A PRE-READ FOR BOARDS & EXECUTIVE TEAMS

The World Has Changed.

What boards and executive teams need to understand about the AI era — before the session begins.

- TRAINING → INFERENCE
- CAREMARK / MARCHAND
- MCKINSEY · LILLI
- NESTED LEARNING

AUTHORS

Magnus Tagtstrom · Jon Nordmark
Iterate.ai Executive Team

EDITION

Handout 01
Board AI Governance training

SUBJECT

The Governance Shift
Director & executive pre-read

ABOUT THIS PAPER

This is not a technology briefing. It is an argument.

It is written for board directors and senior executives who are already broadly aware of AI, and who want a clearer understanding of what has actually changed — and why that change is a *governance* problem, not a technology problem.

It is intended to be read before the session, and argued with during it.

HOW TO READ IT

Most frameworks describe the AI transition as a long, evenly-weighted list — nine shifts, twelve dimensions, five risk categories. This paper argues that one shift matters more than the rest, and organizes everything around it. Disagree with that framing; that is the point of the session.

WRITTEN FOR

Board directors and senior executives who already know AI matters, and want the mental model to govern it.

PUBLISHED BY

Iterate.ai — San Jose, CA & Denver, CO. Private AI infrastructure for the enterprise.

CONTENTS

What's inside.

One shift, four consequences. Read it front to back — the McKinsey case on p.7 and the memory section on p.12 are the two most directors ask to revisit.

01

THE ONE SHIFT THAT MATTERS

AI moved from supporting human decisions to making them.

That single change is what alters the nature of the board's job — not just its agenda. Everything else in the AI-governance conversation follows from it.

It changes what governance means at its root.

THE IT ERA — PASSIVE SYSTEMS

Think about the systems companies have relied on for thirty years. A payroll system that calculated paychecks. A customer database. An ERP like SAP or Oracle that tracked orders. An email server. A financial reporting tool.

Every one of them did exactly what it was told. It stored things, moved things, or calculated things. **It never decided anything.** A human always made the call; the system carried it out.

Governing them was straightforward, even when it took real effort: make sure they worked, stayed secure, weren't misused. Check in quarterly. Protective, periodic, manageable. The key word is **passive** — these systems had no opinions, didn't learn, and didn't change behavior based on what they had seen.

THE AI ERA — SYSTEMS THAT LEARN

AI systems do not follow fixed rules written by a programmer. **They learn.** A hiring tool is not given a checklist of a good candidate — it studies thousands of past decisions and builds its own model of "good." A loan AI is trained on millions of past loans and forms its own sense of creditworthiness.

This is powerful — and the source of entirely new risk. Because behavior comes from what was learned, not explicit rules, a system can develop biases, blind spots, and errors that no one programmed and no one can easily see. And at machine speed, a problem that affects one employee's ten decisions a day can affect an AI's ten thousand decisions an hour, before anyone notices.

The board's old questions — "were we hacked?" "did we stay on budget?" "are we compliant?" — were built for a world where technology stored things and humans decided things. That world is gone.

02 THE WORD THAT LOCATES THE GAP

There is one word boards should understand: inference.

The AI story so far has been about *training* — building better models. The operational reality is about *inference* — trained models running, deciding, and acting every minute of every day. That is where the board's attention has not yet moved.

PERIODIC · VISIBLE

Training

What happens when a model **learns** — it studies data, develops its judgment, and is prepared for deployment. It happens in a controlled environment, periodically, with some visibility.

CONTINUOUS · IN THE DARK

Inference

What happens **after** — the trained model is deployed and actively running, making decisions, generating outputs, and taking actions, continuously, in production.

The board's oversight gap is almost entirely at inference.

That is where cost accumulates, where errors compound, where bias scales, where manipulation is possible, and where accountability gets blurry. It is also the part of the AI lifecycle that gets the least board attention. We are moving from a world dominated by training to one dominated by inference — where trained models are the operational infrastructure of the enterprise.

WHAT THIS LOOKS LIKE IN PRACTICE, RIGHT NOW

- **JPMorgan Chase** built COiN to review commercial loan documents — doing in seconds what took **360,000 hours** of lawyer time per year.
- **UnitedHealthcare** faced a lawsuit alleging its AI denied post-acute care claims at a rate no human reviewer would have approved.
- At major **banks**, AI approves or flags loan applications before a human ever sees them.
- At **insurance companies**, AI is deciding which claims to pay.

In most of these cases there is technically a human "in the loop." But that human is reviewing hundreds of decisions per hour, under pressure to move quickly, with tools that make it easy to approve and hard to push back. **That may not be oversight. It may be rubber-stamping with extra steps.**

03

A DIFFERENT KIND OF ATTACK

The target is no longer your data. It is your decisions.

Consider what happens when someone attacks one of these systems — not to steal information, but to corrupt judgment.

THE OLD WORLD

Stolen data

A cyberattack meant leaked records: customer files, credit card numbers. Bad, expensive — and recoverable. An alarm goes off. There is a trail.

THE NEW WORLD

Corrupted decisions

An attacker manipulates what your AI **sees** — feeding false signals so it approves fraudulent loans, clears bad actors through compliance, or mis-prices risk. No files stolen. No alarm. The system just starts deciding differently.

Researchers have already demonstrated that AI fraud-detection systems can be fooled into approving transactions they should block, simply by crafting inputs the system was not trained to recognize. By the time anyone notices, months of damage may have accumulated — and the trail is nearly impossible to follow.

> **The attack surface on an AI decision system is not just the data it holds. It is the behavior it produces.**

This is the governance problem in a sentence. The new questions — *who is the AI making decisions for, on what basis, with what checks, and could those decisions be manipulated without anyone knowing* — are not yet on most board agendas. The next page is what this looks like when it actually happens.

04

A REAL EVENT · FEBRUARY 28, 2026

An autonomous agent walked through McKinsey's front door.

A white-hat security firm, CodeWall, unleashed an autonomous AI agent on the public web. It found McKinsey's internal AI platform, **Lilli** — and 22 unauthenticated endpoints. No credentials. No insider knowledge. No human in the loop. In two hours, it had full read *and write* access to the entire production database.

22	46.5M	728K	57K	95
OPEN DOORS (UNAUTH. ENDPOINTS)	CHAT MESSAGES, IN PLAINTEXT	CONFIDENTIAL FILES	USER ACCOUNTS	SYSTEM PROMPTS — ALL WRITABLE

THE DETAIL THAT SHOULD STOP EVERY DIRECTOR

Those 95 system prompts controlled how the AI responded to every one of McKinsey's **43,000 consultants**. A malicious actor with write access could have silently rewritten what Lilli told them — subtly altering financial models, strategic recommendations, and risk assessments going to clients worldwide.

No code change. No deployment. No security alert. Just a single database update — and 43,000 consultants start receiving poisoned advice they have every reason to trust, because it comes from their own internal tool.

THE VULNERABILITY

SQL injection — a technique documented since 1998.

The platform had run in production for two years. McKinsey's own scanners never found it. It patched within 24 hours of disclosure and reported no evidence of access beyond the researchers. The point still stands.

05 MAKE IT CONCRETE

Four failures. One pattern.

It is easy to treat AI governance as an abstract risk-management exercise. These numbers make it concrete — and in every case, the AI was doing exactly what it was designed to do.

$881M Zillow

In 2021, Zillow's AI-powered home-buying algorithm badly overestimated property values — it decided which homes to buy and at what price. The result: an **$881 million loss**, 2,000 layoffs, and the shutdown of the business unit.

Nobody had adequate oversight of what it was actually learning, or how its judgment was drifting from market reality.

45 min Knight Capital

A faulty trading algorithm lost **$440 million in 45 minutes** — a firm profitable for years, effectively destroyed in under an hour.

The algorithm made thousands of decisions per second. No human could have reviewed them fast enough to stop it.

90% UnitedHealth · nH Predict

A class action alleges an AI used to determine post-acute care coverage for elderly patients had a **90% error rate on appeals** — nine of ten denials were overturned when a human looked.

The suit claims it was optimized for cost savings, not medical accuracy, and systematically overrode physicians.

100+ Workday · Mobley

In *Mobley v. Workday*, a plaintiff alleges the AI hiring tool discriminated against applicants over 40, rejecting him from **100+ jobs within minutes**.

In May 2025 a federal court granted preliminary certification as a nationwide class action. Every employer using that tool is potentially exposed.

The pattern is consistent: **the AI did what it was built to do.** The problem was that nobody with real authority was watching closely enough, asking hard enough questions, or had the structures in place to catch the drift before it became a disaster. That is a governance failure — and the board is the last line of defense against it.

06 THE COST, AND THE DOCTRINE

When a business problem becomes a director-liability problem.

$4.88M
PER DATA BREACH

The average breach now costs $4.88 million.

From IBM's 2024 Cost of a Data Breach report — covering response, lost business, and penalties. In AI environments the cost is expected to be higher: the attack surface is larger and the damage harder to contain. A compromised AI doesn't just leak data — it corrupts decisions in real time, often for months before discovery.

$67B
IN 2024

Hallucinations cost an estimated $67 billion.

The business cost of AI systems producing confident, plausible-sounding wrong answers — in legal documents, financial analysis, customer communications, and compliance filings. Not from headline failures, but from the quiet accumulation of errors nobody caught until the damage was done.

THE CAREMARK STANDARD

When AI is mission-critical, oversight becomes personal.

Under *Caremark* (Del. Ch. 1996), boards face **personal liability** when they fail to implement and monitor systems for overseeing mission-critical risks. *Marchand v. Barnhill* (2019) refined it: for risks central to the business, oversight must be "more rigorously exercised."

AI is moving fast into mission-critical territory — pricing, hiring, lending, compliance, financial reporting. As it does, the argument that a board had no responsibility to establish AI oversight becomes harder to sustain.

07

THE DEEPER PROBLEM

The gap is not structural. It is epistemic.

Most descriptions of AI governance assume the fix is the right committee, the right reporting lines, the right metrics. That assumption is wrong — and it matters that it is wrong. The deeper problem is one of knowledge: most boards do not yet know enough to know what they don't know.

WHAT IT LOOKS LIKE IN THE ROOM

Management presents a quarterly AI risk update: tools in use, policies in place, controls implemented. The board asks whether the tools are secure and the company is compliant. Management says yes. The board moves on.

What the board did not ask: Is the model *drifting* — has its behavior changed since deployment? Is the training data *clean*? Do we control our own AI — hardware, models, and data — or does a vendor? If the vendor changes the model tomorrow, did they tell us, and what happens to our decisions?

Those questions don't require a technical degree. They require a basic understanding of how AI works that most directors have not yet had reason to build. That gap is not a personal failing — AI moved faster than governance education. But it is a real risk.

> **Oversight without understanding is not oversight. It is ratification — approving things you cannot fully evaluate.**

A board that reads a report it cannot critically examine, asks three general questions, and moves on has not governed AI. It has created a paper trail that *looks like* governance. If something goes wrong, that trail does not protect the board — it becomes evidence that the board had the information and still did not ask hard enough questions.

08

A TENSION WORTH NAMING

The pressure to deploy fast is in direct conflict with the oversight responsible governance requires.

The risk varies dramatically with *which* tools a company is actually using. These are not hypotheticals — they are happening now, often without board-level visibility.

Shared models CLAUDE · GITHUB COPILOT	Shared models and shared compute for software development — as many dev teams now use — can quietly create intellectual-property risks that neither the team nor the board has thought through.
ChatGPT SHADOW AI	Easy to use in any browser — which is exactly what makes it a shadow-AI risk. Sensitive information typed into a consumer tool is not confidential. It is a shared model running on shared hardware.
Microsoft Copilot LOOKS SECURE	Looks secure because it's Microsoft — but most deployments still run on shared models and shared hardware, which creates its own exposure.
OpenClaw AUTONOMOUS AGENT	An open-source autonomous agent that went viral in January 2026, scaling to 1.5M agents in weeks and connecting itself to corporate email, Slack, Workspace, and file systems — often without security's knowledge. An audit found 512 vulnerabilities. Researchers called it "shadow AI with elevated privileges." When it is compromised, the attacker doesn't need to breach you — the agent handed over the keys.

A knowledgeable board does not slow the company down. It lets the company move faster with less risk.

A board that cannot govern AI becomes a bottleneck one of two ways: it rubber-stamps decisions it cannot evaluate (liability), or it slows deployment out of unspecific concern (competitive drag). Both come from the same gap — the knowledge to make fast, calibrated, confident decisions about AI risk.

09 THE HARDER PROBLEM

The memory problem.

Most discussions focus on what AI systems do *today*. The harder problem is what they become over time. Memory is the ultimate enabler of these decision-making, action-taking systems — and the dimension of governance boards are least equipped to think about, because they have never had to.

Think about what it would mean to manage a *person's* memory. It is not like managing a filing cabinet — you cannot open a drawer and delete a folder. Think of a judge telling a jury to disregard a comment. Can they? Of course not. The information is in there; it shapes what they think, even if they don't realize it. That is the problem with AI memory — except at enterprise scale, in real time, across thousands of decisions a day.

There was no equivalent in the IT era. A database stored what you put in it — you could see it, export it, delete it. AI memory is different. It is not a file. It is a **pattern** — something the system learned, baked into how it thinks, not stored in a way you can easily find or remove.

> **"We are in the GPT-2 era of memory, but the time will come when AI remembers every detail of your life and personalizes itself based on all of that — not just the facts, but also the little preferences you didn't even think to mention."**
>
> **Sam Altman, CEO of OpenAI · podcast with Alex Kantrowitz, December 2025**

Altman is talking about consumer AI — but the same dynamic is playing out inside enterprises right now. What should make boards sit up is his follow-on: **"I don't think we know yet how far we should allow this to go."** If the CEO of one of the world's most prominent AI companies is uncertain about the limits of a technology he is actively deploying, boards should be asking hard questions.

10 THE TECHNICAL FRONTIER · NEURIPS 2025

When the model *is* the memory, you cannot delete it.

Today's models have two kinds of memory: what they were trained on (frozen) and the current conversation window (which vanishes when the session ends). Everything in between does not exist — which is why current AI does not truly learn from your organization's ongoing use.

Nested Learning, published by Google Research at NeurIPS 2025, changes that. It structures the model as a stack of nested layers, each updating at a different speed — fast layers for immediate context, mid layers for recent experience, slow deep layers for stable knowledge. Google's prototype is called **HOPE**. The result: a model that learns continuously from use, without forgetting and without retraining from scratch.

WHY IT'S A GOVERNANCE PROBLEM

When memory lives in a separate database, you can find it, audit it, and delete it.

When memory is baked into the model — distributed across layers that update continuously — you cannot. There is no file to open. The model is the memory.

"A continuously learning model could behave differently for different users, which raises security and consistency issues." — IBM, Chief Architect of AI Open Innovation

THIS MATTERS FOR GOVERNANCE IN THREE SPECIFIC WAYS

1 **Memory is where small errors become systemic.**

A bias encoded in early interactions does not stay contained. An AI that learns to respond differently to certain zip codes — because the data reflected historical underservice — quietly shapes every future interaction there. Six months later it shows up in retention numbers, but nobody connects it back. The causal chain is untraceable by then.

2 **Memory creates an accountability problem no framework addresses.**

With a traditional system, a log shows who changed what and when. With memory-enabled AI, a decision today may stem from something learned eighteen months ago, from interactions nobody logged, through patterns nobody intended to teach it. There is no audit trail. You cannot roll it back.

3 **Memory is a strategic asset boards aren't yet treating as one.**

Every conversation, every deal won or lost, every compliance decision — an extraordinarily valuable institutional knowledge base. But if that memory lives inside a third-party vendor's model, who owns what the AI has learned? Boards have asked "who owns our data?" for a decade. The next version is "who owns what our AI has learned from our data?"

A DIRECT STATEMENT

What has actually changed.

For those who prefer the summary before the discussion.

	IT ERA	AI ERA
What is governed	Boards governed **systems**.	Boards must govern **decisions and outcomes**.
Shape of risk	Periodic and event-driven.	Continuous, adaptive, and compounding.
Accountability	Human and traceable.	Blended, sometimes untraceable.
Technology's role	Served strategy.	Executes strategy — and sometimes makes it.
Governance lag	Lagged deployment by quarters.	Lag creates legal exposure in real time.
Vendors	Supplied software.	May control model behavior, training, and memory.
The new frontier	Data governance — what is stored.	Memory governance — what has been learned, and by whom.

> WHAT THIS SESSION ASKS OF YOU
>
> It is not designed to make you feel better about AI governance. It is designed to make you more *capable* of it — a different thing. The most useful posture is not skepticism, and not confidence that your current oversight is adequate. It is a willingness to ask harder questions, to hold management to a higher standard of specificity, and to be honest about what your board does and does not know.
>
> **AI is expected to be as transformative as electricity. It is not an old-fashioned IT issue. It is a governance issue — and it is already here.**

ABOUT THIS PAPER

A companion to the session — and to Handout 02.

This paper was prepared by a combination of Iterate.ai's team, Claude (Anthropic), and Iterate's *Generate* platform. It is intended to frame today's session in strategic terms, and is a companion to **Handout 02: AI Risk Disclosure and the Regulator in the Room**, which frames it in legal terms.

It reflects publicly available information as of April 2026 and is intended for educational purposes. It is not legal advice and should not be relied upon as such.

AUTHORS

Magnus Tagtstrom — CORP VP

Worked on products requiring GDPR compliance in Europe for Couche-Tard (Circle K).

Jon Nordmark — CEO

Served on Colorado's governor- and legislature-appointed AI Task Force during the development of the first broad-sweeping US AI law.

REFERENCES

01 JPMorgan Chase, *COiN (Contract Intelligence)* — commercial-loan document review.

02 Zillow Offers — iBuying unit shutdown and write-down, 2021.

03 Knight Capital Group — algorithmic trading loss, August 2012.

04 UnitedHealth / *nH Predict* — post-acute care coverage class action.

05 *Mobley v. Workday* — preliminary class certification, May 2025.

06 IBM, *Cost of a Data Breach Report 2024* ($4.88M average).

07 AI hallucination business-cost estimate, 2024 ($67B).

08 McKinsey "Lilli" / CodeWall disclosure — Feb 28, 2026.

09 *In re Caremark* (Del. Ch. 1996); *Marchand v. Barnhill* (Del. 2019).

10 Google Research, *Nested Learning* (HOPE) — NeurIPS 2025.

11 Sam Altman — podcast with Alex Kantrowitz, December 2025.

A note on scope. Memory is discussed at greater length in the booklets distributed during the general IterateOn sessions. This handout is the strategic companion to Handout 02, which addresses the SEC, Caremark, and disclosure exposure in legal detail. Together they are intended to frame today's session before the discussion begins.

SERIES
Handout 01
Board AI Governance

HEADQUARTERS
San Jose, CA
& Denver, CO

ON THE WEB
iterate.ai
hello@iterate.ai

EDITION
April 2026