Mental Models and High-Dimensional Representation

ECE 57000 — September 2, 2026

David I. Inouye

Monday turned MAGE into six durable operating principles

Make thinking explicit

1. Preserve consequential context.
2. Build structure at useful levels.
3. Deliver the right context to each decision.

Turn judgment into a system

4. Define value, requirements, and evidence.
5. Govern invariants and preserve lessons.
6. Formalize repeatable checks when worthwhile.

Make intent explicit → delegate → inspect evidence → govern → preserve the lesson

Mental Models for Wise Use

Use familiar histories to reason about an unfamiliar medium

LLMs are human artifacts at unprecedented scale

- Pretraining learns statistical structure from vast amounts of human-created language and other media.
- Post-training adds human demonstrations, evaluations, objectives, and engineered feedback.
- Scale and composition can yield surprising capabilities without making the system a foreign intelligence.

> **LLMs transform patterns in human artifacts through human-designed systems; they do not create intelligence from nothing.**

Two historical models make AI less conceptually foreign

AI as a tool

- Extends human capacity for knowledge work
- Can be used skillfully, carelessly, or destructively
- Directs attention to leverage, control, and responsibility

AI as a book

- Distills human knowledge into a dynamic, personalized medium
- Can inform, persuade, omit, distort, or invent
- Directs attention to trust, provenance, and discernment

Neither model is complete; together they give us useful inherited intuition.

AI inherits the same core challenges as tools and books

From tools

- Extend human capacity and reduce effort
- Amplify both good and poorly chosen goals
- **Human task:** choose the purpose and retain control

From books and media

- Preserve, transmit, and distill human knowledge
- Amplify both warranted and misleading claims
- **Human task:** judge the source, evidence, context, and value

AI does not remove these old challenges; it requires both responsibility and discernment.

Discussion: an advising agent raises two ordinary questions

A university proposes an AI system that explains degree requirements, recommends a semester schedule, and—after student approval—submits registration changes.

1. **When should a student trust its explanation of degree requirements enough to act on it?**

2. **Which registration changes should it be allowed to submit, and what must remain reversible or require advisor approval?**

What is new is the combination—and the compressed transition

1. **Book + tool in one system:** AI can generate knowledge-like content and immediately use it to act, revise, decide, or create.
2. **A faster adaptation problem:** capabilities and practices are changing quickly enough that individual and institutional judgment may lag.
3. **Large-scale amplification:** AI can reproduce work and influence broadly, although earlier waves—including the internet—also reached enormous scale.

> **The novelty increases the coupling, speed, and reach. It does not replace the core human challenges we already recognize.**

As a tool, AI reallocates knowledge work

- Engines and machines transformed physical work; LLMs and agents transform parts of knowledge work.
- A tool can remove low-value mechanics—or displace the judgment and practice the work was meant to develop.
- The central question is not whether AI saves effort, but **which effort and attention it saves**.

> **Is AI freeing your attention for the most consequential choices—or helping you avoid them while you optimize details that matter less?**

As a book, AI is dynamic—not authoritative

- Oral tradition → written text → printing press → internet → search → generative AI
- Unlike a fixed book, an LLM adapts its explanation and synthesizes across patterns on demand.
- Like every medium, it inherits human knowledge, blind spots, incentives, and errors.
- Unlike most books, it can improvise persuasive content without stable authorship or provenance.

Fluency is a property of the medium—not evidence that a claim deserves trust.

Trust becomes confidence engineered for the stakes

- Perfectly checking every claim or action would erase much of AI’s leverage—and is often impossible.
- Confidence should come from the whole system: explicit intent, bounded authority, evidence, tests, review, monitoring, and recovery.
- Use manual inspection where human judgment is essential; formalize recurring checks when worthwhile.
- Increase assurance as consequences, uncertainty, irreversibility, and exposure increase.

> **The goal is not certainty about every output. It is a managed system whose outputs deserve the confidence you place in them.**

Discernment chooses the right assurance structure

Situation	Appropriate confidence-building structure
Exploratory and reversible	Use freely; label uncertainty; retain promising ideas for later checking
Important but bounded	Define acceptance evidence; sample outputs; independently check consequential parts
Repeated or scaled	Add structured inputs, tests, monitoring, ownership, and feedback loops
High-impact or hard to reverse	Bound authority; require stronger evidence and human gates; prepare fallback and recovery

Responsibility means designing proportionate control—not personally redoing every action.

Fluent output can create an illusion of competence

AI can make three different things look better than they are:

1. **The artifact:** polished form can conceal weak evidence or reasoning.
2. **Your capability:** producing with AI can feel like being able to produce or explain independently.
3. **The system:** several successful runs can feel like evidence of dependable behavior.

Artifact quality, human learning, and system reliability require different evidence.

Test the capability you actually need to preserve

When the goal is learning

- Can you explain the idea without the AI?
- Can you transfer it to a new case?
- Can you diagnose or revise a flawed answer?

When the goal is delegation

- Can you specify what good means?
- Would the system expose important failure?
- Can you intervene, recover, and improve the process?

> **Do not demand unaided performance for every delegated task; preserve the understanding and control the larger system requires.**

The mental models clarify four questions for wise use

1. **Nature:** Does AI support the purpose of this work—or quietly change what the work is?
2. **Attention:** What am I choosing not to think about, and is my attention moving to what matters most?
3. **Capability:** Am I strengthening the judgment I need—or hiding a gap that will matter later?
4. **Confidence:** What structure, evidence, and control justify trusting the result at these stakes?

Wise use combines intuition and engineered control

> **Treat AI as a human-made tool and knowledge medium. Then build the environment that directs its leverage, exposes its failures, and preserves the human judgment the work requires.**

- The **tool** model asks where attention and responsibility should move.
- The **book** model asks what confidence the medium and its evidence deserve.
- **MAGE-derived principles** turn those judgments into structures that can govern real work.

AI should optimize your thinking—not replace the parts of thinking that make the work valuable.

Dimensionality Reduction: Motivation and Problem Setup

A table can be computable while remaining impossible to inspect

Suppose we measure **20,000 gene-expression values** for each of **500 cells**.

$$500 \text{ cells} \times 20{,}000 \text{ measurements per cell}$$

- A computer can store and manipulate the table.
- A person cannot directly see its geometry in 20,000 dimensions.
- We still want to find groups of similar cells, flag unusual cells, reveal related or connected structures, and remove statistical noise.

Dimensionality reduction serves two practical goals

1. Visualize hidden structure

Create a 2D view where we can look for:

- clusters of similar cells;
- unusual cells or outliers; and
- connected, branching, or other structures.

2. Improve further analysis

Create a lower-dimensional representation that:

- requires less computation; and
- can yield more reliable statistical estimates by ignoring spurious variation caused by finite samples or measurement noise.

Both goals replace many measured coordinates with fewer useful coordinates.

Goal 1: visualize hidden structure

Original observations

- One point per cell
- Thousands of measured coordinates
- Structure exists beyond direct human vision

This is a representation of handwritten digits—not cells—but it illustrates the kinds of clusters and other structure we might seek in cellular data.

Goal 2: improve further analysis

A lower-dimensional representation can provide:

- **computational efficiency:** fewer coordinates to store and process;
- **statistical efficiency:** fewer irrelevant directions to estimate from limited data; and
- **denoising:** less sensitivity to measurement noise or spurious sample variation.

> **These benefits require the discarded variation to be less useful than the structure retained. Reduction is not automatically an improvement.**

Both goals lead to the problem of dimensionality reduction

We seek a new representation with far fewer coordinates:

$$\text{high-dimensional observations} \longrightarrow \text{low-dimensional representation}$$

But “make it simpler” does not say what a faithful simplification should retain.

- Which points should remain near one another?
- Which directions of spread should remain large?
- Which distinctions may be treated as noise?
- Should the original measurements be approximately recoverable?