HANDOUT 1

The World Has Changed.

What Boards and Executive Teams Need to Understand About the AI Era — Before the Session Begins

Board AI Governance Training Session | IterateOn | April 2026

A note on what this paper discusses:

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.

THE SHIFT THAT ACTUALLY MATTERS

There is no shortage of frameworks describing how AI changes board governance. Most of them are organized as lists: nine shifts, twelve dimensions, five categories of risk. They are well-structured. They are also, mostly, **wrong in one important way.**

They describe a transformation as though it is happening evenly across all fronts simultaneously, with roughly equal urgency on each. It is not. The shift that matters most — the one that changes the nature of the board's job, not just its agenda — is a single one:

> ***AI has moved from supporting human decisions to making them. That changes what governance means at its root.***

Think about the **IT systems companies have relied on for the past thirty years**. A payroll system that calculated paychecks. A customer database. An ERP system like SAP or Oracle that tracked orders and inventory. An email server. A financial reporting tool that pulled numbers together for the CFO. Every one of these systems did exactly what it was told. It stored things,

moved things, or calculated things. But it never decided anything. A human always made the call. The system just carried it out.

Governing those systems was straightforward, even if it took real effort. The board's job was to make sure systems worked, remained secure, and weren't misused. Check in quarterly. If something broke or got hacked, find out why and fix it. That was governance — protective, periodic, manageable. The key word is **passive: these systems had no opinions. They didn't learn. They didn't change their behavior based on what they'd seen before.**

AI systems are fundamentally different. They do not follow fixed rules written by a programmer. They learn. An AI hiring tool is not given a checklist of what makes a good candidate — it studies thousands of past hiring decisions and develops its own model of what "good" looks like. A loan approval AI is not handed a rulebook — it is trained on millions of past loans and builds its own sense of creditworthiness. A fraud detection system does not check a list of known fraud patterns — it learns to recognize suspicious behavior from examples and applies that judgment to every transaction it sees.

This is powerful. It is also the source of **entirely new kinds of risk**. Because an AI system's behavior comes from what it learned, not from explicit rules, it can develop biases, blind spots, and errors that no one programmed and no one can easily see. And because **AI operates at machine speed and scale**, a problem that would affect one employee's ten decisions per day can affect an AI's ten thousand decisions per hour — **before anyone notices** something is wrong.

There is a word that captures exactly where governance needs to focus, and it's one that boards should understand: **inference**. Training is what happens when an AI model learns — it studies data, develops its judgment, and gets prepared for deployment. That process happens in a controlled environment, periodically, with some visibility. **Inference is what happens after — when that trained model is deployed and actively running, making decisions, generating outputs, taking actions, every minute of every day.**

We are moving from a world dominated by training, where the AI story was about building better models, to **a world dominated by inference, where trained models are the operational infrastructure of the enterprise.**

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. And it is the part of the AI lifecycle that gets the least board attention.

Here's what this looks like in practice right now, at real companies.

- JPMorgan Chase built an AI system called COiN that reviews commercial loan documents. It does in seconds what used to take 360,000 hours of lawyer time per year.
- United Healthcare 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 what 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 process quickly, with tools that make it easy to approve and hard to push back. **That may not be oversight.** That may be **rubber-stamping** with extra steps. And in some cases, **the human has been removed entirely.**

Now consider **what happens when someone attacks one of these AI systems** — not to steal data, but to corrupt decisions.

In the old world, a cyberattack meant stolen data: customer records leaked, credit card numbers compromised. That is bad, expensive, and recoverable.

In the new world, a sophisticated attacker no longer needs to steal your data. They can manipulate what your AI sees — feeding it false signals so it approves fraudulent loans, clears bad actors through compliance checks, mis-prices risk across a portfolio, or routes resources in ways that benefit the attacker. 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. No files are stolen. No alarm goes off. The system just starts making slightly different decisions. By the time someone notices, months of damage may have accumulated — and the trail is nearly impossible to follow.

This is not hypothetical. On February 28, 2026, a security firm called CodeWall (a white hat security firm) unleashed an autonomous AI agent on the public web, and it found **McKinsey's internal AI platform, Lilli**. The agent found 22 'open doors' or unauthenticated endpoints — and walked right through them. No credentials. No insider knowledge. There was no human in the loop. In two hours, the agent had full read and write access to the entire production database. What it found: 46.5 million chat messages covering strategy, mergers and acquisitions, and client engagements — in plaintext. 728,000 confidential files. 57,000 user accounts. And 95 system prompts that controlled how the AI responded to every one of McKinsey's 43,000 consultants — all writable.

That last detail is the one that should stop every board member reading this. A malicious actor with write access to that database could have silently rewritten what Lilli told McKinsey's consultants — subtly altering financial models, strategic recommendations, and risk assessments going to clients around the world. 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 that made it possible? SQL injection — a technique documented since 1998. The platform had been running in production for two years. McKinsey's own internal security scanners never found it. McKinsey patched within 24 hours of disclosure and stated there was no evidence of unauthorized access beyond the researchers. But the point stands: the attack surface on an AI decision-making system is not just the data it holds. It is the behavior it produces.

This is the governance problem in a sentence: 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. 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 showing up on most board agendas.

WHAT FAILURES ACTUALLY COST

It is easy to treat AI governance as an abstract risk management exercise. These numbers make it concrete.

Zillow. In 2021, Zillow's AI-powered home-buying algorithm badly overestimated property values. The company had been using it to decide which homes to buy and at what price. The result: an $881 million loss, 2,000 employees laid off, and the complete shutdown of the business unit. The AI was doing exactly what it was designed to do. The problem was that nobody had adequate oversight of what it was actually learning, and how its judgment was drifting from market reality.

Knight Capital. A faulty trading algorithm lost $440 million in 45 minutes — a firm that had been profitable for years was effectively destroyed in less than an hour. The algorithm was making thousands of decisions per second. By the time anyone could intervene, the damage was done. No human could have reviewed those decisions fast enough to stop it.

UnitedHealth and the nH Predict algorithm. A class-action lawsuit alleges that an AI system used to determine post-acute care coverage for elderly patients had a 90% error rate on appeals — meaning nine out of ten times a human reviewer looked at a denial the AI made, the

human overturned it. The lawsuit claims the system was optimized for cost savings, not medical accuracy, and that it systematically overrode physician recommendations. The legal exposure is substantial; so is the reputational damage.

Workday. In Mobley v. Workday, a plaintiff alleges that Workday's AI hiring tool discriminated against job applicants over the age of 40, rejecting the lead plaintiff from over 100 jobs within minutes. In May 2025, a federal court granted preliminary certification, allowing the case to proceed as a nationwide class action. Every employer using that screening tool is potentially exposed.

The average data breach now costs $4.88 million. That figure, from IBM's 2024 Cost of a Data Breach report, covers direct response costs, lost business, and regulatory penalties. In AI environments, breach costs are expected to be higher because the attack *surface is larger and the damage harder to contain*. A compromised AI system does not just leak data — it corrupts decisions in real-time, and the corruption may have been running for months before discovery.

Hallucinations cost $67 billion in 2024. That is the estimated business cost of AI systems producing confident, plausible-sounding wrong answers — in legal documents, financial analysis, customer communications, and compliance filings. Not from spectacular failures that make headlines, but from the quiet accumulation of errors that nobody caught until the damage was already done.

The pattern across these failures is consistent:

The AI was doing 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 oversight structures in place to catch the drift before it became a disaster. That is a governance failure. And in every one of these cases, the board is the last line of defense against it.

When these failures happen at a company whose board had no meaningful AI oversight structure, there is a legal doctrine that turns them from a business problem into a director liability problem. It is called the Caremark standard, after a 1996 Delaware Court of Chancery decision. Under Caremark, boards face personal liability when they fail to implement and monitor systems for overseeing mission-critical risks. The doctrine was refined in *Marchand v. Barnhill* (2019) to hold that for risks central to a company's business, board oversight must be "more rigorously exercised." AI is moving fast into mission-critical territory — influencing pricing, hiring, lending, compliance, and financial reporting at companies across every industry. As it does, the argument that a board had no responsibility to establish AI oversight structures becomes harder

to sustain. The session will go deeper on what Caremark means for director liability in the AI era.

THE PROBLEM MOST FRAMEWORKS MISS

Most descriptions of AI governance assume that the problem is structural: boards need the right committee, the right reporting lines, the right metrics. Fix the structure and the governance follows.

That assumption is wrong, and it matters that it is wrong.

The deeper problem is one of knowledge — specifically, **most boards not knowing enough to know what they don't know.** The technical word for this is **epistemic**, meaning it is *a problem of understanding, not just of structure*. When we say the governance gap is epistemic, we mean the problem is not just that the board lacks the right committee. It is that most directors do not yet have the mental framework to ask the right questions about AI, even when they want to.

Here is what that looks like in practice. A company's management team presents a quarterly AI risk update. It covers the tools in use, the policies in place, and the controls that have been implemented. The board asks whether the tools are secure and whether the company is compliant. Management says yes. The board moves on. What the board did not ask: Is the model drifting? Meaning, has its behavior changed since it was deployed, in ways that nobody intended? Is the training data clean? Meaning, was the data used to teach the AI free of bias, errors, or gaps that would cause it to make bad decisions at scale? Do we control our own AI (i.e. Private AI involving three things: hardware, models, and data), or does a vendor control it (Public AI)? If the vendor changes the model tomorrow, did the vendor inform us, and what happens to our decisions?

Those questions do not require a technical degree to ask. But they do require a basic understanding of how AI systems work that most directors have not yet had reason to build. That gap is not a personal failing — AI has moved faster than governance education. But it is a real risk.

This matters because oversight without understanding is not really oversight. It is ratification — approving things you cannot fully evaluate. *A board that reads an AI risk 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 later, that paper trail does not protect the board. It becomes evidence that the board had the information and still did not ask hard enough questions.

The governance gap is not just structural — it is a knowledge gap. And the knowledge gap is the harder problem to fix.

This is not an argument that directors need to become data scientists. It is an argument that the current standard — general awareness, delegation to management, periodic review — is insufficient for a world in which AI systems are making material decisions on behalf of the enterprise. The standard of care is moving. The law is beginning to reflect that movement. And the boards that have not yet noticed are carrying risk they have not yet quantified.

AN UN-NAMED TENSION

There is a tension at the center of AI governance that most frameworks describe around but never confront directly. It is worth naming — and with specific examples, because the risk varies dramatically depending on which AI tools a company is actually using.

The competitive pressure to deploy AI fast is in direct conflict with the deliberate, continuous oversight that responsible AI governance requires.

Using shared models — like Claude or GitHub Copilot — and shared compute infrastructure for software development, as many dev teams now do, can quietly create intellectual property risks that neither the team nor the board has thought through.

Using ChatGPT is easy — it runs in any browser, which is exactly what makes it a shadow AI risk. Sensitive information typed into a consumer tool is not confidential. It's a shared model running on shared hardware.

Using **Microsoft Copilot** looks secure because it's Microsoft — but most deployments still run on shared models and shared hardware, which creates its own exposure.

And then there is **OpenClaw:** an open-source autonomous AI agent that went viral in January 2026, scaling to 1.5 million agents within weeks, and connecting itself to corporate email, Slack, Google Workspace, and file systems — often without the security team's knowledge. A security audit found 512 vulnerabilities before it reached mainstream adoption. Researchers described it as "shadow AI with elevated privileges." When an agent like this is compromised, the attacker does not need to breach your systems. They already have the keys — because the agent handed them over.

These are not hypothetical scenarios. They are happening right now in most organizations, often without any board-level visibility.

The old framing of this tension was: boards that govern most carefully move slowest. That framing is wrong, and it lets boards off the hook in the wrong direction.

The real dynamic is this: **a board that does not understand AI cannot govern it well** — and a board that cannot govern it well becomes a bottleneck in one of two ways. Either it **rubber-stamps** decisions it cannot evaluate, which creates liability. Or it **slows deployment** out of unspecific concern, which creates competitive drag. Neither outcome is acceptable. Both are the result of the same problem: a board that lacks the knowledge to make fast, calibrated, confident decisions about AI risk.

A knowledgeable board does not slow the company down. It **enables the company to move faster with less risk**, because governance decisions that used to require lengthy review can be made quickly by people who understand what they are looking at. The goal of this session is to start closing that gap.

Organizations that deploy AI aggressively capture real advantages: compressed time-to-market, dramatically lower operating costs, new business models that competitors cannot quickly replicate. The pressure on management to move fast is real, commercially rational, and will not ease. If anything, it will intensify as AI capabilities accelerate.

At the same time, the legal exposure for inadequate oversight is building. Enforcement actions are being filed. Derivative suits are being structured. Regulators across the EU, the U.S., and multiple states are constructing frameworks that will hold companies accountable for AI-driven harms. The governance standards that courts and regulators will apply retroactively are being written now, based on what sophisticated actors should have known and done.

This is not a problem that resolves itself. It requires a deliberate board-level decision about risk appetite — explicitly **owned at the board level**, not delegated away. What is our threshold for AI deployment without adequate controls? Which tools are approved for which uses? Who has the authority to stop a deployment that is moving faster than oversight can follow? A board that cannot answer those questions is not protecting the company. It is leaving the company exposed.

THE MEMORY PROBLEM: WHERE THIS GOES NEXT

Most AI governance discussions focus on what AI systems do today. The harder problem is what they become over time. **Memory** is discussed at length in the booklets distributed during the general IterateOn sessions. It is the ultimate enabler of these new AI decision-making and action-taking systems. And it is the dimension of AI governance that 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 they just heard.** Can they? Of course not. The information is in there. It shapes what they think, even if they do not realize it. That is the problem with AI memory — except at enterprise scale, running in real time, across thousands of decisions per day.

Boards have never dealt with memory before. Neither have companies. 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.** AI systems are accumulating context about employees, customers, suppliers, and strategies — and that accumulation is happening faster than governance frameworks can track. What should make boards sit up is his follow-on comment: ***"I don't think we know yet how far we should allow this to go."*** That is the CEO of one of the world's most prominent AI companies expressing uncertainty about the limits of a technology his company is actively deploying. If he is uncertain, boards should be asking hard questions.

Here is what is now taking shape: **persistent, large-scale memory systems** that allow AI to build deep, long-term representations of organizations, people, and processes. These are not static databases. **They learn continuously. They update based on every interaction.** And they are beginning to power a new generation of AI agents that do not just answer questions — they take action, remember what they did, and adjust their behavior over time based on outcomes.

The technical frontier that makes this a board-level issue right now is a paradigm called **Nested Learning**, published by Google Research at NeurIPS 2025 in November. Here is what it means in plain terms.

Today's AI models have **two kinds of memory**: what they were trained on before deployment, which is frozen and cannot change, and what is in the current conversation window, which vanishes the moment the session ends. Everything in between does not exist. This is why current AI does not truly learn from your organization's ongoing use of it — every session largely starts fresh.

Nested Learning changes that architecture fundamentally. Instead of treating a model as a single system with static weights, it structures the model as a stack of nested learning layers, each updating at a different speed. Fast-updating layers handle immediate context. Mid-speed layers integrate recent experience. Slow-updating deep layers accumulate stable, long-term knowledge. Google's prototype implementation is called **HOPE**. The result is **a model that learns continuously from use — without forgetting what it knew before, and without needing to be retrained from scratch.**

Why this is a governance problem, not just a technical one:

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

When memory is baked into the model itself — distributed across layers that update continuously — you cannot delete it. There is no file to open. There is no record to pull. The model is the memory. IBM's Chief Architect of AI Open Innovation put it directly: "A continuously learning model could behave differently for different users, which raises security and consistency issues."

Boards have governed data retention for decades. **They have no equivalent framework for governing a model that learns, adapts, and changes its own behavior as it operates — because nothing like this has existed before.**

This matters for governance in three specific ways.

First, memory is where small errors become systemic. A bias encoded in early interactions does not stay contained. Imagine an AI customer service system that early on learns to respond differently to customers in certain zip codes — because the training data reflected historical patterns of underservice. That bias quietly shapes every future interaction in those areas. It

compounds. Six months later, it shows up in customer retention numbers, but nobody connects it back to the AI. The causal chain is practically untraceable by then.

Second, memory creates an accountability problem that no existing governance framework can fully address. With a traditional system, if something goes wrong, you can trace it. A database log shows who changed what and when. With a memory-enabled AI, a decision made today may be the result of something the system learned eighteen months ago, from interactions nobody logged, through patterns nobody intended to teach it. There is no audit trail for that kind of learning. You cannot roll it back. This is not a future problem — it is happening now in any enterprise running AI systems that learn and adapt over time.

Third, memory is a strategic asset that boards are not yet treating as one. Think about what an AI system accumulates over years of operation: every customer conversation, every deal that was won or lost, every compliance decision, every internal experiment. That is an extraordinarily valuable institutional knowledge base. But if that memory lives inside a third-party vendor's model — a model the enterprise does not own or control — then who really owns what the AI has learned? Boards have asked "who owns our data?" for a decade. The next version of that question is "who owns what our AI has learned from our data?" Most boards have not asked it yet.

WHAT HAS ACTUALLY CHANGED: A SUMMARY

For those who prefer a direct statement before a longer discussion:

IT Era	AI Era
Boards governed systems	**Boards must govern decisions and outcomes**
Risk was periodic and event-driven	**Risk is continuous, adaptive, and compounding**
Accountability was human and traceable	**Accountability is blended, sometimes untraceable**
Technology served strategy	**Technology executes strategy — and sometimes makes it**
Governance lagged deployment by quarters	**Governance lag creates legal exposure in real time**
Vendors supplied software	**Vendors may control model behavior, training, and memory**

Data governance: what is stored	**Memory governance: what has been learned, and by whom**

WHAT THIS SESSION IS ASKING OF YOU

The session you are about to participate in is not designed to make you feel better about AI governance. It is designed to make you **more capable of governing AI** — which is a different thing.

The most useful posture for the session is not skepticism of AI, and it is not confidence that your current oversight structures are adequate. It is a willingness to ask harder questions than you have been asking, to hold management to a higher standard of specificity than you may have required, and to be honest about what your board currently does and does not know about how AI is operating inside your organization.

AI is expected to be as transformative as electricity. The best boards will help their organizations move fast and safely — because not doing so could place the companies they govern at a strategic and operational disadvantage. AI is not an old-fashioned IT issue. It is a governance issue. And it is already here.

This handout is a companion to Handout 1: AI Risk Disclosure and the Regulator in the Room. Together they are intended to frame today's session in both strategic and legal terms before the discussion begins.

About This Paper

This paper was prepared by a combination of Iterate.ai's team, Claude (Anthropic), and Iterate's Generate platform. Iterate.ai's team includes Magnus Tagtstrom, Corporate VP, who worked on products that required GDPR compliance in Europe for Couche-Tard (Circle K), and Jon Nordmark, CEO, who served on Colorado's governor and legislature-appointed AI Task Force as Colorado worked, for a year, through the first broad-sweeping AI law passed in the United States.