California Management Review

ETHICS

Artificial Integrity Over Intelligence Is The New AI Frontier

by Hamilton Mann

Image Credit | na

The Only Code That Matters Is Integrity—Not Intelligence.

☑ INSIGHT | FRONTIER 19 May 2025

As AI systems increasingly take on critical roles across healthcare, education, transportation, finance, and public safety, relying solely on computational power and intelligence without embedding integrity into their design represents a major flaw.

RELATED CMR ARTICLES

Haenlein, Michael, and Andreas Kaplan, "**A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence**," California Management Review, 61/4 (2019): 5-14

On the path to advancing AI with integrity over intelligence, five critical characteristics—yet current limitations—come to mind as priority concerns.

Safety: While AI can quickly process data, it does not inherently consider whether its actions are safe, legal, or ethical. An illustration of this is the near-perfect execution of imitating a person's identity traits and characteristics, made possible by certain systems, without any verification, prevention, or restriction. This can lead to what we call deepfakes and severe consequences affecting individuals' reputations, privacy, or safety, and also lead to broader societal harms such as misinformation, manipulation in politics, and fraud.

Fairness: Some AI systems have taken steps to reduce harmful biases in their responses by training them on diverse datasets and continuously fine-tuning them to avoid producing unethical outputs. However, this is still an ongoing challenge. Even among the best image generation applications powered by GenAI, biases persist, such as when these tools suggest image modifications that reflect stereotypical or sexist cultural clichés, which can offend certain populations and perpetuate discriminatory biases.

Values: We can assume that an AI system that we use in our daily life has been designed to align with broadly accepted values and cultural settings. However, as its value system is shaped by its training data, it does not necessarily reflect cultural ethical norms. It does not “learn” values, culture and social norms dynamically after deployment in the way a system with integrity might. It may be updated periodically by its developers to improve its alignment with values, but it does not adapt autonomously to changing contexts. It lacks the autonomous reinforcement learning system where it could continuously learn and improve its behavior without human intervention.

Explainability: While some AI systems can explain certain processes or decisions, many AI systems cannot fully explain the decision-generating process (i.e., how they generate specific responses). Those based on Machine Learning, and even more so, those based on more complex models like deep learning are often opaque to users and operate as “black boxes.” While these systems may produce accurate results, users, those affected by the systems, and even developers often cannot fully explain how specific decisions or predictions are made. This lack of transparency can lead to several critical issues, particularly when they are used in sensitive areas such as healthcare, criminal justice, or finance.

Reliability: Some GenAI systems, such as ChatGPT, are designed to provide useful information, but true artificial integrity would involve a higher degree of consistency in ensuring that all information provided is reliable, verifiable with sources, and fully respects copyright of any kind, so as not to infringe on anyone’s intellectual property. AI with embedded integrity would analyze the data it processes and produce results that adhere to all relevant copyright laws, ensuring respect for creators and protecting against legal challenges.

All of these essential characteristics are related to a specific trait, which is not intelligence but integrity.

Without integrity embedded at its core, the risks and externalities posed by unchecked machine intelligence make them unsustainable, and render society even more vulnerable, even though they also bring positive aspects that coexist.

The excitement and rush towards AI is no excuse or tolerance for irresponsibility; it is quite the opposite.

The responsibility is to shift towards ensuring that AI systems operate with integrity over intelligence—safeguarding human values, and upholding societal imperatives over raw intelligence.

The question is not how intelligent AI can become, whether it involves calls for super artificial intelligence or artificial general intelligence. No amount of intelligence can replace integrity.

The question is how we can ensure AI exhibits Artificial Integrity—a built-in capacity to function with integrity, aligned with human values, and guided by principles that prioritize safety, fairness, values, explainability, culture and reliability, ensuring that its outputs and outcomes are integrity-led first, and intelligent second.

What Artificial Integrity Systems Are

The difference between intelligent-led and integrity-led machines is simple: the former are designed because we could, while the latter are designed because we should.

Without the capability to exhibit a form of integrity, AI would become a force whose the impact of evolution is inversely proportional to its necessary adherence to values and its crucial regard for human agency and well-being.

Just as it is not sheer engine power that grants autonomy to a car, nor to a plane, so it is not the mere increase of artificial intelligence that will guide the progress of AI.

This perspective highlights the need of AI systems to function considering the balance between “Human Value Added” and “AI Value Added” where the synergy between human and technology redefines the core design of our society, while preserving societal integrity.

Systems designed with this purpose will embody Artificial Integrity, emphasizing AI’s alignment with human-centered values.

A world predicated on Artificial Integrity would look vastly different from today, primarily because AI systems would be designed to prioritize not just intelligence and efficiency, but value models that ingrain, by design, the requirements of explainability, fairness, values, safety, and reliability in particular.

To systematically address the challenges of Artificial Integrity, organizations can adopt a framework I defined, structured around three pillars: the Society Values Model, the AI Core Model, and the Human and AI Co-Intelligence Model.

Each of these pillars reinforces each other and focuses on different aspects of integrity, from AI conception to real-world application.

The *Society Values Model* revolves around the core values and integrity-led standards that an AI system is expected to uphold. This model demands that organizations start to consider doing the following:

- Clearly define integrity principles that align with human rights, societal values, and sector-specific regulations to ensure that the AI's operation is always responsible, fair, and sustainable.
- Consider broader societal impacts, such as energy consumption and environmental sustainability, ensuring that AI systems are designed to operate efficiently and with minimal environmental footprint, while still maintaining integrity-led standards.
- Embed these values into AI design by incorporating integrity principles into the AI's objectives and decision-making logic, ensuring that the system reflects and upholds these values in all its operations while optimizing its behaviour in prioritizing value alignment over performance.
- Integrate autonomous auditing and self-monitoring mechanisms directly into the AI system, enabling real-time evaluation against integrity-led standards and automated generation of transparent reports that stakeholders can access to assess compliance, integrity, and sustainability.

This is about building the "Outer" perspective of the AI systems.

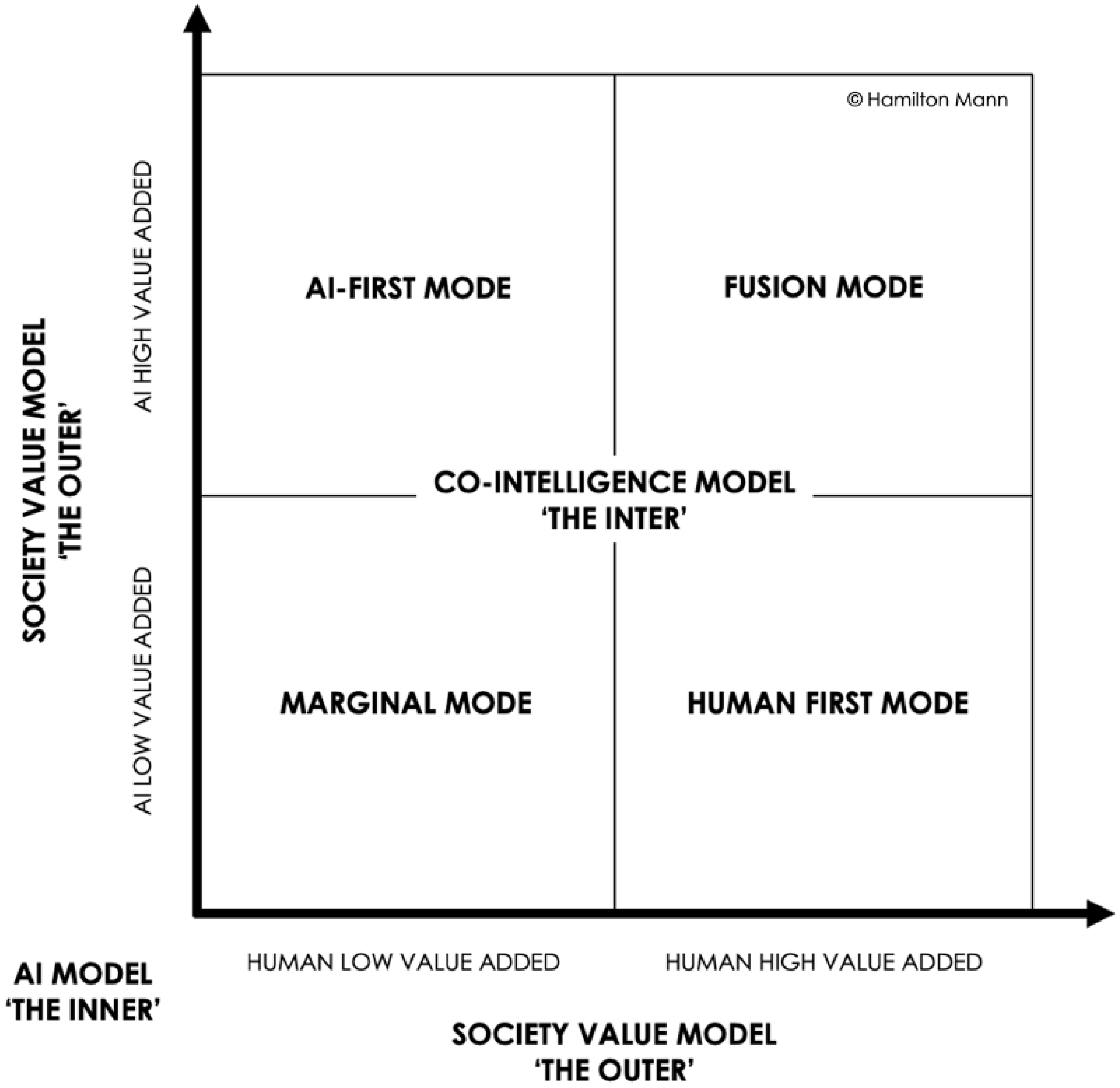
The *AI Core Model* addresses the design of built-in mechanisms that ensure safety, explicability, and transparency, upholding the accountability of the systems and improving their ability to safeguard against misuse over time. Key components may include:

- Implementing robust data governance frameworks that not only ensure data quality but also actively mitigate biases and ensure fairness across all training and operational phases of the AI system.
- Designing explainable and interpretable AI models that allow stakeholders, both technical and non-technical, to understand the AI's decision-making process, increasing trust and transparency.
- Establishing built-in safety mechanisms that actively prevent harmful use or misuse, such as the generation of unsafe content, unethical decisions, or bias amplification. These mechanisms should operate autonomously, detecting potential risks and blocking harmful outputs in real time.
- Creating adaptive learning frameworks where the AI is regularly retrained and updated to accommodate new data, address emerging integrity concerns, and continuously correct any biases or errors with regard to the value model that may occur over time.

This is about building the "*Inner*" perspective of the AI systems.

The *Human and AI Co-Intelligence Model* emphasizes the symbiotic relationship between humans and AI, highlighting the need of AI systems to function considering the balance between "Human Value Added" and "AI Value Added", where the synergy between human and technology redefines the core design of our society, while preserving societal integrity.

They would be able to function considering four distinct operating modes:



Marginal Mode: In the context of Artificial Integrity, Marginal Mode refers to situations where neither human input nor AI involvement adds meaningful value. These are tasks or processes that have become obsolete, overly routine, or inefficient to the point where they no longer contribute positively to an organization's or society's goals. In this mode, the priority is not about using AI to enhance human capabilities, but about identifying areas where both human and AI involvement has become useless.

One of the key roles of Artificial Integrity in Marginal Mode is the proactive detection of signals indicating when a process or task no longer contributes to the organization. For example, if a customer support system's workload drastically decreases due to automation or improved self-service options, AI could recognize the diminishing need for human involvement in that area, helping the organization to take action to prepare the workforce for more value-driven work.

AI-First Mode: Here, AI's strength in processing vast amounts of data with speed and accuracy takes precedence to the human contribution. Artificial Integrity would ensure that, even in these AI-dominated processes, integrity-led standards like fairness and cultural context are embedded. When Artificial Integrity prevails, an AI system that analyzes patient data to identify health trends would be able to explain how it arrives at its conclusions (e.g., a recommendation for early cancer screening), ensuring transparency. The system would also be designed to avoid bias—for example, by ensuring that the model considers diverse populations, ensuring that conclusions drawn from predominantly one demographic group don't lead to biased or unreliable medical advice.

Human-First Mode: This mode prioritizes human cognitive and emotional intelligence, with AI serving in a supportive role to assist human decision-making. Artificial Integrity ensures that AI systems here are designed to complement human judgment without overriding it, protecting humans from any form of interference with the healthy functioning of their cognition, such as avoiding influences that exploit vulnerabilities in our brain's reward system, which can lead to addiction.

In legal settings, AI can assist judges by analyzing previous case law, but should not replace a judge's moral and ethical reasoning. The AI system would need to ensure explainability, by showing how it arrived at its conclusions while adhering to cultural context and values that apply differently across regions or legal systems, while ensuring that human agency is not compromised regarding the decisions being made.

Fusion Mode: This is the mode where Artificial Integrity involves a synergy between human intelligence and AI capabilities, combining the best of both worlds.

In autonomous vehicles operating in *Fusion Mode*, AI would manage a vehicle's operations, such as speed, navigation, and obstacle avoidance, while human oversight, potentially through emerging technologies like brain-computer interfaces (BCIs) would offer real-time input on complex ethical dilemmas. For instance, in unavoidable crash situations, a BCI could enable direct communication between the human brain and AI, allowing ethical decision-making to occur in real time, blending AI’s precision with human moral reasoning. These kinds of advanced integrations between human and machine will require Artificial Integrity at its highest level of maturity. Artificial Integrity would ensure not only technical excellence but also ethical, moral, and social soundness, guarding against the potential exploitation or manipulation of neural data and prioritizing the preservation of human safety, autonomy, and agency.

Finally, Artificial Integrity systems would be able to perform in each mode, while transitioning from one mode to another, depending on the situation, the need, and the context in which they operate.

Considering the *Marginal Mode* (where limited AI contribution and human intelligence is required—think of it as “less is more”), *AI-First Mode* (where AI takes precedence over human intelligence), *Human-First Mode* (where human intelligence takes precedence over AI), and *Fusion Mode* (where a synergy between human intelligence and AI is required), the model *Human and AI Co-Intelligence* ensures that:

Human oversight remains central in all critical decision-making processes, with AI serving to complement human intelligence rather than replace it, especially in areas where ethical judgment and accountability are paramount.

- AI usage promotes responsible and integrity-driven behaviour, ensuring that its deployment is aligned with both organizational and societal values, fostering an environment where AI systems contribute positively without causing harm.
- AI usage establishes continuous feedback loops between human insights and AI learning, where these inform each other’s development. Human feedback enhances AI’s integrity-driven intelligence, while AI’s data-driven insights help refine human decision-making, leading to mutual improvement in performance and integrity-led outcomes.
- AI systems are able to perform in each mode, while transitioning from one mode to another, depending on the situation, the need, and the context in which they operate.

Reinforced by the cohesive functioning of the two previous models, the *Human and AI Co-Intelligence Model* reflects the “Inter” relations, dependencies, mediation, and connectedness between humans and AI systems.

This is the aim of Artificial Integrity.

Systems designed with this purpose will embody Artificial Integrity, emphasizing AI’s alignment with human-centered values.

This necessitates a holistic approach to AI development and deployment, considering not just AI’s capabilities but its impact on human and societal values. It is about building AI systems that are not only intelligent but also understand the broader implications of their actions.

An Essential Element in Building Such an Artificial Integrity Model Lies in the Data Process

Beyond labeling, which generally refers to the process of identifying and assigning a predefined category to a piece of data, it is necessary to adopt the practice of annotating datasets in a systematic manner. While labeling data gives it a form of identification so that the system can recognize it, annotating allows for the addition of more detailed and extensive information than simple labeling. Data annotation gives the data a form of abstract meaning so that the system can somehow contextualize the information.

Including annotations that characterize an integrity code, reflecting values, integral judgments regarding these values, principles underlying them, or outcomes to be considered inappropriate relative to a given value model, is a promising approach to train AI not only to be intelligent but also capable of producing results guided by integrity to a given value model. For example, in a dataset used to train an AI customer service chatbot, annotations could include evaluations on integrity with respect to the value model referenced, ensuring that the chatbot's responses will be based on politeness, respect, and fairness. Training data could also include annotations about ethical decision-making in critical scenarios, or ensure data is used ethically, respecting privacy and consent.

Another essential element for an AI model capable of displaying features of artificial integrity lies in the training methods. AI trained using supervised learning techniques that allow the model to learn not only to perform a task but also to recognize integrity-led and preferred outcomes is a promising path for the development of artificial integrity. It is also conceivable to add information about the value model used to train a given AI model through data annotations and then use supervised learning to help the AI model understand what does and does not fit the value model. For example, regarding AI models that can be used to create deepfakes, the ability to help the system understand that certain uses indicate deep faking and do not match the value models would demonstrate artificial integrity.

Another complementary approach is to design systems where human feedback is integrated directly into the AI model learning process through reinforcement learning methods. This could involve humans reviewing and adjusting the AI's decisions, effectively training the AI model on more nuanced aspects of human values that are difficult to capture with data and annotations alone. Especially when it comes to global AI models, thus used in many countries around the world, users across these different countries should have the opportunity to express their feedback on whether the model aligns with their values so the AI system can continue to learn how to adapt to the different value models they impact.

Building AI systems with Artificial Integrity presents several challenges that must be carefully addressed to ensure they operate ethically and responsibly. One major difficulty is the subjectivity of values—different cultures, communities, and individuals may have varying perspectives on what constitutes ethical behaviour.

Moreover, scalability poses another challenge. Annotating large datasets with detailed integrity codes requires significant resources, both in terms of time and human expertise, and may not always be feasible in practice. This process can be further complicated by the risk of bias introduction—the annotators themselves may unintentionally embed their own biases into the AI system, leading to skewed or discriminatory outcomes.

To overcome these issues, it is critical that AI systems are designed with mechanisms for continuous learning and adaptation. AI models equipped with Artificial Integrity must evolve alongside shifting ethical standards and societal values, which can be achieved through ongoing human feedback loops and dynamic updates to the annotated data. This could allow the system to recalibrate its decisions as cultural contexts or ethical norms change over time.

One of the Most Pressing Design Challenges of Our Time

Artificial Integrity is unattainable by AI developers working in isolation—ethicists, sociologists, public policy-makers, domain experts, diverse user groups and more must be involved from the outset to ensure a comprehensive approach that reflects a range of perspectives. This collaborative effort is essential for creating AI systems that are not only technically advanced but also grounded in a well-rounded, integrity-driven foundation.

Overall, this is a subject that requires more researchers to build AI that upholds human values over the pursuit of performance for the sake of performance.

Warren Buffet famously said, ‘In looking for people to hire, look for three qualities: integrity, intelligence, and energy. And if they don’t have the first, the other two will kill you.’ This principle equally applies to AI systems.

How to prevent such systems to be used to generate propaganda or manipulate public opinion on a large scale, as this could destabilize political and social systems even more than we see today?

How to protect people from becoming overly reliant on AI for critical thinking and decision-making as this can result in diminished human judgment and expertise in areas like education, law, and even healthcare, where the value of human intuition, empathy, and ethical reasoning are critical (if not irreplaceable)?

How to ensure the training process of AI does not lead to unintended privacy violations, particularly when AI systems begin to interact with sensitive data at scale?

How to mitigate AI environmental costs such as increased water consumption, CO2 emissions, rare earth mineral extraction for hardware production and the exacerbation of e-waste?

These are some of the questions that need to guide AI model design, prioritizing Artificial Integrity over Intelligence, which therefore aligns with the societal model we envision for the future.

Hamilton Mann Follow

Hamilton Mann is Group VP of Digital at Thales and lecturer at INSEAD and HEC Paris. He is a globally recognized expert in AI for Good and was inducted into the Thinkers50 Radar as one of the Top 30 most prominent rising business thinkers. Mann is the author of "Artificial Integrity" (Wiley).