What Should Responsible AI Be About—If We Put the Noise Aside?

Dr. Anastassia Lauterbach-Lang

ORCID ID 0009-0006-6988-7406

Responsible artificial intelligence (RAI) may be the most important concept of this decade—and it's currently one of the least precisely defined. The term is used interchangeably to mean governance, compliance, ethics, technical safety, and a vague commitment to "doing good." This ambiguity is a strategic liability: when the definition of responsible AI is blurred, so are the criteria for funding it, addressing it with process guidelines and policies, and defining clear principles on how to hold anyone accountable for failing at it.

AI Edutainment is proposing an RAI framework, which goes beyond the practicalities of developing and using AI tools and technologies. It does not recycle compliance checklists. It does not provide summaries of existing guardrail debates or repackage AI ethics principles that have been circulating since 2016.

The framework consists of four principles, ordered by urgency and scope. The first principle is **f**ocused on improving and changing AI architectures. It applies to large companies and research laboratories building the foundational AI systems that everyone else will use.

The second, focused on radical transparency around AI capabilities and vulnerabilities, applies to every organization deploying AI—from a multinational to a startup running a customer chatbot.

The third and fourth principles, focused on conceptual AI literacy and fusing STEM and humanities to enhance human agency, are the most neglected—and almost certainly the most consequential. They address the long-term relationship between humans and the AI systems we are building and scaling at speed.

Who Decides About Responsible AI?

A paper from Carnegie Mellon from March 2026 found that the unresolved tensions between the AI safety and AI ethics communities are directly slowing progress on governance priorities—not because the problems are unsolvable, but because imprecise language allows that disagreements over power and risk remain unresolved beneath a surface of apparent consensus.[1]

The general public believes that RAI occurs within the large technology companies that build foundational models and deliver mass-market AI applications and services. However, organizations that cannot deploy billions into AI development are not without influence: the institutions, educators, researchers, and public voices that insist on clarity—that refuse to accept fog as a substitute for principle—are doing work that money alone cannot buy. What follows here is a contribution to that effort.

Unlike most technologies with large-scale societal impact, AI occupies a unique and uncomfortable position: it can both strengthen and degrade the very cognitive capacities—reasoning, judgment, creativity—that humans rely on to evaluate it. No previous general-purpose technology has carried that dual risk.

A hammer does not make your hands weaker. A calculator does not erode your ability to reason numerically if used appropriately. Electricity is important for running any factory, and its costs are imperative for pricing of your products and services, but it does not dictate what these will look like. AI, even when deployed without understanding, *can* be used as a shortcut to solve problems. This capability, however, can also seduce an organization into thinking it can "AI" its way out of *any* problem. This leads to the waste of funds and additional risks around safety and usability.

The danger of treating AI as a universal problem-solver—forcing it onto every challenge regardless of fit—is already visible in the data: MIT's Project NANDA (2025) found that 80–95% of enterprise AI pilots deliver zero measurable return on investment, with the

[1] Bálint Gyevnár and Atoosa Kasirzadeh, "Bridging the Gap in the Responsible AI Divides," *arXiv* (submitted March 15, 2026), https://doi.org/10.48550/arXiv.2603.14495.

primary cause being not model quality but organizational learning gaps.[2] This is what happens when organizations deploy AI without understanding what it can and cannot do.

The deeper problem is the lack of general public knowledge about AI, in particular, its development and deployment. Without broad *genuine understanding of AI*—not *tool familiarity* (currently often referred to as *AI fluency*), but *conceptual literacy*—there can be no informed societal dialogue about where AI should be deployed, where the risks genuinely lie, how public and private investment in AI should be scrutinized, or what the future of education and labor markets looks like when AI transitions from a productivity tool to a shared infrastructure.

Principle 1: Improve and Change Current AI Architectures

Making AI more responsible requires making it more intelligent, not just more capable. Simply building larger AI models to increase capabilities that might be on par with those of biological brains is not enough. Without the ability to understand *cause and effect*, to *model the world*, and to *understand other minds*, AI systems cannot be safely trusted with high-stakes decisions. These are not optional upgrades. They are the foundations that truly responsible AI cannot be built without.

The Problem of Causality in Plain Language

Today's most widely used AI systems—ChatGPT, Gemini, Claude, and their cousins—are all built on the same underlying technology: large language models (LLMs). These systems are extraordinarily good at one thing: predicting which word is likely to come next in a sentence, based on patterns learned from enormous amounts of text.

That is a genuine achievement. But it has a ceiling—and that ceiling matters.

Judea Pearl, one of the most important thinkers in the history of AI and a winner of the Turing Award (the equivalent of the Nobel Prize in computing), put it directly: *“Machines'*

[2] Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari, The GenAI Divide: State of AI in Business 2025 (Project NANDA, 2025), https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.

lack of understanding of causal relations is perhaps the biggest roadblock to giving them human-level intelligence." [3] [4]

Pearl describes three levels of reasoning, which he calls the Ladder of Causation:[5]

1. **Seeing**—noticing patterns: *"People who carry umbrellas tend to stay dry."*
2. **Doing**—understanding cause and effect: *"If I carry an umbrella, I will stay dry."*
3. **Imagining**—reasoning about alternatives: *"If I had carried an umbrella yesterday, would I have stayed dry?"*

Current LLMs operate only at level one. They are excellent pattern-recognition machines. They cannot *intervene* in the world, and they cannot *imagine* how things might have been different. Everything they appear to know about causes and consequences is borrowed from human text—not derived from genuine understanding.

Hallucinations Cannot Be Engineered Away

One of the most common questions about AI is: *"Why does it make things up, and will that get better over time?"*

Unfortunately, hallucinations will not be fully solved by making models larger or training them on more data. This is based on foundational mathematics, not a lack of effort or funding.

---

[3] Judea Pearl and Dana Mackenzie, The Book of Why: The New Science of Cause and Effect (Basic Books, 2018).

[4] Quantum Evangenlist, "Judea Pearl and the Dawn of 'Causal AI,'" Quantum Zeitgeist, January 29, 2026, https://quantumzeitgeist.com/judea-pearl-and-the-dawn-of-causal-ai/.

[5] Darko Matovski, "Judea Pearl on the Future of AI, LLMs, and Need for Causal Reasoning: A Fireside Chat with Judea Pearl," causaLens, presented at the Causal AI Conference, June 4, 2024, https://causalai.causalens.com/resources/blog/judea-pearl-on-the-future-of-ai-llms-and-need-for-causal-reasoning/.

Three independent research teams, publishing between 2024 and 2025, have each proved this from different angles: [6] [7] [8]

- **Every LLM will produce false outputs on some inputs**—provably and unavoidably, regardless of size or training data.
- **Every part of the LLM pipeline**—how it is trained, how it retrieves information, how it generates responses—carries an irreducible probability of error.
- **No LLM can simultaneously be truthful, complete, and consistent**—there is a formal mathematical trade-off that cannot be engineered away. The structure of how transformers process information creates an unavoidable gap between what the model "knows" and what it outputs.

This does not mean LLMs are useless. It means they should not be used as sources of verified facts without human oversight—especially in medicine, law, finance, or any domain where errors carry real consequences.

Three Vulnerabilities that Make AI Different from Biological Intelligence

One approach to responsible AI involves aligning it with features enabled by evolutionary breakthroughs in biological brains. But there are three ways in which artificial intelligence fundamentally differs from animal (and human) intelligence, and furthermore, AI researchers might be missing something about how brains actually operate.

[6] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models," *arXiv* (February 13, 2025), https://doi.org/10.48550/arXiv.2401.11817.

[7] "Inevitable Hallucination in LLMs," Emergent Mind, updated December 31, 2025, https://www.emergentmind.com/topics/inevitable-hallucination-of-llms.

[8] Michal P. Karpowicz, "On the Fundamental Impossibility of Hallucination Control in Large Language Models," *arXiv* (submitted June 4, 2025; revised October 15, 2025), https://doi.org/10.48550/arXiv.2506.06382.

AI Cannot Learn After It Is Built

Once an LLM is trained, its knowledge is fixed. It is, in a sense, frozen in time. If the world changes—new laws, a new disease, a geopolitical shift—the AI does not know, unless it is retrained from scratch.

Human intelligence works differently. We integrate new information continuously through experience. We update our understanding in real time. We remember what worked and what did not.

This capacity is called *continual learning*. The challenge of achieving it in AI—without the system forgetting everything it already knew when it learns something new—is called *catastrophic forgetting*. It was identified as a fundamental problem in 1989 by researchers Michael McCloskey and Neal J. Cohen,[9] and it has not been fully solved in the three and a half decades since.[10] [11]

AI Cannot Understand Other Minds

Theory of mind (ToM) is the human (and some primates') ability to understand that other people have beliefs, intentions, and knowledge that are different from our own. It is how we know when someone is confused, when they are lying, when they need help, or when they have misunderstood us.

Current AI systems—including GPT-4, Gemini, and Claude—do not have this capacity. They can *appear* to demonstrate it, because they have been trained on vast amounts of human text that describes mental states. But a 2025 research paper presented at NeurIPS (the world's leading AI research conference) concluded that these systems are

[9] Michael McCloskey and Neal J. Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem," *Psychology of Learning and Motivation* 24 (1989): 109–165, https://doi.org/10.1016/S0079-7421(08)60536-8.

[10] Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi, "Continual Learning and Catastrophic Forgetting," *arXiv* (March 8, 2024), https://doi.org/10.48550/arXiv.2403.05175.

[11] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang, "An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning," *arXiv* (January 5, 2025), https://doi.org/10.48550/arXiv.2308.08747.

doing something much simpler: pattern-matching on familiar phrases, not genuinely modeling other minds.[12]

The difference matters. A medical AI that appears empathetic but cannot actually model what a patient understands and fears should be used only by those who understand current AIs' limitations, and who can mitigate potential errors in AI advice with solid human expertise and judgment. An AI assistant that appears to grasp context but cannot track what the user actually believes might not be reliable for complex tasks.

AI Cannot Reason About "What If"

Counterfactual reasoning—the ability to ask "what would have happened if things had been different?"—is central to how animals learn from scenarios, and how humans assign blame and make ethical judgments.

Consider a self-driving car that causes an accident. A court, an insurance company, and the family of the victim all need an answer to the same question: *"If the car had braked half a second earlier, would the pedestrian have survived?"* Current AI systems cannot answer this. They cannot reason about what would have happened under different circumstances.

Cognitive scientists call this capacity the "imaginarium"—the ability to mentally simulate alternative realities.[13] Without it, AI cannot be trusted to reason about responsibility, consequences, or ethical trade-offs.

[12] Xiaoyun Yin, Elmira Zahmat Doost, Shiwen Zhou, Garima Yadav, and Jamie Gorman, "When Researchers Say Mental Model/Theory of Mind of AI, What Are They Really Talking About?", NeurIPS, OpenReview (published September 23, 2025; last modified February 17, 2026), https://neurips.cc/virtual/2025/129698, https://openreview.net/forum?id=pkKO5Kph3o.

[13] Max S. Bennett, A Brief History of Intelligence (William Collins, 2023).

AI Alignment Misses How Human Values Actually Work

A 2025 position paper reviewed at NeurIPS analyzed over 400 research papers on AI alignment—the field dedicated to making AI pursue human goals.[14] Its finding was striking: almost all of this research treats human values as fixed targets to optimize against, as if human values were a stable list that could simply be programmed in.

But human values are not static. They are negotiated, contextual, and shaped by conversation and experience. What a community values today may differ from what it values in ten years. A responsible AI system must be able to participate in that ongoing negotiation—not just comply with a snapshot of values taken at the moment of training.

Where the Research Is Heading

There is ongoing research tackling cause-and-effect reasoning, world modeling, and theory of mind, with some promising results.

Self-Improving AI: The Gödel Machine and Its Descendants

In 2003, computer scientist Jürgen Schmidhuber published the theoretical blueprint for a new kind of AI: a system that models itself and its environment, and rewrites its own code when it can *prove* the rewrite will make it better.[15] He called it the *Gödel Machine*—named after the mathematician Kurt Gödel, who proved that no sufficiently powerful system can fully verify its own completeness from within itself.

This is the kind of recursive self-awareness that a genuine theory of mind would require: an AI that can represent its own beliefs as objects of reasoning.

---

[14] Hua Shen et al., "Position: Towards Bidirectional Human-AI Alignment," NeurIPS, OpenReview (published September 25, 2025; presented December 3, 2025; last modified October 28, 2025), https://neurips.cc/virtual/2025/poster/121940, https://openreview.net/forum?id=PgA9rZoMY8.

[15] Jüergen Schmidhuber, "Goedel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements," *arXiv* (submitted September 25, 2003; last revised December 17, 2006), https://doi.org/10.48550/arXiv.cs/0309048.

In May 2025, researchers at Sakana AI built a practical version—the *Darwin Gödel Machine*—that autonomously rewrites its own code based on performance testing, improving its score on software engineering benchmarks from 20% to 50% without human intervention.[16]

LeCun's AMI Labs: Building AI That Understands the World

Yann LeCun—one of the three AI scientists who won the Turing Award—has argued consistently that LLMs are a dead end for genuinely intelligent systems. His new company, AMI Labs (launched in 2025 with $1 billion in seed funding),[17] is building AI with *persistent memory*, *hierarchical reasoning*, and the ability to *plan sequences of actions*—an architecture grounded in his 2022 theoretical paper "A Path Towards Autonomous Machine Intelligence."[18] [19] [20]

This approach is built on *world models*—internal simulations based on a deep understanding of how reality works, which allow an AI to imagine consequences before acting, rather than simply predicting the next word.

---

[16] Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune, "The Darwin Gödel Machine: AI That Improves Itself by Rewriting Its Own Code," sakana.ai, May 30, 2025, https://sakana.ai/dgm/.

[17] "Yann LeCun's AMI Labs Launches with a $1B seed @ $4.5B to Build World Models Around JEPA," Latent.Space, AINews: Weekday Roundups, March 11, 2026, https://www.latent.space/p/ainews-yann-lecuns-ami-labs-launches.

[18] Yann Lecun, "A Path Towards Autonomous Machine Intelligence," OpenReview, June 27, 2022, https://openreview.net/forum?id=BZ5a1r-kVsf.

[19] Anna Dawid and Yann LeCun, "Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence," *arXiv* (submitted June 5, 2023), https://doi.org/10.48550/arXiv.2306.02572.

[20] Mahmoud Assran et al., "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," *arXiv* (submitted January 19, 2023; last revised April 13, 2023), https://doi.org/10.48550/arXiv.2301.08243.

DeepMind: Proving That Causality Is Not Optional

Google DeepMind has produced one of the most important theoretical results in recent AI research. Its 2024 paper "Robust Agents Learn Causal World Models" contains a formal mathematical proof: [21]

> *Any AI agent that can reliably generalise to new situations it has never seen before—the most basic requirement for real-world deployment—must have learned an approximate causal model of the world.*

In plain language: if you want an AI that does not break the moment conditions change, that AI must understand cause and effect. Embedded causality is not a feature, but a structural necessity.

This is why DeepMind's research agenda—from AlphaFold[22] (which won Demis Hassabis and John Jumper the Nobel Prize in Chemistry in 2024) to the Genie world model series—is converging on systems that learn causal structure, not just patterns. DeepMind's Genie 3 (August 2025) is the first AI capable of generating real-time interactive 3D environments from a description—described by DeepMind itself as “a key stepping stone on the path to AGI.” [23]

Silver and Sutton's 2025 paper "Welcome to the Era of Experience"—co-authored at DeepMind—argues that the next leap in AI will not come from larger models trained on human text, but from agents that learn continuously through interaction with the world, building internal models from experience.[24]

---

[21] Jonathan Richens and Tom Everitt, "Robust Agents Learn Causal World Models," *arXiv* (submitted February 16, 2024; last revised July 19, 2024), https://doi.org/10.48550/arXiv.2402.10877.

[22] Liwei Chang and Alberto Perez, " AlphaFold2 Knows Some Protein Folding Principles," *bioRxiv* preprint (August 26, 2024), https://doi.org/10.1101/2024.08.25.609581.

[23] Jack Parker-Holder and Shlomi Fruchter, "Genie 3: A New Frontier for World Models," Google DeepMind, August 5, 2025, https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

[24] David Silver and Richard S. Sutton, "Welcome to the Era of Experience," April 26, 2025, https://storage.googleapis.com/deepmind-media/Era-of-Experience /The Era of Experience Paper.pdf.

IBM Research: Causal AI for Business Decisions

IBM Research has developed a Causal AI lab that applies Pearl's causal reasoning framework to enterprise decision-making systems. Its published work explicitly brings intervention calculus into business software, and its 2025 watsonx announcements positioned "decision intelligence"—powered by causal AI—as the layer above generative AI for high-stakes enterprise decisions.[25] [26]

Who Is Teaching This

The academic world is catching up, and here are several (but not the only) examples:

- **CMU's CLeaR Group** (Carnegie Mellon): dedicated causal learning and reasoning research, with annual publications at top venues including NeurIPS, the International Conference on Machine Learning (ICML), and *Nature* [27]
- **Stanford's Causal Science Center**: applied causal inference for science and policy[28]
- **Leiden University (LIACS)**: causal discovery via reinforcement learning, dedicated graduate course for computer scientists [29]

---

[25] Kim Martineau, "All Decisions Have Trade-Offs. IBM's Wei Sun Is an Expert at Weighing Them," *IBM Research Blog*, August 6, 2025, https://research.ibm.com/blog/wei-sun-operations-research.

[26] Yuval David, Fabiana Fournier, Lior Limonad, and Inna Skarbovsky, "The WHY in Business Processes: Unification of Causal Process Models," *arXiv* (submitted May 28, 2025), https://doi.org/10.48550/arXiv.2505.22871.

[27] "CMU-CLeaR Group," Carnegie Mellon University, accessed April 14, 2026, https://www.cmu.edu/dietrich/causality/.

[28] "Stanford Causal Science Center," Stanford University, accessed April 14, 2026, https://datascience.stanford.edu/causal.

[29] "About Us," Universiteit Leiden, accessed April 14, 2026, https://www.universiteitleiden.nl/en/science/computer-science/about-us.

- **Swiss Engineering Master (MSEngineering)**: structural causal models, interventions, and counterfactuals with hands-on projects [30]
- **Zurich University of Applied Sciences (ZHAW Zurich)**: active Swiss National Science Foundation (SNSF)-funded research project on learning world models through physical interaction, with neuroscience grounding and robotics applications[31]
- **University of California Berkeley Cognitive Science**: computational models of belief attribution and theory of mind as a formal problem [32]

Where Applications Stand Today

Some of the research is already being applied in practice.

Causal AI—Already Arriving

Causal AI is the most practically mature of the three capabilities (understanding cause and effect, reflecting the world model, and understanding other minds) discussed in this principle. Industry analysts project the global causal AI market will grow at 42.5% per year from 2026 to 2034, reaching nearly $2 trillion.[33]

Deployment is already happening:

- **Causaly** (UK) has a causal AI platform for clinical trial design in healthcare.
- **DataRobot** has introduced causal modeling with human-in-the-loop explainability.
- **Aitia's Gemini Digital Twins** apply causal AI for drug discovery.

---

[30] "Causal AI (TSM_CausAI)," Master of Science in Engineering, accessed April 14, 2026, https://www.msengineering.ch/de/theoriemodule/2026-2027-tsm-causai.

[31] Pascal Sager, "Learning World Models Through Actionable Representation for Next-Generation AI," ZHAW, accessed April 14, 2026, https://www.zhaw.ch/en/research/project/78451.

[32] "Cognitive Science," UC Berkeley, accessed April 14, 2026, https://cogsci.berkeley.edu.

[33] "Causal AI Market Overview," Fortune Business Insights, last updated March 30, 2026, https://www.fortunebusinessinsights.com/causal-ai-market-112132.

- **IBM watsonx** has positioned causal decision intelligence as its enterprise AI frontier.[34]

World Models—Promising but Not Yet Deployed at Scale

The gap between research and real products remains wide. DeepMind's Genie 3 exists as a research system. AMI Labs is building, not yet shipping. Robotics applications (NVIDIA's GROOT, Google DeepMind's DreamerV3) are leading deployment in physical systems. At AI Edutainment, we don’t believe that the large deployment of a general-purpose world model AI will happen before 2030.

Theory of Mind AI—Early Stage

Current AI can approximate some surface behaviors associated with ToM—detecting frustration, adjusting communication tone, personalizing responses—in narrow applications like customer service and mental health tools. But the fundamental architecture for genuine mental state modeling is not yet solved. Two separate NeurIPS 2025 papers confirmed that the research community has not yet agreed on what a genuine ToM in AI would even mean, let alone how to measure it.[35] [36]

---

[34] Scott Hebner, "IBM Advances watsonx with a Trio of Agentic AI Innovations Tuned for the Enterprise," LinkedIn, *Next Frontiers of AI*, October 9, 2025, https://www.linkedin.com/pulse/ibm-advances-watsonx-trio-agentic-ai-innovations-tuned-scott-hebner-qfdoe .

[35] Xiaoyun Yin, Elmira Zahmat Doost, Shiwen Zhou, Garima Yadav, and Jamie Gorman, "When Researchers Say Mental Model/Theory of Mind of AI, What Are They Really Talking About?", NeurIPS, OpenReview (published September 23, 2025; last modified February 17, 2026), https://neurips.cc/virtual/2025/129698, https://openreview.net/forum?id=pkKO5Kph3o.

[36] Hua Shen et al., "Position: Towards Bidirectional Human-AI Alignment," NeurIPS, OpenReview (published September 25, 2025; presented December 3, 2025; last modified October 28, 2025), https://neurips.cc/virtual/2025/poster/121940, https://openreview.net/forum?id=PgA9rZoMY8.

Principle 2: Radical Transparency About What Works and What Doesn't in AI

There are several aspects of AI that need to be addressed, and being forthright about these problems is a necessary first step to solving them. The language used to define the problems is often imprecise or even misleading; the data used to train models often carries historical bias, and the solutions being currently explored carry their own risks; and there is a lack of clarity about who is accountable for the issues that arise.

The Language Problem: When Words Mislead by Design

The vocabulary used to describe AI systems is not neutral. It is the product of decisions made by companies and their executives, social media influencers, consultants, and policy makers. These decisions sometimes overstate capabilities and misrepresent the risk.

Consider the word "hallucination." A person who has never studied computer science might think that AI hallucinates because something temporarily went wrong. The reality is structurally different: a large language model does not have a truth-seeking mechanism to begin with. It has a next-token prediction architecture. There is no internal state—and no world model—in which the system tries to be accurate and fails. Calling the output of that process a hallucination is not a technical description. It is a rhetorical frame that protects the product from the more accurate description: the system generated plausible-sounding text with no grounding in verified fact. Its underlying mathematics allowed it to happen.

The same problem applies to *bias***.** In the public (and policy-centric) discourse around AI, bias functions as a correctable engineering defect—something that exists in a model but can be removed by a diligent team with better training data, more diverse teams, or new algorithms. This framing is misleading, as we cannot change the historical context that shaped the datasets—or change the underlying mathematics of a model overnight—to eliminate bias and mistakes.

The landmark paper “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, authored by Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell, established that bias removal from internet-scale corpora is not an engineering challenge with an available solution.[37] The data encodes historical power asymmetries—in what was written, by whom, in which language, and under which social conditions—that cannot be excised by post hoc filtering. The structure of the archive *is* the bias.

As an example to highlight such asymmetries in practice, medical datasets underrepresent women because clinical trials historically excluded them.[38] An AI system trained on these datasets will reproduce these asymmetries not as a flaw in its implementation but as a faithful reflection of its inputs. No post hoc intervention changes what the data encoded in the first place.

The Data Problem: Scarcity, Synthetic Solutions, and the Transparency They Require

There is a structural reason why AI systems are more dependent on questionable data than their developers prefer to acknowledge: data scarcity is one of the industry's most persistent and underreported problems. Building capable AI requires enormous volumes of labeled, high-quality training data. In many of the domains where AI could deliver the greatest value—healthcare, finance, law, social care—the most relevant data is, precisely, the most protected. Patient records, legal case files, and financial transaction histories are governed by privacy regulations, including GDPR in Europe and HIPAA in the United States, that restrict their use without explicit consent and robust legal

[37] "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (March 1, 2021): 610–623, https://dl.acm.org/doi/10.1145/3442188.3445922.

[38] Emily Springer and Gratiana Fu, *The Gender Data Foundations of Responsible and Equitable AI,* Data2X (2025), https://data2x.org/wp-content/uploads/2025/09/Gender-Data-and-Equitable-AI-1.pdf.

justification.[39] [40] In other domains, the data simply does not exist in sufficient volume: rare medical conditions, edge-case driving scenarios, or languages spoken by small populations cannot produce the quantity of labeled examples that current AI architectures require.[41]

Generative AI has produced a technically valid response to this problem: synthetic data—artificially generated datasets that replicate the statistical properties of real-world data without containing actual records of real individuals.[42] The practical value is genuine. A healthcare company building a diagnostic AI can train on synthetic patient records that were never created from real patients, bypassing the consent and regulatory burden that would otherwise block development. A financial institution can simulate rare fraud scenarios that occur too infrequently in real data to train a reliable detection model. Gartner has projected that by 2030, synthetic data will be the dominant input for most AI model training.[43]

This is where the transparency argument becomes urgent. We do not yet have a reliable understanding of the long-term consequences of training AI systems at scale on

---

[39] Andrios Robert, "GDPR Synthetic Data Generation: Privacy-Compliant AI Training Without Real User Data," Hoop, October 12, 2025, https://hoop.dev/blog/gdpr-synthetic-data-generation-privacy-compliant-ai-training-without-real-user-data.

[40] Waseem AL Rousan, "How Synthetic Data Is Solving Privacy Challenges in AI Training," Datahub Analytics Blog, July 15, 2025, https://datahubanalytics.com/how-synthetic-data-is-solving-privacy-challenges-in-ai-training/.

[41] Datahub Analytics Team, "Synthetic Data Generation: Addressing Data Scarcity in Machine Learning," Datahub Analytics Blog, March 18, 2025, https://datahubanalytics.com/synthetic-data-generation-addressing-data-scarcity-in-machine-learning/.

[42] Danielle Shanley et al., "Getting Real About Synthetic Data Ethics," *EMBO Reports* 25, no. 5 (February 22, 2024): 2152–2155, https://doi.org/10.1038/s44319-024-00101-0.

[43] Haritha Murari, "How Generative AI Is Revolutionizing Training Data with Synthetic Datasets," Dataversity, August 13, 2025, https://www.dataversity.net/articles/how-generative-ai-is-revolutionizing-training-data-with-synthetic-datasets/.

AI-generated data. Research published in *Nature* has identified a phenomenon called *model collapse*: when AI models are trained repeatedly on their own outputs, without fresh grounding in real-world data, the quality of those outputs degrades progressively across generations.[44] [45] Errors and biases encoded in synthetic data are amplified with each training iteration. Over thirty generations in one documented experiment, outputs that initially resembled real-world examples degraded into indistinct, useless noise.[46] The mechanism becomes clear once formulated with radical honesty: a model trained on outputs generated by a model trained on synthetic data is, in effect, learning from a copy of a copy—with each generation removing it further from the reality it is supposed to represent.

Beyond model collapse, statisticians and data scientists have identified a more subtle problem: *synthetic data does not behave statistically like real-world data, even when it is designed to.* Simple generation methods capture only basic distributional properties—averages, variances, broad correlations—while missing the complex multivariate relationships, rare edge cases, and what researchers call the "unknown unknowns" of real datasets: the subtle patterns that exist in reality but were not anticipated when the synthetic generation process was designed.[47] [48] More sophisticated generation methods improve fidelity but introduce their own failure modes—including the

---

[44] Ilia Shumailov et al., "AI Models Collapse When Trained on Recursively Generated Data," *Nature* 631 (July 24, 2024): 755–759, https://doi.org/10.1038/s41586-024-07566-y.

[45] Sascha Brodsky, "Examining Synthetic Data: The Promise, Risks and Realities," IBM *Think* [newsletter], accessed April 17, 2026, https://www.ibm.com/think/insights/ai-synthetic-data.

[46] Aatish Bhatia, "When A.I.'s Output Is a Threat to A.I. Itself," *The New York Times,* August 25, 2024, https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html.

[47] "Real Data vs. Synthetic Data," *Introduction to Synthetic Data for Machine Learning*, ApX Machine Learning (2026), https://apxml.com/courses/intro-synthetic-data-ml/chapter-1-understanding-synthetic-data/real-vs-synthetic-data.

[48] Edwin Kooistra, "How Reliable Is Synthetic Data?" BlueGen, October 16, 2025, https://bluegen.ai/how-reliable-is-synthetic-data/.

risk that highly realistic synthetic data inadvertently encodes statistical fingerprints of the real individuals it was designed to replace, creating privacy risks through the same back door the synthetic approach was intended to close.[49]

None of this means that synthetic data should not be used. In many contexts, it is the only viable path. But it does mean that organizations using synthetic data to train AI systems have a transparency obligation that they are not currently meeting. Being honest about whether a system was trained on synthetic data, what generation method was used, what quality validation was performed, and what the known statistical divergences between synthetic and real-world distributions are—this is not a technical "nice-to-have-or-disclose." It is the minimum level of honesty required to allow regulators, customers, and downstream users to build and calibrate appropriate trust.

As a suggestion, technology companies could be honest about what they are working on to improve their models and why. Overpromising to fix the issues in the next release and doubling down on the rhetoric of the greater good for AI technologies seem like easy ways out of uncomfortable discussions. If we, however, want AI to touch every aspect of our life, every industry, and our education, we can't pretend there aren't any structural problems to fix.

The Accountability Problem: Who Reports What, and Who Doesn't

Transparent failure reporting—the systematic, public disclosure of what AI systems cannot do, where they fail, and under what conditions they become dangerous—is the minimum requirement to increase accountability around AI. It does not currently exist.

Red-teaming is the industry practice of adversarially testing AI systems before release to surface failure modes. OpenAI, Google, and Anthropic all conduct red-teaming exercises.

---

[49] Haritha Murari, "How Generative AI Is Revolutionizing Training Data with Synthetic Datasets," Dataversity, August 13, 2025, https://www.dataversity.net/articles/how-generative-ai-is-revolutionizing-training-data-with-synthetic-datasets/.

None of them publishes the findings systematically.[50] OpenAI has published a paper describing its approach to external red-teaming, [51] but the document describes the process without disclosing what it found. The distinction matters: a methodology document without results is, in accountability terms, equivalent to a pharmaceutical company publishing its clinical trial protocol while withholding the trial data.

The transparency deficit extends to safety commitments. OpenAI whistleblowers have filed a complaint with the U.S. Securities and Exchange Commission claiming that the company made employees sign agreements requiring them to waive their federal right to whistleblower compensation and prohibiting them from disclosing information to federal authorities without prior permission.[52] This looks like a structural suppression mechanism instead of a governance policy.

Model Cards—the documentation standard introduced by Mitchell, Gebru, and colleagues at Google Research in 2019—were a genuine attempt to create a public record of model limitations and bias profiles.[53] They are now widely used by Google, Meta, and the broader open-source community. But Model Cards are voluntary. They contain no standardized requirements for explaining why debiasing failed, what the magnitude of known failure modes is, or what the system should not be used for in high-stakes environments.

---

[50] "Transparency & Safety," The OpenAI Files, June 18, 2025, https://www.openaifiles.org/transparency-and-safety.

[51] Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin, "OpenAI's Approach to External Red Teaming for AI Models and Systems," *arXiv* (submitted January 24, 2025), https://doi.org/10.48550/arXiv.2503.16431

[52] Transparency & Safety," The OpenAI Files, June 18, 2025, https://www.openaifiles.org/transparency-and-safety.

[53] Margaret Mitchell et al., "Model Cards for Model Reporting," *arXiv* (submitted October 5, 2018; last revised January 14, 2019), https://doi.org/10.48550/arXiv.1810.03993.

What Partial Regulation Has and Has Not Solved

The EU AI Act's Article 13 mandates transparency about AI limitations for high-risk systems.[54] It requires providers to make systems "sufficiently transparent" and to supply documentation including "characteristics, capabilities and limitations of performance."

The EU's definition of "limitations" is predominantly technical: accuracy metrics, robustness thresholds, cybersecurity properties. It does not require epistemological honesty about what the system fundamentally does and does not do. A company can comply with Article 13 by disclosing that its medical AI has a 94.3% accuracy rate on a defined benchmark while remaining entirely silent on the fact that the benchmark dataset excluded women under 40, that the system cannot reason causally about treatment interactions, or that its outputs were produced by a model trained substantially on synthetic data whose statistical divergence from clinical reality has not been publicly validated.

Compliance with Article 13 in its current form is consistent with the continuation of all three failure modes described above. The regulation ensures that bureaucrats and law firms have jobs to check whether companies are compliant. It did not make current AIs better, more transparent, or beneficial for all of us.

What Radical Transparency Would Actually Require

Genuine transparency about AI capabilities and limitations is not a communications exercise. It requires structural commitments that most organizations currently resist because they impose reputational and commercial costs.
At minimum, it requires the following:

- **Honest capability language**: organizations should maintain a public distinction between what their AI systems do (pattern matching, next-token prediction, statistical retrieval) and what they do not do (causal reasoning, genuine

[54] "AI Act Explorer; Article 13: Transparency and provision of information to deployers," The European Commission, AI Act Service Desk, June 13, 2024, https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-13.

understanding, theory of mind). This distinction should be reflected in product naming, marketing materials, and public statements by leadership.

- **Mandatory red-team disclosure**: the findings—not merely the methodology—of adversarial testing should be published before deployment for any AI system used in consequential decision-making, including healthcare, hiring, criminal justice, credit, and education. The format can be standardized; the obligation cannot remain voluntary.
- **Structural bias provenance**: documentation accompanying AI systems should include a description of the historical conditions under which the training data was produced, including known overrepresentation and underrepresentation by demographic groups, language, and geography. This is not a demand for perfection. It is a demand for honesty about the conditions of production.
- **Synthetic data disclosure**: any AI system trained substantially on synthetic data should publicly disclose that fact, the generation method used, the quality validation applied, and the known statistical differences between the synthetic dataset and the real-world distribution it was designed to approximate. Where long-term behavioral effects remain unknown, that uncertainty should be stated explicitly.
- **Failure taxonomy publication**: organizations should publish structured taxonomies of known failure modes, updated on a defined schedule, including failure modes discovered after deployment. This is standard practice in aviation, pharmaceuticals, and nuclear energy. It is not standard practice in AI.

None of these requirements is technically infeasible. They are commercially inconvenient. The gap between what is possible and what is practiced is not an engineering problem. It is a leadership choice.

Principle 3: AI Literacy as Understanding, Not Tool Training

Literacy means understanding the basics about how AI works in comparison to how biological brains work—where AI exceeds human abilities, and where we hold irreplaceable advantages. AI literacy does not imply tutorials on prompt engineering.

Understanding AI properly leads to better decision-making on AI utilization. One cannot "AI" one's way out of every problem; AI is not a universal solver, and knowing this requires understanding what constitutes computability. Crucially, *AI literacy* is not *AI fluency*, which is where most of the current frameworks are focused. A truly beneficial kind of fluency, that is more than just tool familiarity—the ability to deploy AI productively in iterative, generative workflows—is only possible on the foundation of genuine literacy.

Why AI Literacy Matters More Than AI Fluency

The current discourse in corporate and governmental circles has elevated AI fluency, understood as tool familiarity, above true AI literacy. Research by Rogers and Carbonaro (2025) defines AI fluency in a more robust way than is common: as “moving from understanding to creating” with AI.[55] When individuals learn to create with AI before they understand what AI is and is not, they develop dependencies rather than genuine capabilities—they become dependent on a system they cannot interrogate, challenge, or deploy wisely.

Anthropic's AI Fluency Index (2026) identifies eleven directly observable behaviors across 9,830 conversations, finding that iteration and refinement are the strongest predictors of overall AI fluency—users who iterate are 5.6× more likely to question AI reasoning and 4× more likely to identify missing context compared to non-iterative users.[56] Yet iteration cannot be cultivated without the prior capacity to evaluate—to ask whether AI is the right tool, whether the output reflects genuine understanding, or whether the response is a statistically plausible hallucination. That evaluative capacity is what AI literacy provides, and it is structurally absent from "fluency-first" frameworks.

The World Economic Forum's analysis of what constitutes AI readiness acknowledges three interconnected layers: foundational skills (literacy, numeracy, digital fluency),

[55] Thomas Rogers and Mike Carbonaro, "From Understanding to Creating: Bridging AI Literacy and AI Fluency in K-12 Education," *Journal of Teaching and Learning* 19, no. 4 (2025): 20–38, HTTPS://DOI.ORG/10.22329/JTL.V19I4.9420.

[56] Kristen Swanson et al., *Anthropic Education Report: The AI Fluency Index*, Anthropic, February 23, 2026, https://www.anthropic.com/research/AI-fluency-index.

human-centric skills (critical thinking, collaboration), and adaptability.[57] The WEF's *Future of Jobs Report 2023* predicted that 44% of workers' core skills would be disrupted by 2027; its 2025 edition revised this to 39% by 2030—yet 63% of employers still identify skill gaps as the single biggest barrier to business transformation, with no consistent national systems in place to measure which skills are actually being gained or lost.[58] That measurement vacuum is itself a literacy failure: if we cannot assess what people actually understand about AI, all fluency credentials are unverifiable.

The Widening Gap Between AI Literacy Programs and Genuine Literacy

Most corporate and governmental AI literacy programs—including those from Google, Microsoft (AI Skills), and the EU (DigiComp framework)—teach *operational competence*: how to use Copilot, how to write prompts, how to check outputs. None systematically teaches the cognitive science gap between statistical pattern completion and genuine conceptual understanding.

The U.S. Department of Labor (DOL)'s AI Literacy Framework (February 2026) defines an AI-literate person as someone who can "understand, explore, direct, evaluate, and responsibly use AI."[59] At no point does the framework define literacy as the capacity to refuse AI—to recognize when the tool is inappropriate, insufficient, or harmful. The concept of not using AI is structurally absent from its vocabulary. There is no content area

[57] Amit Sevak, "3 Vital Truths About AI Literacy That Will Define the Future," World Economic Forum, October 15, 2025, https://www.weforum.org/stories/2025/10/ai-literacy-3-vital-truths/.

[58] *The Future of Jobs Report* project team, *The Future of Jobs Report 2025*, World Economic Forum, January 7, 2025, https://www.weforum.org/publications/the-future-of-jobs-report-2025/digest/.

[59] U.S. Department of Labor, Employment and Training Administration, Training and Employment Notice No. 07-25, the U.S. Department of Labor's Artificial Intelligence Literacy Framework, issued February 13, 2026, https://www.dol.gov/sites/dolgov/files/ETA/advisories/TEN/2025/TEN%2007-25/TEN%2007-25%20(complete%20document).pdf.

covering what AI costs: the environmental footprint of data centers,[60] [61] [62] the systematic biases baked into training sets,[63] [64] or the well-documented cognitive effects of AI-assisted offloading—flattery-induced intellectual stagnation, erosion of critical thinking, and atrophied judgment.[65] The DOL framework treats AI as something you apply to your work. It does not acknowledge that AI also applies itself to you.

The WEF's framework similarly conflates readiness with tool adoption. Company-sponsored AI training—the dominant delivery channel identified in the WEF's *Future of Jobs* analysis—prioritizes using AI tools to boost productivity: improving Excel use, automating routine tasks, and deploying AI-powered applications.[66] [67] This produces

[60] Léon Pieyre, "AI, Data Centres, and Their Expanding Impact on the Environment," Bertrand Piccard, the Solar Impulse Foundation, June 15, 2025, https://bertrandpiccard.com/articles/ai-data-centres-and-their-expanding-impact-on-the-environment.

[61] David Nutt, "'Roadmap' Shows the Environmental Impact of AI Data Center Boom," *Cornell Chronicle*, November 10, 2025, https://news.cornell.edu/stories/2025/11/roadmap-shows-environmental-impact-ai-data-center-boom.

[62] Roselyne Min, "AI Data Centres Could Have a Carbon Footprint That Matches Small European Country, New Study Finds," Euronews, December 20, 2025, https://www.euronews.com/next/2025/12/20/ai-data-centres-could-have-a-carbon-footprint-that-matches-small-european-country-new-stud

[63] Timnit Gebru, et al., "Datasheets for Datasets," *Communications of the ACM* 64, no. 12 (November 19, 2021): 86–92, https://doi.org/10.1145/3458723; *arXiv* (submitted March 23, 2018; last revised December 1, 2021), https://doi.org/10.48550/arXiv.1803.09010.

[64] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (March 1, 2021): 610–623, https://dl.acm.org/doi/10.1145/3442188.3445922.

[65] Michael Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies* 15, no. 1 (published January 3, 2025; corrected December 10, 2025) https://doi.org/10.3390/soc15010006.

[66] *The Future of Jobs Report* project team, *The Future of Jobs Report 2025*, World Economic Forum, January 7, 2025, https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf.

[67] Galina Misheva, "'The Great Skills Reset' in WEF's Future of Jobs Report 2025: a Catch-22 for the Future of Work?", The European Union, Digital Skills and Jobs Platform, updated February 24, 2026, https://digital-skills-jobs.europa.eu/en/latest/news/great-skills-reset-wefs-future-jobs-report-2025-catch-22-future-work.

workers who are useful to AI makers, but who lack the conceptual grounding to navigate AI's failure modes, challenge its outputs, or contribute meaningfully to its governance.

A 2025 WCET analysis of over fifty AI literacy frameworks found that overlapping categories and inconsistent definitions produce "framework fog"—a proliferation of guides with varying definitions and design assumptions that leaves institutions unable to build coherent, values-aligned strategies.[68] The result is *reactive compliance* rather than a proactive literacy culture.

Current Frameworks Are Not Sufficient: The Evidence

Evidence for framework inadequacy is accumulating rapidly:

- A 2024 report from TeachAI and EY found that *almost half of Gen Z scored poorly on evaluating and identifying critical shortfalls with AI technology*—including whether AI systems can fabricate facts.[69] This is the population that has had the most AI tool exposure of any generation in history. Fluency without literacy produces exactly this outcome: confident users who cannot detect failures.
- The 2024–2025 Syracuse University *Fluency Report* documented a deepening AI digital divide tied not to hardware access but to disparities in digital literacy, critical thinking skills, and institutional investment in structured education.[70] Organizations that prioritize fluency over literacy are systematically widening this divide.

---

[68] Angela Gunder, *AI Literacies in Focus: From Frameworks to Action*, WICHE Cooperative for Educational Technologies (WCET), September 2025, wcet.wiche.edu/wp-content/uploads/sites/11/2025/09/2025-WCET-AI-Literacies-Lit-in-Focus-1.pdf.

[69] Beatriz Sanz Sáiz, Gillian Hinde, and Laura Turkington, "How Can We Upskill Gen Z as Fast as We Train AI?", EY, December 10, 2024, https://www.ey.com/en_gl/about-us/corporate-responsibility/how-can-we-upskill-gen-z-as-fast-as-we-train-ai

[70] Regina Luttrell, Jason Davis, Carrie Welch, and Chris Fiegel, *The 2024–2025 Fluency Report: Bridging the AI Digital Divide*, May 2025, the Emerging Insights Lab at the S.I. Newhouse School of Public Communications, Syracuse University, https://newhouse.syracuse.edu/research/research-spaces/emerging-insights-lab/2024-25-fluency-report-bridging-the-ai-digital-divide/.

- McKinsey's 2025 State of AI Survey found that while 80% of companies set efficiency as the objective of AI initiatives, the organizations achieving the most value from AI set growth and innovation as additional objectives—and systematically redesign workflows rather than merely layer AI onto existing processes.[71] [72] At the AI Edutainment enterprise consulting practice, we saw that workflow redesign requires the kind of structural understanding of AI capabilities that only genuine literacy can provide.
- IBM's *Race for ROI* report (2025) found that while 66% of EMEA organizations report significant productivity gains from AI, small and midsized enterprises (SMEs) and public sector organizations are falling behind larger private firms.[73] The most plausible structural explanation: large firms invest in deep AI capability development and hire AI experts; smaller organizations adopt fluency-first tools without building the foundational literacy that enables meaningful workflow transformation.
- McKinsey's long-term analysis sizes the AI productivity opportunity at $4.4 trillion in added corporate value.[74] The gap between those who capture this value and those who do not will not be determined by who has access to the best AI tools—it will be determined by who understands them well enough to use them wisely.

---

[71] Alex Singla et al., "The State of AI: How Organizations Are Rewiring to Capture Value," QuantumBlack, AI by McKinsey, March 12, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value.

[72] Alex Singla et al., "The State of AI in 2025: Agents, Innovation, and Transformation," QuantumBlack, AI by McKinsey, November 5, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[73] IBM conducted a survey in partnership with Censuswide, involving over 3,500 senior executives across 10 EMEA countries in September 2025, with a report produced in October 2025 by Hans Peter Dalen and Sebastian Weir, *The Race for ROI: How AI is Transforming Enterprise Productivity in EMEA*, https://filecache.mediaroom.com/mr5mr_ibmnewsroom/199848/IBM%20EMEA%20Report%20-%20The%20Race%20for%20AI%20.pdf.

[74] Michael Chui et al., *The Economic Potential of Generative AI: The Next Productivity Frontier,* McKinsey & Company, June 14, 2023, https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

Where Genuine Literacy Work Is Emerging

- Research comparing STEM and humanities students on AI literacy needs found that humanities students might be better positioned to engage with AI's ethical and epistemological dimensions—but are systematically underserved by tool-focused curricula.[75] [76] [77]
- The OECD AI Literacy Framework for Primary and Secondary Education (May 2025 review draft) is one of the few frameworks to explicitly state that "AI excels at pattern recognition and automation but lacks emotions, ethical reasoning, context, and originality"—acknowledging AI's structural limits as a core literacy competency, not a footnote.[78]
- The AI Edutainment Romy & Roby AI series[79] is genuinely doing what almost no corporate literacy program does: explaining the *architecture* of intelligence by comparing biological and artificial systems in terms accessible to young people, their parents, and non-specialists alike.

---

[75] Amara Atifa, Lisa-Angelique Limb, A. Baki Kocaballia, and Anna Lidfors Lindqvist, "Navigating AI Literacy: STEM Students' Perceptions of Generative AI," presented at the 36th Australasian Association for Engineering Education Annual Conference (2025), https://aaee2025.org/microsite/pdf/full-paper_175.pdf.

[76] Cristo Leon, James Lipuma, and Xavier Oviedo-Torres, "Artificial Intelligence in STEM Education: A Transdisciplinary Framework for Engagement and Innovation," *Frontiers in Education* 10 (July 8, 2025), https://doi.org/10.3389/feduc.2025.1619888.

[77] Angela Gunder, *AI Literacies in Focus: From Frameworks to Action*, WICHE Cooperative for Educational Technologies (WCET), September 2025, https://wcet.wiche.edu/wp-content/uploads/sites/11/2025/09/2025-WCET-AI-Literacies-Lit-in-Focus-1.pdf.

[78] Cathy Adams et al., "Knowledge Statement: AI's Capabilities and Limitations," K4.1, p. 20, *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*, review draft, OECD/European Commission/Code.org, May 2025, https://ailiteracyframework.org/wp-content/uploads/2025/05/AILitFramework_ReviewDraft.pdf.

[79] You can explore the first book in the series, *Romy, Roby and the Secrets of Sleep*, at www.romyandroby.ai; visit www.aiedutainment.ai to learn more about AI Edutainment and the rest of the series.

Discussion Points

- Understanding where AI exceeds humans (processing speed, pattern recall, consistency across millions of cases, no fatigue) versus where humans are irreplaceable (embodied experience, genuine curiosity, moral considerations, contextual judgment, counterfactual reasoning) is the foundational literacy question—and it is almost never framed this way in existing frameworks.
- Teaching people only to use AI tools without understanding the underlying mechanism creates dependency—analogous to teaching someone to drive without knowing what an engine is, then asking them to diagnose a failure. The immediate experience is seamless; the long-term vulnerability is structural.
- Marvin Minsky's *Society of Mind*, originally published in 1986[80] and now being revisited in light of 2025 developments in AI architecture, provides a far richer framework for comparing brain and AI architecture than anything currently present in corporate literacy curricula.[81] [82] [83]
- Stephen Wolfram's work on computational irreducibility highlights a further foundational insight: some problems are computationally irreducible, meaning there

[80] Marvin Minsky, *The Society of Mind* (Simon & Schuster, 1986).

[81] AI engineer Sutha Kamal, in his June 2025 essay, “Revisiting Minsky's Society of Mind in 2025,” published on Substack and LinkedIn and discussed on Hacker News, argues, "The AI field is running into the limits of gigantic, monolithic models—like today’s large language models (LLMs) that try to do everything in one go—and increasingly looking toward modular, multi-agent approaches. Techniques that once sounded fanciful in Society of Mind, like collections of specialized 'mini-AIs' and internal self-monitoring agents, are re-emerging as practical strategies for building more robust, scalable, and aligned AI."

[82] Robert Engels, "Rediscovering ‘Society of Mind’: Marvin Minsky’s Timeless Lessons on AI and Collective Intelligence," Capgemini, June 20, 2024, https://www.capgemini.com/ch-en/insights/expert-perspectives/marvin-minskys-timeless-lessons-on-ai-and-collective-intelligence/.

[83] The technical connection is real: mixture-of-experts models (used in GPT-4, Gemini, Mistral), multi-agent frameworks (AutoGen, HuggingGPT), and AI alignment self-reflection techniques are all structural realizations of Society of Mind principles; see Sutha Kumal's essay “Revisiting Minsky's Society of Mind in 2025,” https://www.linkedin.com/pulse/revisiting-minskys-society-mind-2025-sutha-kamal-qbnmc/.

is no shortcut to understanding them. AI cannot think its way around them; only humans who understand the underlying computational logic can recognize where that boundary lies.[84] [85] [86] [87]

Principle 4: Defend the Human Creative Purpose Through the Fusion of STEM and the Humanities

If companies developing foundational AI architectures claim to be good citizens, they should look into how to actively defend why humans write, learn mathematics, and create art. Understanding what enables lazy cognition, particularly in teenagers, is a must. This is not an argument against AI technologies.

Why Writing Is a Thinking Technology, Not a Communication Tool

This is the most neglected principle in the entire AI responsibility landscape. No major AI company has a program that answers the question “Why should a human write an essay

---

[84] See Stephen Wolfram's book, *A New Kind of Science* (Wolfram Media, 2002), page 739: “Whenever computational irreducibility exists in a system it means that in effect there can be no way to predict how the system will behave except by going through almost as many steps of computation as the evolution of the system itself,” https://www.wolframscience.com/nks/p739--computational-irreducibility/.

[85] Wolfram wrote about this on his blog, *Writings*, in his March 2024 article, "Can AI Solve Science?": “We can't expect to systematically 'jump ahead' and predict or 'solve' the system; it inevitably takes a certain irreducible amount of computational work to figure out what exactly the system will do.” https://writings.stephenwolfram.com/2024/03/can-ai-solve-science/.

[86] Susan Stem quotes Wolfram in her January 2026 article "Stephen Wolfram’s View of AI: My Biggest Takeaway in 2026 Is the World’s Computational Irreducibility—and That Our Goal Is to Use AI to Find 'Pockets Of Reducibility,'" on her blog, *Susan STEM's Entropy Control Theory*: “there just won’t be any way—with AI or otherwise—to shortcut… step by step.” Stem claims, "AI cannot let us systematically bypass irreducibility," https://www.entropycontroltheory.com/p/stephen-wolframs-view-of-ai-my-biggest.

[87] In November 2023, Northeastern University's Institute for Experiential AI's executive director, Usama Fayyad, had a talk and fireside chat with Wolfram as part of the Institute’s Distinguished Lecturer Seminar series. He argued that "...computational irreducibility is a monster at the gate there that will prevent some things from being possible” for LLMs attempting scientific prediction. You can read more about the interview in the December 2023 article, "Stephen Wolfram Shares His Vision for Humanity’s AI Future," https://ai.northeastern.edu/news/stephen-wolfram-shares-his-vision-for-humanitys-ai-future.

when AI can do it?"—not as a policy restriction but as a genuine argument for human cognitive flourishing.

Naomi Baron, author of the books *Who Wrote This? How AI and Lure of Efficiency Threaten Human Writing* (2023), argues that writing is not primarily a communication technology—it is a thinking enabler.[88] [89] [90] When a student outsources writing to AI, they are not skipping a task; they are skipping the cognitive process through which ideas become coherent.

Writing activates and strengthens the neural circuits for sustained attention, working memory, and higher-order reasoning. Flannery O'Connor's famous observation—"I write because I don't know what I think until I read what I say"[91]—is echoed across literary history by Horace Walpole, E.M. Forster, Arthur Koestler, George Bernard Shaw, William Faulkner, and Joan Didion.[92]

This matters for AI companies specifically because their tools are the primary vector through which cognitive offloading is now occurring at scale.[93] Speaking up in favor of developing writing skills and celebrating clearly written scientific publications could be

---

[88] Naomi S. Baron, *Who Wrote This? How AI and the Lure of Efficiency Threaten Human Writing* (Stanford University Press, 2023).

[89] Naomi S. Baron, *Words Onscreen: The Fate of Reading in a Digital World* (Oxford University Press, 2015).

[90] Naomi S. Baron, "Reading and Writing Skills in the Age of AI," Children and Screens Institute of Digital Media and Child Development, July 2025, https://www.childrenandscreens.org/learn-explore/research/reading-and-writing-skills-in-the-age-of-ai-naomi-baron-phd/.

[91] Timothy Taylor, "I Don't Know So Well What I Think Until I See What I Say," Conversable Economist, August 29, 2018, https://conversableeconomist.com/2018/08/29/i-dont-know-so-well-what-i-think-until-i-see-what-i-say/.

[92] Naomi S. Baron, "Why Human Writing Is Worth Defending in the Age of ChatGPT," Literary Hub, September 12, 2023, https://lithub.com/why-human-writing-is-worth-defending-in-the-age-of-chatgpt/.

[93] Justin Jackson, "Increased AI Use Linked to Eroding Critical Thinking Skills," Phys.org, January 13, 2025, https://phys.org/news/2025-01-ai-linked-eroding-critical-skills.html.

seen as a sign of thought leadership, not just for the sake of AI development, but for enhancing human agency in the world of AI.

The Humanities–STEM Divide as a Responsible AI Failure

The exclusion of humanists from AI development is not a cultural oversight—it is a structural risk. Ethicists, historians, linguists, and artists understand AI's social risks more acutely than engineers do, yet they are almost entirely excluded from model development, evaluation, and governance. This produces AI systems that are technically sophisticated and socially naive: capable of generating grammatically perfect text but structurally unable to account for the human context in which that text will land.

The defense of human creative purpose must therefore be articulated not as a restriction on AI capability but as a design principle: AI systems built by organizations that are genuinely interested why humans write, create, and reason will produce different products than those built by organizations that are not.

Where This AI Literacy Framework Sits Relative to the Industry

The four principles outlined in this paper operate at a deeper level than any corporate or regulatory standard currently in existence.

Dimension	Corporate RAI Frameworks	U.S. DOL/WEF Frameworks	These Principles
Causal reasoning & theory of mind	Absent	Absent	Core safety prerequisite
Transparency of failure	Partial, PR-filtered	Partial, compliance-framed	Radical and structural

Literacy focus	Tool usage	Workforce productivity	Cognitive architecture understanding
AI literacy vs. fluency distinction	Collapsed	Collapsed	Explicitly argued and ordered
Cost of AI (human, ecological, cognitive)	Not addressed	Not addressed	Named as literacy competency
STEM–humanities fusion	Siloed	Siloed	Explicitly required
Defense of human creative purpose	Not addressed	Not addressed	Central ethical obligation
Writing as cognitive development	Not addressed	Not addressed	Core to education principle

Responsible AI is not a constraint on AI development—it is a commitment to preserving what is distinctly human while enhancing human capabilities, precisely because AI is becoming so capable.