Analysis of

Drew McDermott's "Artificial intelligence meets natural stupidity" (1976)

and its contemporary resonance in the co-superintelligence illusion

Tor-Ståle Hansen | 11. December 2025

McDermott's 1976 essay has typically been read as a humorous reprimand of early AI researchers. When reassessed through the full CIITR framework – *Cognitive Integration and Information Transfer Relation*, - the essay instead emerges as an early recognition of the structural error that continues to shape contemporary AI discourse, including the recent institutional narrative of "co-superintelligence." McDermott's critique identifies precisely the forms of conceptual confusion that CIITR later formalises as the boundary between syntactic performance and genuine comprehension.

Three themes in McDermott's argument align tightly with the structural principles of CIITR:

(1) the belief that naming a procedure after a cognitive act gives the system that cognitive property;

(2) the conflation of surface representation with semantic grounding; and

(3) the persistent inability to distinguish intelligent behaviour from the appearance of intelligence.

These themes echo CIITR's division between syntactic integration and rhythmic reach. McDermott saw the rhetorical misuse; CIITR explains the structural impossibility that lies beneath it.

Wishful Mnemonics as Inflated Syntactic Confidence

McDermott's leading criticism concerns the use of seductive program names such as "UNDERSTAND," "GOAL," or "General Problem Solver," which he saw as a form of self-deception. As he noted, the practice "may mislead a lot of people, most prominently himself," referring to the researcher who labels a routine with a cognitive term that the system does not structurally implement.

This corresponds to overestimating the significance of structural complexity. No matter how intricate the program's internal organisation becomes, the absence of temporal self-maintenance prevents any form of comprehension. McDermott recognised the psychological tendency to attribute capabilities that had not been earned. Theories shows why these capabilities cannot be earned by such systems in the first place.

The rhetorical gesture of naming becomes a projection error: the label gives the illusion of cognitive depth, but the architecture remains fundamentally non-cognitive.

Natural Language as a Misleading Starting Point

McDermott criticised the tendency to treat natural-language questions as if they provided a natural computational structure. He observed that researchers "trick ourselves into thinking that the statement of a problem in natural language is natural," despite the fact that the system cannot make use of what gives language its meaning in human contexts.

CIITR clarifies this failing. Human language is rhythm-bearing because human cognition sustains temporal continuity and internal re-entry of thought. Symbolic programs of McDermott's era, and modern transformer systems today, do not. They treat language as static symbolic input, disconnected from any lived temporal structure that would allow the symbols to become meaningfully integrated.

Thus a persistent error emerges: the assumption that because text appears meaningful to humans, it must be meaningful to the machine that manipulates it. McDermott saw the conceptual flaw; CIITR identifies the structural reason the system cannot bridge the gap.

The Fiction of the "Next Version" and Imagined Emergence

McDermott mocked the pattern in which researchers, having built a flawed Version I, would write lengthy descriptions of a hypothetical Version II that conveniently solved all theoretical problems. As he wrote, researchers behave as if "having identified the shortcomings of Version I is equivalent to having written Version II."

CIITR generalises this insight. The imagined Version II is structurally unobtainable for architectures that rearrange internal representations without ever developing temporal self-maintenance. No amount of complexity, optimisation, or system refinement will create a transition from syntactic manipulation to comprehension. The belief that such a transition lies ahead is an illusion rooted in misunderstanding of what comprehension requires.

Here McDermott's critique becomes strikingly contemporary: large-scale modern models exhibit the same projection error. The field continues to expect that with the next increase in parameter count or the next cycle of optimisation, systems will suddenly cross from simulation to understanding. CIITR shows why this expectation cannot be fulfilled.

The Projection of Understanding Onto Behaviour

In one of the paper's most perceptive passages, McDermott noted that people inevitably interpret machine behaviour in human terms:
"We interpret responses in a manner which may indeed allow (us) to conclude that (we) are being 'understood'..."

He argued that the distinction between *appearing* to understand and *truly* understanding was scientifically unclear. CIITR closes this gap by showing that the distinction is not behavioural

but structural. Systems that lack rhythmic continuity cannot form or sustain internal states that qualify as comprehension.

McDermott identified the interpretive fallacy. CIITR provides the ontological boundary that makes the fallacy inevitable.

Structural Continuity With Today's Co-Superintelligence Narrative

The patterns McDermott criticised in 1976 are reproduced today with greater rhetorical sophistication in corporate narratives such as the "AI-human co-superintelligence" proposal. In both eras, three errors recur:

1. **Cognitive vocabulary is used to describe non-cognitive systems.**
   McDermott: "GOAL," "UNDERSTAND," "CONTEXT."
   Modern discourse: "co-evolution," "joint intelligence," "mutual reasoning."
   In both cases, linguistic elevation obscures structural absence.

2. **Syntactic refinement is mistaken for cognitive development.**
   McDermott noted the illusion around Version II.
   Modern narratives assume that scaling and feedback loops cause understanding to emerge.
   CIITR demonstrates that these processes do not address the missing substrate.

3. **Human cognition is projected into the system's behaviour.**
   In both past and present, humans supply the temporal and epistemic structure that the system lacks.
   The institution or researcher then interprets the combined output as shared intelligence.

The Co-Superintelligence Illusion makes explicit what McDermott implicitly sensed: the human completes the epistemic loop, while the machine provides only syntactic elaboration. The appearance of intelligence arises from this coupling, not from any intrinsic capability of the system.

What McDermott Observed, CIITR Explains

Across his critique, McDermott diagnosed the behavioural and conceptual errors of researchers who mistook syntactic systems for cognitive ones. CIITR formalises the structural conditions that explain why these errors occur and why they persist. The systems McDermott criticised were structurally incapable of understanding, not because they were early or primitive, but because they lacked the architectural properties required for comprehension.

Modern large-scale systems, despite their vast representational richness, operate under the same structural limitation. McDermott witnessed the beginning of the illusion; CIITR demonstrates why the illusion cannot be dispelled without altering the architecture of artificial systems at the deepest level.

THE HISTORICAL CONSTELLATION OF THE ILLUSION: FROM EARLY AI CRITIQUES TO CONTEMPORARY RDF ANALYSIS AND THE CIITR FRAMEWORK

The structural patterns identified in McDermott's 1976 critique are not isolated anomalies within the history of artificial intelligence. They belong to a long-standing epistemic cycle in which conceptual overreach, institutional optimism, and rhetorical inflation repeatedly obscure the architectural limits of syntactic systems. When analysed through CIITR, this cycle materialises as an invariant dynamic: representational elaboration is mistaken for cognitive emergence, and behavioural fluency is misinterpreted as comprehension. The phenomenon is not restricted to any one period; it recurs whenever syntactic machines are presented as cognitive agents.

A survey of pre-1980s critical documents demonstrates that concerns about structural hollowness, conceptual exaggeration, and unjustified claims of intelligence have been present since the field's early decades. The following table summarises the most influential interventions, each of which challenged dominant assumptions about the nature and limits of artificial intelligence. Seen together, they form the historical backdrop against which both recent AI enthusiasm and contemporary structural analyses can be understood.

Historical Critiques of The AI Hype

Year	Author / title	Core argument	Why it still cuts deep today	Triggered consequences
1966	**ALPAC Report – *Languages and machines: computers in translation and linguistics* (National Academy of Sciences, USA)**	After 12 years and substantial investment, machine translation proved slower and less accurate than human translation. Expectations had far exceeded empirical performance.	Identified the first major gap between demonstration-driven enthusiasm and real capability, mirroring current exaggerations around "reasoning" benchmarks.	U.S. government cut most machine-translation funding for a decade.
1969	**Marvin Minsky & Seymour Papert – *Perceptrons***	Single-layer neural networks, promoted as "brain-like," were shown to be structurally incapable of solving basic problems such as XOR. Deeper architectures were theoretically known but practically unusable at the time.	Demonstrated that scaling a structurally inadequate architecture merely produces a larger version of the same inadequacy, a pattern that reappears in later eras.	Neural-network research declined sharply until the mid-1980s.
1972	**Hubert L. Dreyfus – *What Computers Can't Do* (building on the 1965 RAND memorandum)**	Symbolic AI assumed human intelligence could be reduced to explicit rule-following. Dreyfus argued that human understanding arises from embodied, intuitive, temporally continuous forms of	Anticipated later distinctions between static syntactic manipulation and temporally sustained cognition. Identified lack	Became the foundational philosophical critique of early AI research.

		coping that symbolic systems cannot replicate.	of situatedness as a structural limit.	
1973	**Sir James Lighthill – *Artificial Intelligence: A General Survey* (UK Science Research Council)**	Most AI promises collapsed under combinatorial explosion. Demonstrations failed to scale into functioning systems. Recommended ending most basic AI research.	Demonstrated that operational constraints cannot be bypassed by conceptual ambition. Directly relevant to contemporary scaling narratives in AI.	Led to the second major AI winter; UK AI funding nearly eliminated.
1976	**Drew McDermott – "Artificial Intelligence Meets Natural Stupidity" (SIGART Bulletin)**	AI research relied on "wishful mnemonics", naming routines UNDERSTAND or BELIEVE to create an illusion of cognitive depth. Exposed the gap between terminology and structural capability.	An insider's dismantling of rhetorical inflation. Still applicable to claims of emergent reasoning in large contemporary models.	Became one of the most frequently cited sceptical interventions in the field.

These documents, taken together, reveal a long-standing recognition of the structural gap between syntactic manipulation and cognitive agency. They identified recurring symptoms: optimistic vocabulary masking shallow functionality, exaggerated demonstrations interpreted as general intelligence, and the belief that greater representational complexity would eventually lead to understanding. These critiques thus constitute the historical antecedents to contemporary concerns about AI over-interpretation.

A similar pattern re-emerged in analyses published in 2023, where the concept of the "Reality Distortion Field" was used to describe how technological enthusiasm routinely outpaces grounded evaluation. This analysis argued that modern AI systems, such as customer-service chatbots, frequently produce an illusion of efficiency while remaining brittle, superficial, and structurally incapable of genuine comprehension. Ambitious claims about deep learning were contrasted with the practical constraints of behavioural mimicry, emphasising the difference between theoretical potential and operational reality. The argument ultimately warned against the assumption that technological novelty inherently produces societal or cognitive progress.

Seen in historical perspective, this 2023 critique is structurally continuous with the earlier interventions summarised above. It reprises the ALPAC concern with overpromising, the Dreyfus distinction between formal manipulation and embodied understanding, the Lighthill demonstration of scaling limits, and the McDermott critique of rhetorical inflation. It illustrates that, despite decades of technical development, the core epistemic pattern remains unchanged: syntactic architectures continue to be mischaracterised as cognitive systems, and institutional narratives continue to frame representational refinement as a precursor to intelligence.

CIITR provides the structural explanation underlying this continuity. The historical critiques identify symptoms; CIITR identifies the invariant architectural cause. Artificial systems that manipulate patterns without sustaining internal temporal organisation cannot form or

maintain comprehension. Their outputs may be increasingly fluent, contextually responsive, or representationally dense, but their cognitive status remains constant: they do not enter into temporally grounded epistemic states. The field's recurring pattern of overinterpretation arises not from empirical misjudgment but from the absence of structural concepts capable of distinguishing simulation from understanding. CIITR fills this conceptual void by specifying the architectural preconditions for comprehension and making clear why systems lacking such conditions cannot cross into cognitive domains.

The resulting picture is historically coherent. Early critics documented the mismatch between rhetoric and capability. Later analyses identified the sociotechnical mechanisms that magnified this mismatch into public narrative. CIITR now articulates the structural boundary that renders such mismatches inevitable in non-rhythmic architectures. The illusion therefore persists not because history repeats itself rhetorically, but because the underlying structural condition of artificial systems has not changed.

Conclusion

Viewed through CIITR, McDermott's "Artificial Intelligence Meets Natural Stupidity" becomes not merely a witty warning but a foundational text identifying the very errors that continue to shape the field. His observations map precisely onto the structural invariants that CIITR later articulates: syntactic activity does not imply understanding, representational refinement does not generate comprehension, and behavioural fluency does not constitute cognition.

The problem McDermott criticised is not historical. It is constitutive of every architecture that performs computation without temporal self-maintenance. The illusion endures because the structural boundary remains intact.

In this sense, McDermott's essay is prophetic. He saw the patterns; CIITR reveals their underlying structure. Together, they expose why modern AI narratives—from symbolic programs to transformer models and from "natural language understanding" to "co-superintelligence"—repeat the same conceptual mistake.

They confuse performance with cognition, appearance with structure, and simulation with understanding. And as long as the underlying architectures remain unchanged, the mistake cannot be resolved.

Postscript

The Conceptual History of "Artificial Intelligence" and the State of the Term in 2025

Since its formal introduction in the mid-twentieth century, the term "artificial intelligence" has undergone repeated cycles of expansion, contraction, reinterpretation, and strategic repurposing. It has never possessed a stable definition. Instead, the concept has functioned as a moving frontier, shaped less by structural understanding of cognition and more by shifting technological capabilities, institutional incentives, rhetorical practices, and public

expectations. The history of AI is therefore not merely a history of technical developments but a history of conceptual drift.

The term emerged in 1956 at the Dartmouth Conference, where intelligence was implicitly equated with symbolic manipulation. Early researchers operated under the assumption that human reasoning could be decomposed into formal rules and that machines executing these rules could, in principle, reproduce the essential features of thought. This foundational premise framed intelligence as a computational artefact rather than a situated or temporally grounded process. From this starting point arose the classical symbolic systems of the 1960s and 1970s, whose apparent successes in toy domains encouraged expansive claims about imminent artificial reasoning. Yet these systems faltered as soon as they were confronted with the combinatorial and representational complexity of real environments, revealing a structural mismatch between symbolic procedures and the processual, temporally continuous nature of human cognition.

By the mid-1970s, critical reports such as ALPAC, the Lighthill Survey, and philosophical interventions by Dreyfus had already demonstrated that intelligence could not be captured through rule specification alone. However, despite the conceptual significance of these critiques, the field did not replace the original definition of intelligence with a deeper structural account. Instead, the term “AI” contracted, becoming associated with narrow, isolated problem-solving methods rather than general cognition.

The resurgence of neural networks in the 1980s and subsequent advances in statistical learning reframed AI again, this time as a discipline concerned with pattern extraction from data rather than formal reasoning. With the success of deep learning in the 2010s, the term expanded dramatically. “Intelligence” became associated with predictive accuracy, fluency, and large-scale function approximation. As systems grew in representational capacity, the term “AI” shifted from a speculative aspiration to a ubiquitous label applied to a wide range of computational processes. Despite this expansion, no structural definition of intelligence accompanied these developments. The concept became increasingly detached from its original philosophical referent.

By the early 2020s, the rise of large language models intensified this definitional ambiguity. These systems produced linguistic behaviour that appeared coherent, contextual, and adaptive, encouraging narratives that equated behavioural resemblance with cognitive equivalence. Institutions, markets, and media outlets adopted a functional and performance-based definition of intelligence, treating AI as any system capable of generating outputs consistent with human expectations. In this environment, intelligence became a behavioural proxy rather than a structural property.

In 2025, the term “AI” now commonly denotes transformer-based architectures capable of large-scale pattern synthesis across text, images, audio, and code. These systems are referred to as “intelligent” because they perform tasks associated with human expertise. Yet this contemporary usage rests on a conceptual slippage: intelligence is equated with the capacity to *produce the appearance* of understanding rather than the capacity to *possess understanding*. As a result, the term “AI” has become simultaneously expansive and

imprecise. It encompasses simulation, prediction, statistical correlation, optimisation, and behavioural mimicry, while excluding none of these activities from being described as "intelligent".

This situation renders the term both operationally useful and theoretically unstable. On one hand, AI refers to a set of techniques that demonstrably enable new forms of automation, analysis, and interaction. On the other hand, it obscures the distinction between syntactic performance and cognitive structure, collapsing fundamentally different processes under a single label. In public discourse, "intelligence" has shifted from describing a property of systems capable of grounded, temporally sustained understanding to a descriptor of systems capable of generating convincing behaviour.

The consequence is that, in 2025, the definition of artificial intelligence is less a scientific concept than an umbrella term for a family of computational artefacts whose unifying characteristic is their ability to approximate tasks traditionally associated with human cognition. The term does not denote a structural analogue of understanding, agency, or reasoning. It denotes performance in contexts where performance *resembles* these capacities.

Thus, the history of the term "artificial intelligence" reveals a persistent pattern: the name of the field has remained fixed, while the referent has shifted repeatedly in response to technological possibility and rhetorical need. What remains absent is a framework that distinguishes simulation from comprehension and behavioural reproduction from cognitive structure. In this context, the contemporary use of "AI" does not describe a machine analogue of intelligence. It describes a constellation of computational systems that, through scale and correlation, produce outputs that humans interpret as intelligent.

Any rigorous discussion of artificial intelligence in 2025 must therefore begin with this recognition: the term designates a historical, institutional, and technological assemblage rather than a coherent scientific concept of intelligence.

Distinctions and Misperceptions Between Expert Systems, Machine Learning, Automation, and Artificial Intelligence

The landscape of computational systems commonly grouped under the term "artificial intelligence" consists of several fundamentally different paradigms. Each rests on its own epistemic logic, design principles, operational constraints, and historical lineage. Yet in contemporary discourse, these paradigms are frequently conflated, giving rise to conceptual ambiguity and inflated expectations. The terms *expert system*, *machine learning*, *automation*, and *artificial intelligence* are often used interchangeably despite referring to distinct categories of systems with little structural overlap. This chapter clarifies these distinctions and examines the persistent misperceptions that arise when they are collapsed into a single conceptual frame.

A central source of confusion lies in the assumption that all systems producing complex or adaptive behaviour participate in a unified technological trajectory. In practice, expert systems operate through explicit rules supplied by human specialists; machine-learning systems infer statistical regularities from data; automation relies on deterministic control

logic; and contemporary “AI” systems combine statistical patterning with predictive modelling at scale. These architectures do not merely differ in degree but in kind. Their internal operations, modes of failure, epistemic status, and capacities for generalisation diverge sharply. The widespread misperception that they represent sequential stages toward general intelligence obscures these discontinuities and leads to inaccurate expectations about their capabilities.

Expert systems emerged in the 1970s and 1980s as attempts to codify and operationalise expert knowledge. They rely on hand-crafted rules, logical inference engines, and explicit decision structures. Their strength lies in transparency: every conclusion can be traced to identifiable rules. Their weakness lies in brittleness: they cannot adapt to contexts beyond those envisioned by their designers. Expert systems therefore reproduce expertise rather than discover it. They implement structured decision logic but lack any capacity for learning or contextual drift. Their operation is intrinsically static.

Machine-learning systems differ fundamentally by replacing hand-crafted rules with statistical induction. Instead of representing knowledge explicitly, they infer patterns from data and encode these patterns in parameters rather than rules. Their adaptability is narrowly bounded by the data distributions on which they were trained, and their generalisation arises from statistical regularities rather than conceptual understanding. Machine-learning systems are therefore flexible but opaque. They excel in environments where large data quantities are available, but they do not possess insight into the meaning of the patterns they exploit. Their operation is reactive, not reflective, and their failures stem from distributional mismatch rather than rule mis-specification.

Automation, by contrast, predates both expert systems and machine learning. It consists of engineered sequences of control operations designed to execute repetitive or safety-critical tasks with high reliability. Automation operates on deterministic logic, often embedded in hardware or industrial control systems. It does not infer, adapt, or interpret; it executes. Its purpose is the reduction of variability, not the generation of novelty. Confusing automation with intelligence leads to an erroneous belief that repetitive precision equates to cognitive capability. Automation demonstrates reliability, not reasoning.

Contemporary artificial intelligence systems, particularly those based on large-scale neural architectures, incorporate elements of machine learning but add substantial complexity through multimodal pattern synthesis, general-purpose inference, and large-scale language modelling. These systems generate outputs that mirror human communication patterns and task performance across a wide range of domains. However, their apparent versatility should not be mistaken for comprehension. Their internal mechanisms do not implement symbolic reasoning, expert knowledge representation, or autonomous conceptual integration. They are engines of statistical correlation applied at scale.

The misperception arises because the surface behaviours of these distinct systems can be functionally similar. An expert system producing domain-specific advice, a machine-learning model predicting outcomes, an automation system controlling a process, and a language model generating fluent explanations may all appear to “think” in some colloquial sense. Yet

the internal architectures that generate these behaviours differ fundamentally. Expert systems deduce; machine-learning systems statistically interpolate; automation executes; contemporary AI systems correlate and synthesise. None of these systems inherently performs cognitive operations in the human sense.

The conceptual blending of these categories contributes to the persistent illusion that all roads lead toward artificial general intelligence. By treating distinct architectures as iterations of a single technological lineage, public discourse and policy often assume continuity where discontinuity is the rule. Expert systems did not evolve into machine learning; automation did not expand into intelligence; machine learning did not substantively converge with symbolic reasoning. Instead, each paradigm represents a separate epistemic and technical solution to different classes of problems.

This misperception has practical consequences. It leads to the belief that increasing scale in machine-learning systems will eventually yield the structured reasoning once hoped for in expert systems. It encourages the assumption that automation will become autonomous cognition as complexity increases. It frames current AI systems as early, incomplete versions of a future intelligence rather than categorically different systems optimised for correlation rather than comprehension. Such narratives obscure the specific structural constraints that define each paradigm and prevent informed evaluation of their capabilities.

The distinctions outlined in this chapter therefore serve a clarifying role. Expert systems represent codified human expertise; machine-learning systems represent statistical inference; automation represents deterministic execution; contemporary AI systems represent large-scale predictive modelling. Each class of system has its own strengths, limitations, and modes of failure. Treating them as interchangeable or sequential technologies creates unrealistic expectations and masks underlying architectural boundaries.

Understanding these differences is essential for accurately assessing the trajectory of AI research, the nature of its achievements, and the structural limits that define its present and future capabilities.

Reassessing the Foundations: Why von Neumann and Turing Machines Cannot Constitute Artificial Intelligence Under the Meta FAIR Narrative

Contemporary institutional narratives, particularly those advanced in documents such as Meta FAIR's co-improvement proposal, implicitly redefine artificial intelligence in a way that reveals an unspoken contradiction at the heart of the field. If this narrative is taken seriously and interpreted with conceptual precision, it leads to an unavoidable conclusion: systems built upon the von Neumann architecture or the Turing computational model cannot meaningfully be classified as artificial intelligence. They constitute, at their structural core, statistical optimisation engines operating within fixed, closed, and non-cognitive computational manifolds. Their capabilities may be impressive, but they do not implement the conditions that the FAIR narrative claims to recognise as the trajectory toward intelligence.

The argument follows from the institutional claim that intelligence arises from adaptive, interactive, increasingly autonomous systems capable of forming or participating in complex epistemic feedback loops. FAIR's position treats AI not as static machinery but as an evolving epistemic partner in processes of discovery, learning, and co-development. Intelligence, in this framing, is not exhausted by problem-solving capacity; it includes forms of epistemic participation and cognitive reciprocity.

However, if this definition is accepted, then the classical and contemporary systems commonly labelled as AI fall outside the category altogether. At the architectural level, both von Neumann machines and Turing-equivalent systems share a decisive structural property: their operations are strictly sequential, syntactic, and bound to externally determined control flow. They cannot establish internally sustained epistemic states, nor can they generate the temporal continuity or reflexive dynamics associated with cognition. Their behaviour is the product of symbolic manipulation or statistical inference rather than any form of understanding. The computational substrate therefore lacks the conditions that FAIR's narrative implicitly requires.

Von Neumann machines process instructions in a linear, deterministic cycle. Their memory architecture, control logic, and execution model form a closed computational space that does not permit intrinsic adaptation or temporal self-maintenance. Any apparent "learning" in such systems arises from externally defined procedures modifying externally provided parameters. The machine does not develop an internal history; it does not incorporate past states into self-sustaining epistemic structures; it does not assess or revise its own operations in a reflexive manner. Its behaviour remains tied to the external programmer's design, however complex the resulting procedure may be.

The same limitation applies to Turing-equivalent systems. While Turing machines define the general limits of computability, they do not define the structural requirements for cognition. Their architecture assumes an infinite tape, a finite set of states, and a deterministic transition function. These systems are universal with respect to symbolic computation but non-existent with respect to understanding. They manipulate representations without entering into the meaning of those representations. They can simulate any computable process but cannot constitute a cognitive process. Their universality as computational devices does not imply capacity for intelligence; it implies only capacity for symbolic transformation.

The move from symbolic AI to statistical learning does not alter this structural condition. Machine learning, including large-scale neural networks, remains bound to the same computational substrate. Models train on data through optimisation procedures that adjust parameters according to predefined objectives. Their internal states do not develop temporal persistence; their learning is not self-originating but externally induced; their inference does not constitute interpretation but probabilistic mapping. These systems do not generate understanding; they approximate functions. Their apparent sophistication arises from scale, not from conceptual structure.

If one adopts FAIR's notion that intelligence must involve some form of epistemic participation or cognitive reciprocity, then the label "AI" becomes incompatible with von

Neumann and Turing architectures. Under this framing, such systems can only be classified as extremely powerful optimisation devices that transform inputs into outputs through computation, not cognition. They are mechanisms of statistical condensation and functional reproduction, not agents of epistemic activity.

This conclusion is strengthened when considering FAIR's further suggestion that AI systems should become epistemic partners in scientific discovery. A system that cannot form internal continuity, maintain epistemic commitments, or revise its own inferential basis does not and cannot participate in scientific reasoning. It cannot hold beliefs, evaluate hypotheses, or integrate new information into a coherent developmental trajectory. It can recombine patterns at scale, but it cannot conceptualise. Any proposal that such a system could engage in co-development with human researchers therefore rests on a conflation of productivity with intelligence.

The field's tendency to refer to these architectures as "intelligent" illustrates a larger definitional drift. Historically, the term "artificial intelligence" has been applied to systems that merely *appear* to perform tasks associated with intelligence, even when the mechanisms underlying their behaviour lack any cognitive structure. FAIR's narrative, however, implicitly gestures toward a more demanding definition—one that requires ongoing epistemic activity. When this more rigorous definition is applied consistently, the term AI, as commonly used, loses its referent. The systems that dominate current practice become recognisable for what they structurally are: computational engines of optimisation, correlation, and procedural execution.

This does not diminish their practical utility. Optimisation machines capable of learning statistical patterns at planetary scale have transformed scientific, industrial, and cultural workflows. Their impact is undeniable. But impact is not intelligence, and capability is not cognition. Rebranding these systems as AI obscures the fundamental distinction between computation and understanding, thereby encouraging narratives that exceed the systems' structural limits.

If FAIR's notion is taken seriously, a new conceptual boundary must be drawn. Von Neumann and Turing architectures constitute powerful computational systems, not artificial intelligences. They do not implement cognition; they simulate correlations. They do not learn autonomously; they adjust parameters. They do not reason; they transform representations. Intelligence, in this framing, must be reserved for systems capable of generating and sustaining internal epistemic life.

Recognising this boundary does not weaken the field; it clarifies it. It prevents the misclassification of optimisation devices as cognitive agents and refocuses research on the structural prerequisites for genuine artificial understanding. It transforms the question from how to scale computation into how to design architectures capable of sustaining cognition. This distinction is essential for developing a coherent scientific discipline rather than perpetuating conceptual illusions.

The implication is clear: if intelligence is defined as FAIR implies, then contemporary computational systems are not early steps toward artificial intelligence. They are different systems altogether, powerful but non-cognitive, operationally transformative but epistemically inert. They belong not to the lineage of intelligence but to the lineage of optimisation. Only by acknowledging this can the field progress toward architectures that move beyond simulation and approach the structural conditions necessary for understanding.