Nationale des Ponts et Chaussées – Institute Polytechnique de Paris. In 2024, he was inducted into the Thinkers50 Radar, and in 2025 he received the Thinkers50 Distinguished Achievement Award for Digital Thinking. His work appears in California Management Review, Stanford Social Innovation Review and Harvard Business Review, among others, and he regularly writes for Forbes. *Hamilton's latest book 'Trias Algorithmica: What Code Rules' (Wiley 2026) will be released in September.*

Explore more from The Kyndryl Institute
kyndryl.com/institute

The AI design challenge: focus on artificial integrity, not intelligence

There is a persistent belief that once a decision is expressed mathematically, it becomes neutral. It suggests discipline. It signals rigor. It gives the impression that once a problem has been translated into numbers, it is free from ambiguity.

In business, policy and increasingly in artificial intelligence, equations are often treated as arbiters of truth — tools that eliminate ambiguity and replace human bias with objective reasoning. Numbers, we assume, don't lie.

But this confidence rests on a fallacy.

An equation can be precise, consistent and scalable. That's what gives it authority. What it can't do, however, is decide what matters, what should be optimized or whose losses count as tolerable. Precision does not remove judgement, in other words. It just stabilizes it — it turns choice into a system.

This is the same fallacy now surrounding AI. We treat models as if they reveal neutral truths about the world, when in fact they operationalize prior decisions about what data matters, what patterns deserve weight and what outcomes should be rewarded.

The crucial misunderstanding is this: we tend to believe that models are 'discovered,' when in fact they are designed. They are, in other words, constructed versions of reality, shaped by assumptions, incentives and omissions. And once deployed, they simply enforce these versions of reality — regardless of their biases, flaws or blind spots.

This is the problem that Artificial Integrity seeks to address: making the value foundations of AI models harder to ignore.

Without safeguarding and prioritizing integrity over intelligence, optimization-first AI can deceive, turn to power-seeking, hack reward signals, indulge in sycophancy, push for homogenization and carry other profound security and adversarial risks that weaponize its power against human cognitive agency and dignity.

Artificial Integrity instead aims to advance AI systems capable not only of analytical performance, but also of preserving alignment with ethical, moral, social and contextual forms of reasoning. It shifts the focus from maximizing intelligence through pattern predictability and optimization toward the more fundamental challenge of maintaining integrity in how systems interpret, prioritize and operationalize human values.

What values sit behind the 'score'?

Imagine a telecommunications company deploying an AI system to manage customer support during a major network outage affecting thousands of customers. With only enough human agents to immediately handle 300 of the 2,000 incoming requests, the system must decide who receives direct assistance first, who is routed to automated support and who waits.

To make those decisions, every interaction is translated into variables: expected handling time (scored 1 to 10), customer lifetime value, churn risk, complaint severity and projected revenue impact. The model then compresses those inputs into a composite priority score that determines whose problem is treated as most urgent.

It is tempting to think that the problem now becomes technical. Rank the customers. Apply the model. Let the numbers decide.

But the numbers never decide.

What decides is **what we ask the numbers to do.**

Let's bring this to life:

Sarah, a long-time customer and single mother working remotely, has lost internet access during an important workday. Her predicted annual customer value is calculated at $1,200. Her churn probability is estimated at only 8% because she has remained loyal for years. Her issue is classified as technically complex and likely to require 45 minutes of human assistance.

Elena, an elderly customer whose medical monitoring equipment depends on network connectivity, generates only $480 in annual revenue. Her churn probability is low at 5%. Her issue may require over an hour of support because she struggles with digital interfaces.

David, a premium business subscriber managing multiple company accounts, generates $18,000 in annual revenue. His churn probability is estimated at 72% because he recently filed several complaints and compared competitor pricing online. His issue can likely be solved within 8 minutes.

If the company optimizes for operational efficiency and revenue protection, revenue value dominates. Churn probability follows. Handling speed plays a secondary role. The model assigns 50% weight to revenue value, 30% to churn probability and 20% to handling speed.

In this scenario, David receives a priority score of 8.9 out of 10 and is immediately routed to a senior human agent. Sarah receives 4.7 and is redirected into a delayed queue because her loyalty makes her statistically unlikely to leave. Elena receives 3.1 because her low revenue value and long handling time reduce her optimization score significantly.

The calculation is clean. Whichever customers are selected as priority out of the three, the outcome can always be made to appear defensible. The mathematics are coherent. The business protects revenue exposure while maximizing the number of resolved tickets per hour.

What if vulnerability mattered more than revenue?

But what if the same company decides to define success differently.

Now, those in the most immediate situational vulnerability are 'programmed' to come first. In this scenario, the model prioritizes exposure to harm and societal impact. The weights shift. Service dependency and human consequences become central. Churn probability recedes. Revenue value now counts for only 15% of the score. Vulnerability indicators and service dependency account for 60%. Handling efficiency falls to 10%. The model is recalculated. A different set of customers is prioritized.

Elena's score rises from 3.1 to 9.4 because medical dependency becomes mathematically central. Sarah rises to 7.8 because employment disruption is weighted as economically destabilizing for vulnerable households. David falls to 5.2 despite his financial value because his situation is no longer considered the highest priority.

Nothing about Sarah, Elena and David's data has changed. Nothing about the mathematics has weakened. Yet the outcome is no longer the same. It becomes entirely different because the organization has changed what the model is designed to value.

Change the values in question, and the equation follows.

Let's consider a third possibility. The same telecommunications company acknowledges that customer support indicators do not fully capture the broader consequences of unresolved customer situations. Metrics such as revenue value, churn probability and handling efficiency may accurately measure short-term operational performance while failing to account for reputational exposure, public amplification,

What emerges is a form of authority without real visibility. Decisions are made in ways that are technically rigorous but conceptually opaque.

customer influence or societal sensitivity. The model is therefore adjusted to incorporate the long-term organizational consequences of customer dissatisfaction and public perception.

As a result, the company is now optimizing for two very different metrics: trust and reputational resilience. The system analyzes which interactions are most likely to influence future public perception, customer advocacy and long-term loyalty. Customer sentiment volatility and reputational contagion are now heavily weighted. A new ranking emerges. Customers who were previously considered operationally low-priority according to traditional efficiency metrics become priority cases.

Sarah's score rises to 8.6 because frustrated remote workers statistically generate significant social amplification online. David receives 7.4 because of his business influence. Elena reaches 8.1 because unresolved cases involving vulnerable individuals create disproportionate reputational damage when publicly shared.

Once again, the mathematics performs flawlessly. It translates intention into decision. What changes is the world being modeled. Nothing else.

In all three cases, the system appears objective. But the model is not discovering what matters. It is operationalizing what the organization has chosen to prioritize.

The illusion of objectivity

This is the point we consistently fail to confront when thinking about AI.

There is no single objectively 'correct' AI model. Each model is internally coherent. Each can be justified. Each produces a different reality. And once a model is deployed, its reality becomes operational.

What is more, the more sophisticated the system becomes, the more effectively it conceals its own assumptions. In simple equations, one can still see the variables, the weights, the thresholds such as in both cases of the telecommunications company. In complex models, such as AI systems, these choices are distributed across layers of design.

What emerges is a form of authority without real visibility. Decisions are made in ways that are technically rigorous but conceptually opaque. It becomes easier to say that the system has "decided." Easier to treat outcomes as inevitable. Easier to mistake the execution of a model for the expression of reality.

But responsibility does not disappear behind computation. The more organizations rely on AI systems to structure decisions, the more accountable leaders become for the assumptions embedded within them.

A system optimized for productivity may normalize surveillance. A system optimized for efficiency may

erode human judgment. In each case, the model executes priorities that leadership has either explicitly designed or implicitly accepted.

This means responsibility cannot be reduced to whether a system performs accurately or efficiently. It requires organizations to remain answerable for the values translated into operational logic: what is being optimized, which trade-offs are being accepted, whose interests are amplified and who bears the cost of those decisions.

What we allow to matter

This is what I call 'matteration.' Matteration is the process through which human judgment determines what is allowed to matter before the first line of code begins constructing the systems that it will later optimize around it.

Some may think the risk is that machines will think in our place. The greater risk is that they will execute our unexamined matteration assumptions, freezing them into permanent, unarguable infrastructure so efficiently that we no longer see them.

And once we stop seeing them, we stop questioning them. We stop debating them. We stop deliberating them.

A hiring system optimized for similarity to previous top performers may progressively eliminate atypical profiles without anyone explicitly deciding to exclude them. A customer service system optimized for speed may normalize the systematic de-prioritization of people requiring more time, explanation, or human care.

Over time, these matteration assumptions stop appearing as visible choices. They become operational norms. Employees adapt to them. Organizations adapt to them. Societies restructure around them. Entire forms of human behavior become progressively less admissible simply because they are less legible to the system.

This is how optimization can invisibly reshape institutions, not by forcing people to obey machines, but by making machine-compatible behavior appear increasingly rational, legitimate and inevitable.

Keeping judgement visible

The real challenge is not to build more intelligent systems, but to build systems whose underlying mattering choices remain visible, interrogable and accountable.

Leaders must ensure that the humans deploying these systems remain capable of recognizing and assuming responsibility for the judgments embedded within their decisions.

In practice, this requires organizations to clearly define where human judgment must remain decisive rather than merely supervisory. It requires establishing both the right and the expectation for employees to question or override AI-generated outputs. Organizations must evaluate systems not only for performance, but also for their effects on autonomy, inclusion, critical thinking and organizational behavior.

They must prevent efficiency metrics from becoming the sole criteria of legitimacy and maintain traceability around how AI-informed decisions are made and operationalized. Employees must be trained not simply to use AI systems, but to understand their limitations, assumptions, blind spots and behavioral influence.

Organizations must also avoid cultures in which disagreement with algorithmic outputs becomes implicitly discouraged and continuously reassess whether these systems are reshaping human decision-making in ways that reduce reflection, accountability or cognitive agency.

Otherwise, organizations risk outsourcing judgment without realizing it while preserving only the appearance of human oversight. What is ultimately at stake is designing effective human and AI collaboration.

Designing for Artificial Integrity

For leaders, questioning the level of Artificial Integrity in current AI systems is a necessity if they want to avoid the progressive erosion of human judgment, agency and institutional responsibility.

As a relatively new concept and a frontier field of research, Artificial Integrity does not begin with a fixed checklist. But this does not mean that, with the current state of knowledge in AI governance, organizational responsibility, cognitive science, ethics and human-centered system design, there is nothing we can do now to identify, mitigate, and prevent artificial integrity gaps.

The challenge is to test and recognize integrity itself as a design and governance priority in whatever AI systems organizations intend to deploy rather than assuming that optimization, efficiency or predictive accuracy are sufficient proxies for responsible human outcomes.

Organizations can already begin by treating AI systems not merely as technical tools to improve efficiency, but as socio-technical systems capable of reshaping managerial decisions, employee behavior, customer relationships, workplace incentives and institutional culture.

Before deploying AI systems, organizations can evaluate whether employees remain genuinely empowered to challenge automated recommendations, whether critical decisions can still be meaningfully explained and contested, whether performance metrics indirectly penalize complexity or vulnerability and whether human judgment is being preserved or progressively reduced to the validation of algorithmic outputs.

They can also monitor whether AI systems create unintended behavioral adaptations over time, such as discouraging collaboration because it is harder to measure, incentivizing employees to optimize for metrics rather than substance, standardizing decision-making at the expense of contextual judgment or normalizing excessive dependence on automated recommendations.

While Artificial Integrity is still a nascent field of research, the need to uphold integrity in increasingly AI-mediated environments is already immediate, whether at the individual level to preserve human agency and cognitive autonomy, at the business level to ensure organizational accountability and responsible decision-making, or at the state level to protect democratic legitimacy and social stability.

Precision in any human setting is not achieved through mathematical accuracy alone, nor through optimization alone, but through integrity grounded in contextual, ethical, moral and social judgment.

And when algorithmic computation blurs one's capacity for reflection toward those dimensions, they become one of the most convincing illusions we have ever created, and one of the most significant forms of power that, if misused, can threaten human agency and further instrumentalize the social fabric of liberty.