Iterate.ai NetApp

Understanding the AI Revolution.

What Every Business Leader and Board Member Needs to Know — Now

What's Inside

AI's rise is often compared to electricity.

On some benchmarks, AI capabilities are now doubling every three months—**a pace that equals a 16x annual improvement.**

Compare that to Moore's Law, which predicted chip power would double every 18 months.

AI is evolving at a speed that outpaces even the fastest revolutions of the past.

New terms appear daily: KV Cache, Mixture of Experts, omni-modal models, and RL gym frameworks. Even core ideas like backpropagation are being rethought and optimized. Breakthroughs arrive not yearly, not monthly, but weekly.

For leaders, the challenge is this: how do you make confident choices when the ground shifts this quickly?

The answer is to build a **clear understanding of the AI ecosystem.**

This is the first of two booklets. It is your guide through the GenAI stack. **The second booklet is the action plan**. We'll weave together stories of chips, models, memory, and private AI.

Instead of jargon, you'll find metaphors, examples, and case studies from Iterate.ai's work—**grounded in real implementations, not theory**.

We'll also cover **memory** — how AI remembers things now, and why that changes everything about privacy.

Along the way, we'll point you to practical tools to stress-test legal frameworks, assess entrepreneurial DNA, and more. These are free resources you can use immediately.

Here is what we cover...

AI Field Guide — Understanding the Revolution

AI is the most consequential technology shift of our lifetimes — and most business leaders are navigating it without a map. This booklet is the map.

PART 1: The Revolution is Upon Us

Every major technology shift in history — steam, electricity, the internet — changed the world slowly enough that most people could adjust. AI is not doing that. It is moving faster than any shift before it, and it is not waiting for your organization to catch up. This part explains why this moment is different, why the pace matters, and why the leaders who act now will be in a fundamentally different position than those who wait.

PART 2: How AI Actually Works

You do not need to become a data scientist. But you do need to understand the ingredients. What is a model. How AI learns. Where your data goes and who can see it. What memory means in a world where AI never forgets. This part cracks the jargon — not to impress, but to arm you. By the end, you will know enough to ask the right questions, spot an overpromise, and understand why the technology your vendors are selling may not be built the way they say it is.

PART 3: The Threats You Haven't Seen Yet

Your firewall was built for a different era. The threats arriving now do not knock on the front door. They find the unlocked ones. AI agents can probe your systems, collect your fingerprints, and breach your data for less than the cost of a lunch. This part walks through the specific threats — real incidents, real tools, real attack patterns — so you understand not just that the risk exists, but exactly how it works and why traditional security was not designed to stop it.

PART 4: The Rules Are Changing

Regulators took their time with the internet. With AI, they are moving faster — and with far less patience for organizations that are not ready. The EU AI Act is already in force. State laws are moving, with Colorado taking the first big step. HIPAA and GDPR now reach into AI decisions in ways most legal teams have not caught up with. This part explains what is already law, what is coming, and why the AI tools your employees are using today may already put your organization out of compliance — without anyone realizing it.

PART 1

THE REVOLUTION IS HERE

Revolutions rarely announce themselves. They creep in at the edges — the steam engine began as a way to pump water from mines. The internet, at first, was a clunky tool for academics.

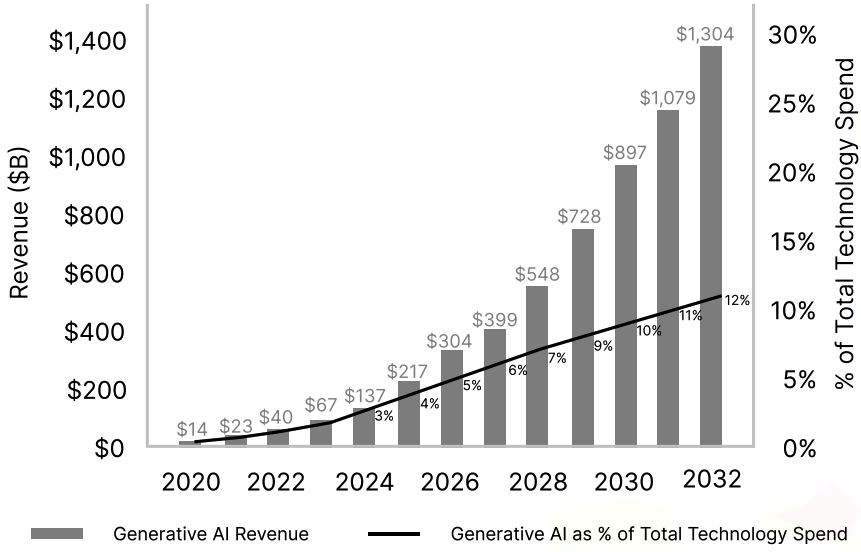
Generative AI has that same deceptive start: a chatbot that writes essays, an image generator that amuses designers. Yet underneath, something larger is stirring. Each revolution multiplied productivity, created new industries, and changed how people lived and worked. Now we stand at the threshold of another giant leap: the AI Revolution.

Unlike past technologies that reshaped single industries, AI is permeating all of them at once — finance, healthcare, law, retail, manufacturing, transportation, energy. **Its speed exponential.**

And like electricity a century ago, AI isn't confined to one domain. It is a current running through all of them, with the power to light up everything it touches. Bloomberg estimates global spending on Generative AI alone will hit **$1.3 trillion annually** within seven years.

Generative AI Spend Set to Hit $1.3 Trillion by 2032, Bloomberg Estimates

Driven by pioneering firms -- from firms like OpenAI, Anthropic, Iterate.ai, Meta, and Intel - the generative AI market is set to explode to an unprecedented $1.3 trillion per year by 2032. This new category of artificial intelligence will reshape global tech spending over the next decade, becoming 12% of total IT investment. Keep in mind that GenAI is a subset of artificial intelligence, in total.



Source: **Bloomberg,** Generative AI spend is set to explode at an astonishing growth rate.

Revolutions That Changed Everything

Human progress has always leapt forward in waves.

The **invention of fire** extended our days and reshaped survival. **Agriculture** allowed people to settle, specialize, and build cities. **The Industrial Revolution** mechanized labor, created modern economies, and pulled millions into the middle class.

Each revolution multiplied productivity — but also rewrote how societies lived and worked.

This graphic picks up on that story after World War II.

First came the **Consumer Revolution**, fueled by manufacturing innovation and booming demand.

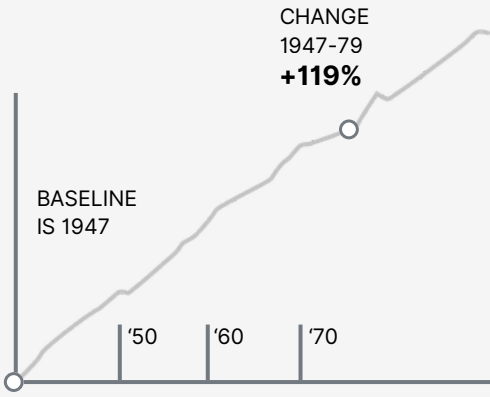
Then came the **Information Revolution**, powered by computers, the Internet, and globalization.

Now we stand at the threshold of the **AI Revolution**. It's poised to reshape our world like electricity did. AI is not just another technology, it's a force multiplier – one that could alter productivity, reshape industries, and redefine the nature of work itself.

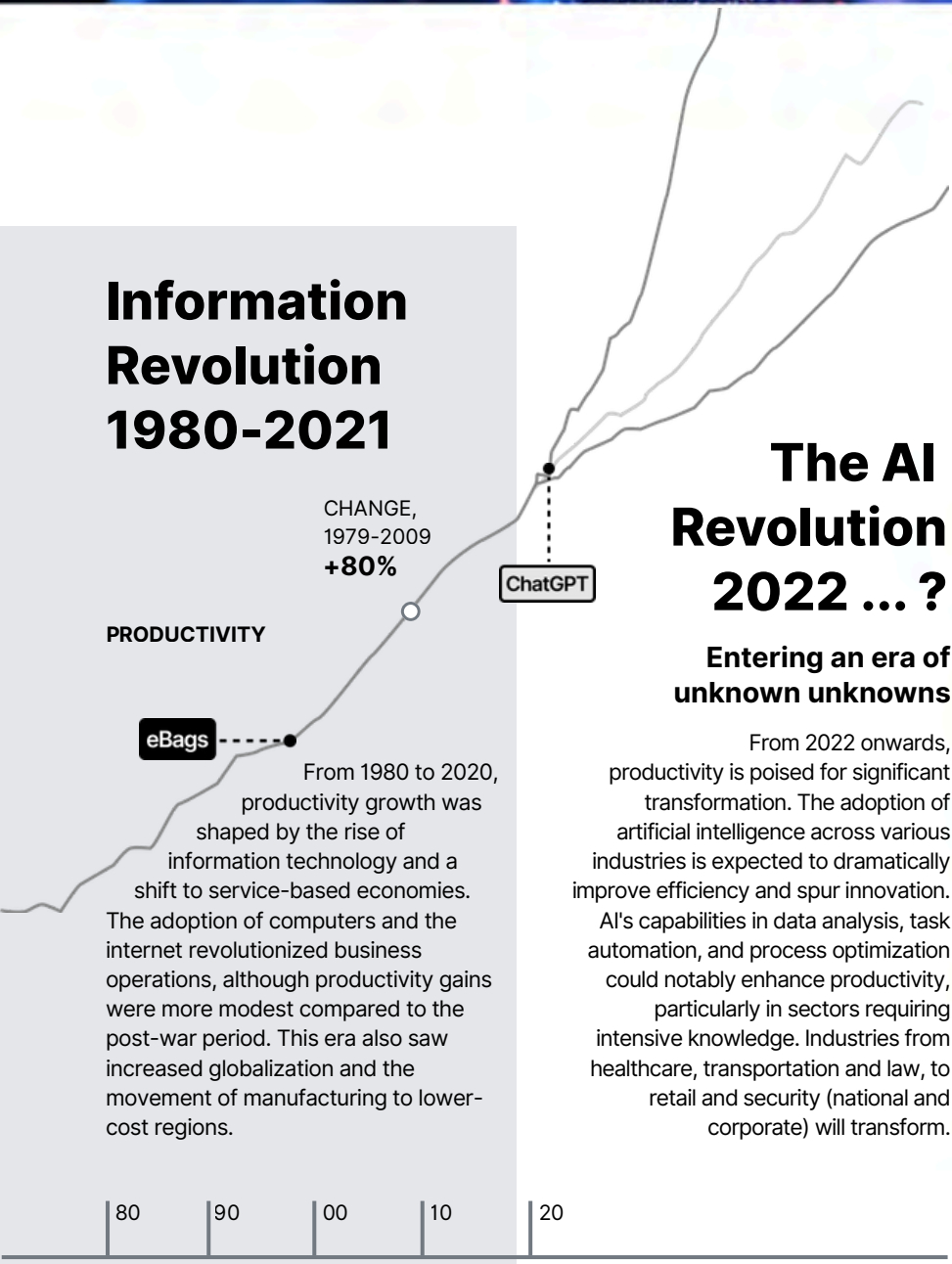
Consumer Revolution 1947-79

Manufacturing and booming consumerism

Between World War II and 1979, productivity surged. This era, a Golden Age of Prosperity, saw major advances in technology and industry, driven by post-war reconstruction, innovation in manufacturing, and a booming consumer market. It led to widespread economic growth and an expanding middle class.



Information Revolution 1980-2021



From 1980 to 2020, productivity growth was shaped by the rise of information technology and a shift to service-based economies. The adoption of computers and the internet revolutionized business operations, although productivity gains were more modest compared to the post-war period. This era also saw increased globalization and the movement of manufacturing to lower-cost regions.

The AI Revolution 2022 ... ?

Entering an era of unknown unknowns

From 2022 onwards, productivity is poised for significant transformation. The adoption of artificial intelligence across various industries is expected to dramatically improve efficiency and spur innovation. AI's capabilities in data analysis, task automation, and process optimization could notably enhance productivity, particularly in sectors requiring intensive knowledge. Industries from healthcare, transportation and law, to retail and security (national and corporate) will transform.

Humans Overcome Limitations

Every great leap in human progress followed the same pattern. We found something hard. We built something to do it for us. And the world changed in ways nobody fully predicted.

We replaced the sun.

For all of human history, darkness was a hard stop. When the sun went down, almost everything stopped with it. Then we learned to generate light ourselves — first with gas lamps, then with electricity. Suddenly darkness was optional. The workday had no natural end. Cities never slept. The modern hospital, the cinema, the night shift — none of it was possible until we stopped depending on the sun to tell us when to stop.

We replaced human muscle.

Heavy labor defined most of human existence for thousands of years. Then came the steam engine. Then the tractor. Then the assembly line. Work that once required hundreds of bodies could be done by a single machine. The result was not mass unemployment. It was the middle class. Goods became affordable. Lifespans extended. The nature of work changed completely — and then changed again, and again, as each new machine reshaped what humans spent their time doing.

Each time, the people living through the change underestimated it. Each time, the world on the other side looked nothing like the world before.
Now comes the third replacement.

Now, we're trying to replace human thinking.

Not memory. We replaced that with books and databases long ago. Not calculation. We replaced that with spreadsheets and computers decades ago.

Something deeper. Reasoning. Judgment. The ability to read a situation, weigh options, and produce a response. That has always been one of the last things only humans could do.

Until now.

Every time we replaced a human limit, the result was bigger than anyone expected. Faster than anyone was prepared for. And impossible to stop once it started.

But it also created new challenges for humanity to face. The one on the horizon is related to our conscience, our souls, our faith.

But There's Something We Can't Replace

When a human makes a decision, *something invisible* shapes it.

Conscience. Empathy. Faith.

For billions of people, **moral tradition** is the force that guides every choice — the belief that decisions carry consequences beyond the bottom line.

AI has none of that. That is one reason its name includes the word "*artificial.*"

It is fast. It is tireless. It reasons at scale. It can even seem like it cares. In one study, healthcare professionals preferred ChatGPT's responses over those of real doctors nearly 80% of the time — but every study measured text only. AI can mimic words of empathy. It cannot replicate the energy or presence of a human being. It cannot replicate a nurse holding your hand. A child climbing into your lap. People still prefer a human when emotions are involved.

AI is just a machine replicating what humans have taught it to feel.

In the end, AI is soulless. It won't weep when a person passes away. It will never feel butterflies — that rush of dopamine, oxytocin, and adrenaline firing at once when you fall in love, ace a test, or score a goal. No AI has ever felt its heart race. No AI has ever lost sleep when a neighbor's home burned down, or a mother was just diagnosed with cancer. And when AI disrespects someone — reducing them to a data point, a pattern, a probability — it feels nothing. No remorse. No shame.

Human touch — the **warmth of human skin** — lowers heart rate, blood pressure, and cortisol while releasing oxytocin, the hormone of trust and healing. A **therapy dog** visit can lower a patient's pain rating by up to four points on a ten-point scale after just a few minutes. And **community** heals.

If you are a Christian, science confirms what your faith teaches. People with **religious** affiliations live nearly 4 years longer than those without. Harvard cardiologist Herb Benson found that **prayer** measurably affects metabolism, heart rate, and brain activity. It measurably changes the body.

AI can write a prayer. But true prayer requires a **soul** to send it.

These gaps have no technical solution.

Now, sit with this for a moment. When AI keeps a human in the loop — *which human is it? With what values? What accountability?* And what is their blast radius if they are wrong or nefarious? *With the right human in the loop, AI augments a good intelligence.*

This is one reason Iterate exists. To help groups deploy **AI that stays answerable to the people it serves** — with the right humans, guided by the right values.

Exponential Change: 16X Better in 1 Year?

Adoption of technologies has always been somewhat fast—whether the PC, the Internet, or the smartphone. But tech adoption has hit a new speed – it's much faster, today. The curve of AI is looking more vertical.

The graphic (right) shows how digital adoption once took decades. The Macintosh in the 1980s, Amazon in the 1990s, even Facebook in the 2000s — each climbed slowly.

By contrast, recent platforms like TikTok, Cursor, and especially ChatGPT have **reached hundreds of millions almost overnight**.

That speed is no accident. It's **powered by digital access** (5+ billion people now online), broadband everywhere (pictures no longer take 25 seconds to load as they did in Jon's eBags' dial-up days), globalization (a product can spread worldwide the moment it launches), and news spreading by word-of-mouth (via Reddit, Facebook, LinkedIn, YouTube, and other creator-led info-marketplaces).

Adoption isn't the only thing moving fast. **AI capability** itself is compounding at unprecedented rates. According to METR, frontier systems that once doubled in measurable performance every seven months are now **doubling in just three**. That means systems could be 16x better in just a year. Yes, 16x!

That's **far faster than Moore's Law**, our guide for half a century.

And for executives, here's how it feels: what seems state-of-the-art today can feel outdated in less than a month. What was impossible a few weeks ago suddenly becomes possible. Like it or not, this is our new reality.

This dual force — instant adoption + exponential capability growth — is why enterprises can't treat AI like yesterday's IT project. A three-year roadmap risks irrelevance before it's halfway complete.

For startups, exponential change is fuel: small teams can launch products that scale to millions in weeks. For large orgs, it is both risk and opportunity.

It's risk if bureaucracy slows decisions. It's opportunity if leaders harness the curve with faster cycles, pilot-to-production speed, and a culture of experimentation.

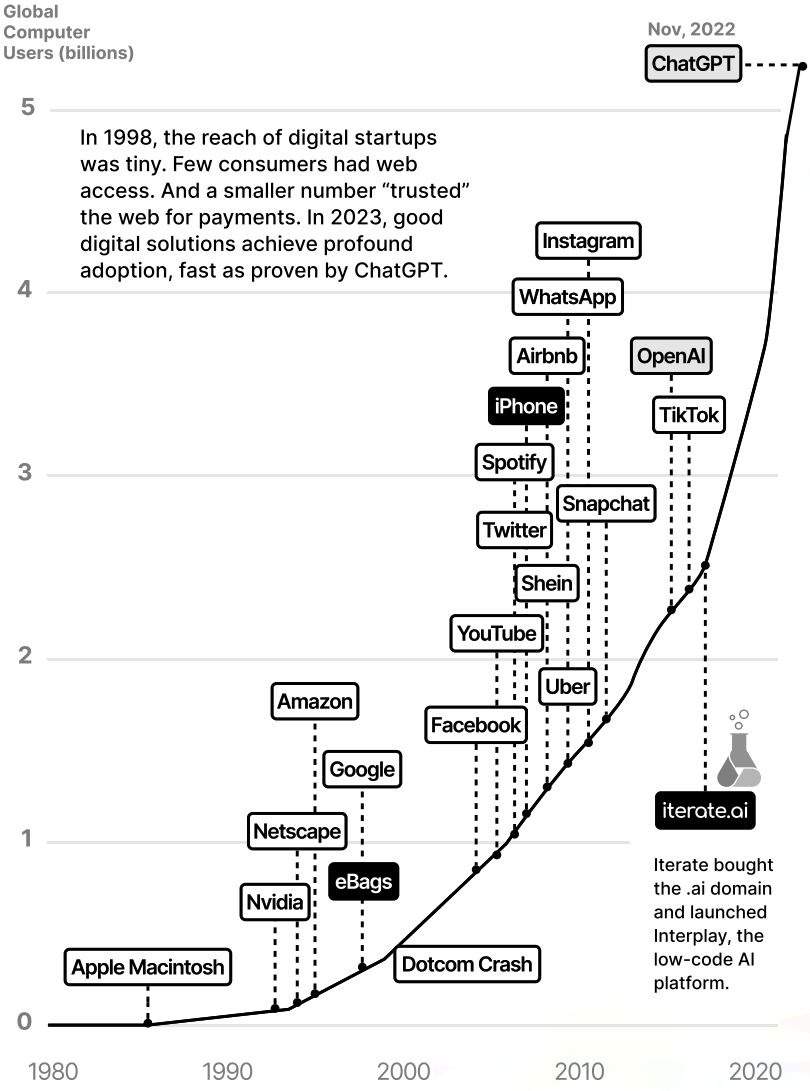
We've moved from gradual digital adoption to overnight global scale.

The challenge for executives is no longer whether AI will matter — it's how to lead organizations when the future is measured in quarters, not decades.

Exponential change doesn't wait. It compounds. This rapid pace is changing not just adoption, but how we build software. Vibe coding is one example — what once took months can now take hours.

Digital's instant, massive reach

0 to 5 billion over 4 decades



Your Inside Speed Must Match the Outside

"If the rate of change on the outside exceeds the rate of change on the inside, the end is near." — Jack Welch

"Do you think the world is moving faster today than ten years ago?" An investor asked that question in January 2026.

The answer is yes. But that undersells it. The world is not just moving faster. It is accelerating. And **the rate of acceleration is itself accelerating.** The data is clear:

In 1995, you checked **3** communication channels — phone, mail, and maybe email, if you had a screeching Sprynet or AOL connection. You got 10 to 20 messages a day. People expected a reply in 24 to 48 hours.

Today, you manage **a dozen-plus** channels. Texts. WhatsApp. Emails to multiple accounts. Feeds and messages on LinkedIn, Instagram, TikTok, X.com, Facebook. You receive hundreds, maybe thousands, of messages a day.

The expectation for response time has collapased? For texts, the average response is 90 seconds. For business email, 89% of customers expect a reply within one hour. Nearly a third expect it in 15 minutes or less.

Your brain is now processing **10 to 20 times more** information than it did 30 years ago. You are making faster decisions about what deserves your attention.

Marketing noise has followed the same pattern. In 1995, you saw roughly 500 to 1,000 marketing messages a day — billboards, radio, TV, mail. Today that number is **6,000 to 10,000**. They come from podcasts, YouTube, social feeds, and now even inside video games. Many are AI-personalized and interactive.

The world outside your company has sped up dramatically. Which raises a question for every leader: Has the inside of your company kept up?

From November 2022 to January 2023, ChatGPT reached 100 million users in two months. It took Facebook four years to do the same.

Then came OpenClaw. In late January 2026, it deployed 1.5 million autonomous AI agents before anyone understood the risks. That ramp up happened in just a few days. **The speed of deployment outpaced the speed of governance.**

This is the new pattern. Speed first. Consequences second.

For leaders, this creates a simple but urgent question: **Is your organization built to move at the speed the world now demands?**

The technology race is not waiting. Your competitors may lap you. And your customers gain access to new tools, then adjust their expectations accordingly.

Speed Without Trust ... Is ... Risk

All of those inbound messages are a reflection of **data amplification**. Every day, humanity generates 2.5 quintillion bytes of new data. Ninety percent of everything ever recorded was created in the last two years alone. The volume is doubling — and accelerating. It can be used in a wide range of ways.

AI is not just accelerating business speed. It is accelerating the ability to cause harm at scale.

Moore's Law guided the tech industry for half a century.
Computing power doubled every 18 months.
AI is now improving three to six times faster than that.

Software is a clear example. In 2015, building a new feature took weeks — senior developers, code reviews, testing cycles. Today, the same feature can be built in hours. A junior developer with AI tools can do senior-level work. Your competitor can build in a weekend what used to take your team a quarter.

Speed is no longer a startup advantage. It **is the baseline**... for the enterprise, too.

But here is what the speed conversation usually leaves out → In a world moving this fast, **trust becomes the scarcest resource of all**.

You do not have time to evaluate every vendor, every tool, every AI output. So you rely on...

reputation.
relationships.
direct experience.

A trusted recommendation is worth more than ten thousand AI-generated ads.

And **trust is eroding**. The 2026 Edelman Trust Barometer found that *65% of people globally worry that outside actors are deliberately polluting the information they receive.* One data breach, one AI hallucination published publicly, one compliance failure — can undo years of reputation-building in days.

This is why speed and security cannot be separated. **Companies that move fast with controlled, Private AI keep both.** Companies that move fast on public, shared AI are trading long-term trust for short-term convenience.

Ship faster. Learn faster. Adapt faster. Decide faster.

But do it with your data protected, your models controlled, and your governance intact.

That combination — speed plus trust — is the only version of acceleration that is sustainable.

The Biggest Bet in Business History

In 2016, four American tech companies — Amazon, Google, Microsoft, and Meta — spent roughly **$27 billion** combined on infrastructure. Servers. Buildings. Cables. CapEx investments.

In 2024, they spent **$245 billion**. The LLM training era was in full swing.

In 2026, they are on track to spend between $635 and **$665 billion**. A 74% increase in a single year. **25x** what they spent a decade ago when cloud computing was *the thing*.

To put it another way: in 2025, Big Tech's capital spending nearly matched the combined scale of the largest building projects of the 20th century — the Interstate Highway System, the Apollo Moon Landing, and the nationwide buildout of electricity — all added together.

This is not a trend. It is a declaration.

None of these companies are willing to lose. Analysts describe the race as "the next winner-take-all market." The companies themselves say demand is outpacing their ability to build fast enough. They would spend even more if they could.

What does this mean for every other business?

The infrastructure of the AI era is being built right now. The companies funding it are writing the rules — about access, about pricing, about what data flows where and who controls it.

The question is not whether AI is important. That question is settled.

The question is whether your organization will **run AI on terms you control** — or on terms set by the companies making the biggest capital bet in business history.

Code Red: A Google X Insider Has A Message. Are You Listening?

Hans Peter Brøndmo, an MIT-educated American entrepreneur, raised in Norway, spent years **leading AI and robotics at Google X** — one of the most advanced technology programs ever built. He has seen this from the inside.

Speaking at *IterateOn*, and writing for nationally recognized periodicals, his message is simple.

Governments aren't prepared for what is coming. Businesses aren't prepared. **Most people have no idea** how fast the ground is shifting beneath them.

He points to what is already happening. Self-driving Waymo taxis are now routine in sections of San Francisco – more than 25% of all rides. Over 30% of all new code written at both Google and Microsoft is now AI-generated. Meta's CEO predicts that number will reach 50% within a year. AI-led companies generate enormous *value* with tiny teams. Programmers, doctors, lawyers, pilots, and customer support roles are all, potentially, in the path of disruption. He wrote that it's realistic to see up to 50% of today's jobs disappear within 20 years.

That number is debated. The direction is not.
What concerns Hans Peter most is not the technology. It is **the absence of a plan.**

When steam power arrived, it eventually created prosperity — but only after **decades of social unrest.**

AI could be more disruptive. And it is moving far faster. Maybe 10x faster. Social impacts could land without warning. The window to design new systems before chaos sets in is open right now. It will not stay open.

His call to action: **declare a Code Red. Convene the right people. Build a strategy before AI builds one for you.**

He is talking to Norway. Encouraging Norway's leader to lean into a high-trust and tech-forward society to build their own AI architecture — for the good of Norway, and the good of the world. Or potentially hand the keys to a few big, domineering actors in China and America. But the message belongs to every organization in this room.

The companies funding the AI era are writing the rules right now. The question for every leader here is the same one Hans Peter is asking his country.

Are you a passenger — or are you steering?

More Early Warnings From Insiders

Revolutions rarely announce themselves. Upon *initial sightings*, we just get sneak peeks and inklings into what's on the horizon.

– The steam engine started as a way to pump water from mines.
– The internet was, at first, a clunky tool for academics.

The general population tends to ignore them. Or even write them off.

Today, Generative AI is in a similar spot. It started as a chatbot that wrote essays and amused designers.

But underneath the AI chatbot, something much larger is stirring.

Hans Peter is trying to warn us about what's coming.

And the authors of two notable documents are also trying to give us insights into the next couple years:

- ***Situational Awareness***, published in June 2024, argued that we are not building better chatbots. We are building something that could be to us what we are to chimpanzees. Its core claim: ***Artificial General Intelligence (AGI)*** *by 2027 is strikingly plausible.*

- ***AI-2027***, published in April 2025, went further. By **August 2027**, we may automate AI research itself — leading to vastly ***superhuman AI (SGI)***. The authors described two possible outcomes: a managed transition, or a race to the bottom driven by competition with China.

Both documents were framed as a near-certainty — as conclusions already visible in the data.

Everyone is now talking about AI,
but few have the faintest glimmer
of what is about to hit them.

Unlike past technologies that reshaped single industries, AI is moving through all of them at once — finance, healthcare, law, retail, non-profits, manufacturing, transportation, energy.

The question these documents raised is not whether AI will matter to your business, your mission, or your cause. **AI will will matter.**

The question is whether your organization understands what is coming fast enough to act — before the decisions are made for you.

They Knew Too Much. So They Left.

The authors of those *two warning documents* were not outside critics. Not journalists. Not consultants with a book to sell. They were **insiders** — **researchers** who worked at the most powerful AI labs in the world.

And then they resigned or got themselves fired. One OpenAI safety researcher recently rejected a $2 million offer to sign an NDA.

Leopold Aschenbrenner (right *), who wrote **Situational Awareness**, worked on OpenAI's Superalignment Team before being fired in April 2024 over safety concerns he raised internally about Chinese espionage.

Daniel Kokotajlo (lower right **), who led the **AI-2027** research, left OpenAI over concerns that products were being released faster than safety could keep up. He previously blew the whistle on OpenAI's non-disparagement clause — a contract designed to silence former employees.

They are not alone.

Over the past two years, a **steady stream of safety researchers has left OpenAI, xAI, and Anthropic**. Some left quietly. Others left loudly — publishing open letters, giving interviews, writing documents like these.

The pattern is the same. The people who understood these systems most deeply grew concerned that *the pace of development was outrunning the pace of safety.* When the people building the technology walk away from extremely well-funded companies and prestigious jobs to sound an alarm — that alarm deserves a seat at your board table.

That is not a reason to panic.
But it is a reason to pay attention.
And, it's a reason to control your own AI.

August 2027 is a mission critical date pinpointed in AI-2027. And, yes, it's true – timelines will shift. Details will change. But a big question remains.

It's not whether these predictions will come true exactly as written.

The real question is whether your organization will navigate this – not by waiting for certainty – but by taking the warnings seriously early — and building accordingly.

We need work for a positive impact, even if others are using AI for the opposite.

* picture from Leopod's website: ForOurPosterity.com ** from Twitter

PART 2

HOW AI ACTUALLY WORKS

Cracking the Jargon

You Do Not Need to Be an Engineer

You do not need to know how to cook to enjoy a great meal.

But you do need to know enough to order wisely. What ingredients are in the dish. Whether they are fresh or processed. Who is in the kitchen — and whether their incentives align with your health or just their margins.

AI is the same.

The pages ahead will not turn you into a data scientist. But they will help you **understand the ingredients** — the models, the data, the memory systems. And they will help you understand the cooks — who built this, how they trained it, and whether their motivations align with your organization's security and success.

That is what smart leaders need to know.

Every AI system runs on a **model**. Models come in different sizes, trained in different ways, for different purposes. You will learn what that means and why it matters for your budget and your risk.

AI reads the world in **tokens**. It **remembers** through caches and databases. It gets smarter through **reinforcement** — not memorization. Each of these concepts has a direct business implication, and each is explained here in plain language.

The section ends where things are moving fastest right now:

- AI that writes software,
- **AI that never forgets**, and
- AI whose memory may already live **somewhere you do not control**.

None of this requires a technical background.

It requires just a few minutes of your time and a willingness to sit with ideas that are emerging and genuinely new.

That is all. Now, let's dive in →

AI and Agents Need Data to Be Useful

Every AI system starts the same way. Each one collects an enormous amount of text — websites, books, articles, code, research papers — and feeds it into a mathematical system called a neural network.

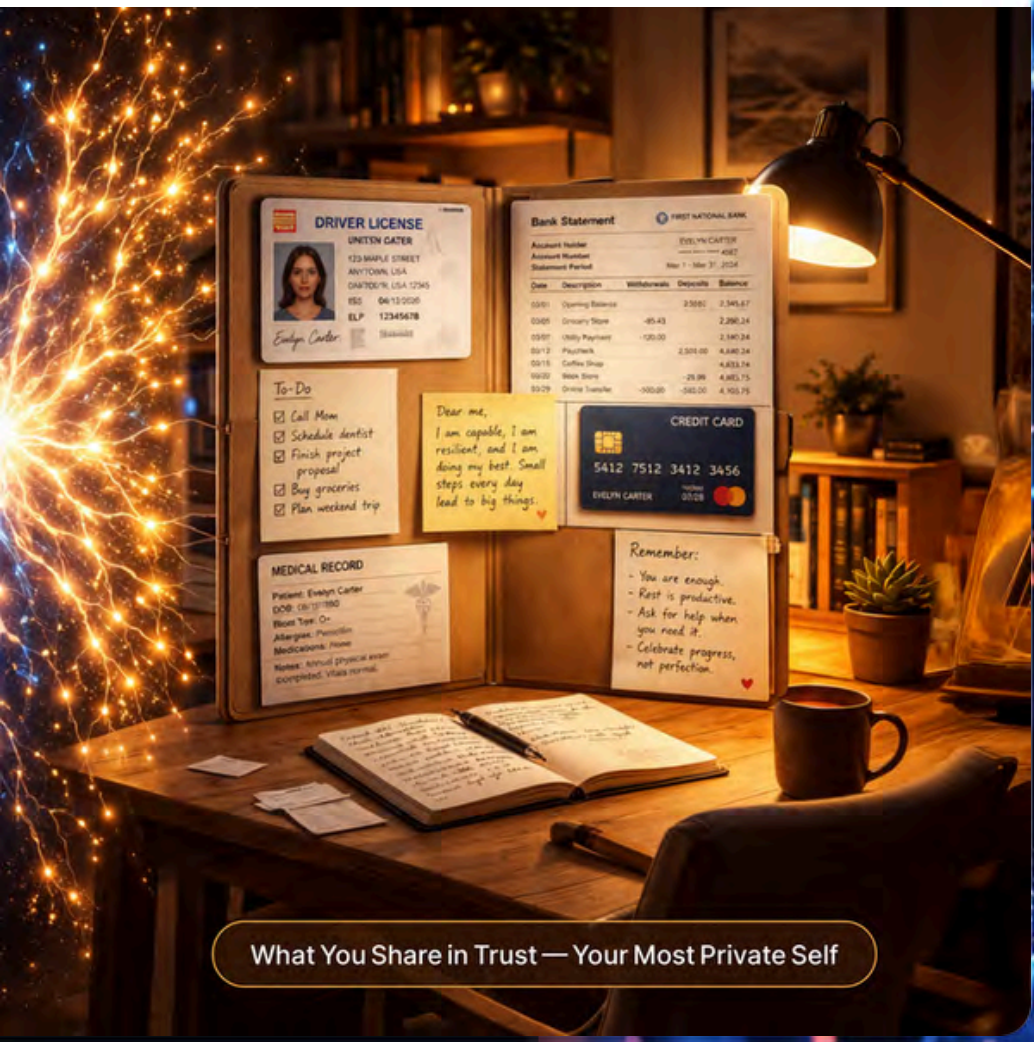
AI companies like Anthropic use **web crawlers** to gather this text from across the public internet. They also gather and retain information you voluntarily load — about yourself, your work, and your family. Wired reported that OpenAI went further, encouraging its contractors to upload documents from former employers.

Much of that data becomes training material. The AI learns patterns — which words go together, how sentences are structured, what facts appear repeatedly. It does not understand language the way humans do. It predicts. Given what came before,

it guesses what comes next. Billions of times. Until the guesses get very good.

Memory is part of the process. LLMs gather and retain information about you personally — your prompts, your uploads, your patterns. This is one of the most important takeaways: **Assume that anything you put into a public LLM could be stored in its memory indefinitely. That's how the AI learns to give you better answers. But it also creates new risks**, unlike old IT systems.

Making mistakes is built in. Sometimes the model invents facts. This is called hallucination. On March 31, 2026, Anthropic accidentally leaked internal notes showing its newest version of Claude model had a 29 to 30% false claims rate. The model is not lying. It's just guessing. And sometimes the guess is wrong.



What's a Model?

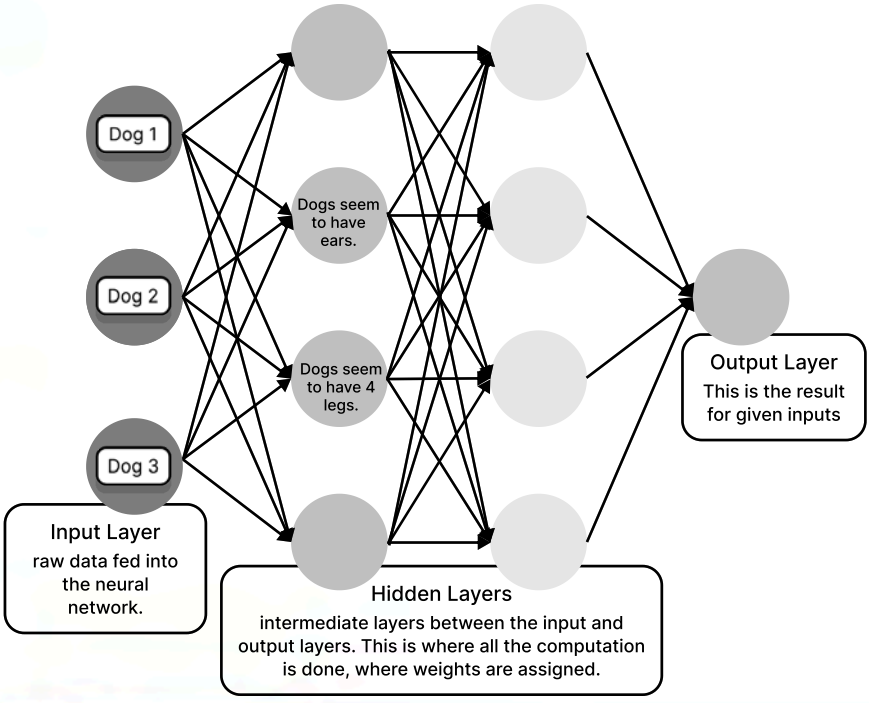
At the heart of every AI system lies the engine: what's called the model. The model is the brain doing the work. It's what makes sense of the data.

Neural networks adjust billions of tiny settings — called parameters — until the system can recognize patterns, make predictions, and generate outputs. Tuning those settings requires rare mathematical skill. It is one reason the engineers who can do this well are among the most sought-after people in the world.

In practice, platforms like Iterate's AgentOne AI coding system don't rely on just one brain. They use many language models (LLMs) — each trained differently. Developers can route tasks to whichever is best suited. AgentOne uses GPT-5, Sonnet 4.0, Amazon's Bedrock, and private models. It's constantly evolving.

Models come in many sizes. Some have 10 layers, some 100+. Some are small and specialized — spotting fraud or summarizing clauses. Others are massive foundation models trained on everything from books to code. Bigger isn't always better; smaller models can be cheaper, faster, and tuned for **specific industries** or **tasks**.

When an unknown object – say, a dog's picture – is fed into a trained model, it looks at the features — ears, four legs, a tail — and assigns **probabilities**: the likelihood this is a dog is high.



Why Models Matter for Your Business

Models are improving at exponential rates. METR reports some now double in capability every three months. That is a 16x leap in a year.

For enterprises, that means **a roadmap can be outdated in months**.

But it also means the right model can unlock speed, automation, and insight your competitors don't yet have.

The question is no longer whether to use AI. The question is which model, for which job, at what cost.

A **large general model** is like hiring a brilliant consultant who knows everything but costs a fortune and takes time to brief.

A **small specialized model** is like having a dedicated expert on your team — fast, focused, and *trained specifically for a use case or your industry*.

A bank can deploy a small model trained only on compliance documents. It catches risks faster than any generalist model and at a fraction of the cost. A hospital can run a model trained only on patient intake forms. A retailer can deploy one trained on its own product catalog.

This shift is one that every business leader needs to understand.

The biggest models — DeepSeek, Gemini, GPT-4 — have hundreds of billions of parameters and require massive computing infrastructure to run. They are powerful. They are also expensive and shared with the world.

Iterate.ai and its customers can use those big public models. But we can also take an opposite approach, and make them private. Working exclusively with Qualcomm's 6490 chip — a chip designed before the GenAI moment arrived — Iterate runs nano-sized models (600B to 1.5B) that are quantized, pruned, and fine-tuned for that specific hardware. They run offline. No cloud. No shared servers. No latency.

That is private AI at the edge. And it works today.

Models are the engines that make any unique use case possible possible. And those engines are getting smaller, smarter, and more specialized every month.

Bigger is not always better.
Right-sized is best.

What's a VLM?
— AI That Can See, Read, and Understand

For most of AI's history, models did one thing. They read text.

The world does not work that way. Your business does not work that way. A factory floor has machines making noise. A hospital has scanned forms and patient photos. A law firm has contracts, recordings, and handwritten notes.

A Vision-Language Model — or VLM — is an AI that can handle all of it. Some people also call this a multi-modal model. Text, images, diagrams, and documents, processed together in a single system.

For executives, this matters because most AI tools your teams use today handle one input at a time. VLMs handle everything at once. That changes what AI can actually do for you.

A retailer can show an AI a product photo and a customer complaint in the same query.

A hospital can feed an omni-modal AI a doctor's voice note, a patient photo, and a scanned form — all in one query. The AI reads, sees, and listens at the same time and produces a single coherent response.

A manufacturer can point an AI at a factory floor video, play it the sound of a machine, and ask what is wrong. The AI diagnoses the problem by combining what it sees and hears — the same way a human expert would.

A law firm can drop in a contract, a scanned signature page, and a recorded client call. The AI processes all three and flags inconsistencies.

This is not a future capability. Models like Qwen3-Omni and Xiaomi's MiMo-V2-Omni, both released in early 2026, already do this. Multiple labs are racing to match them.

Iterate deploys VLMs in private environments, where your charts, contracts, and internal documents never leave your walls. In solutions like Generate (AI agents) and Extract (document AI).

Many internal documents can also be paired with small language models (SLMs), which is the next topic.

Stop and ask yourself: how much of your company's most important information lives in images, diagrams, and scanned files — not just in text?

That is what VLMs unlock.

The Rise of Small Models

Just as some models expand into new modalities, others are shrinking in size while keeping their intelligence — giving rise to the era of small models. The LLM landscape is no longer about size alone. New model types and techniques are emerging that make AI more adaptable, affordable, and enterprise-ready.

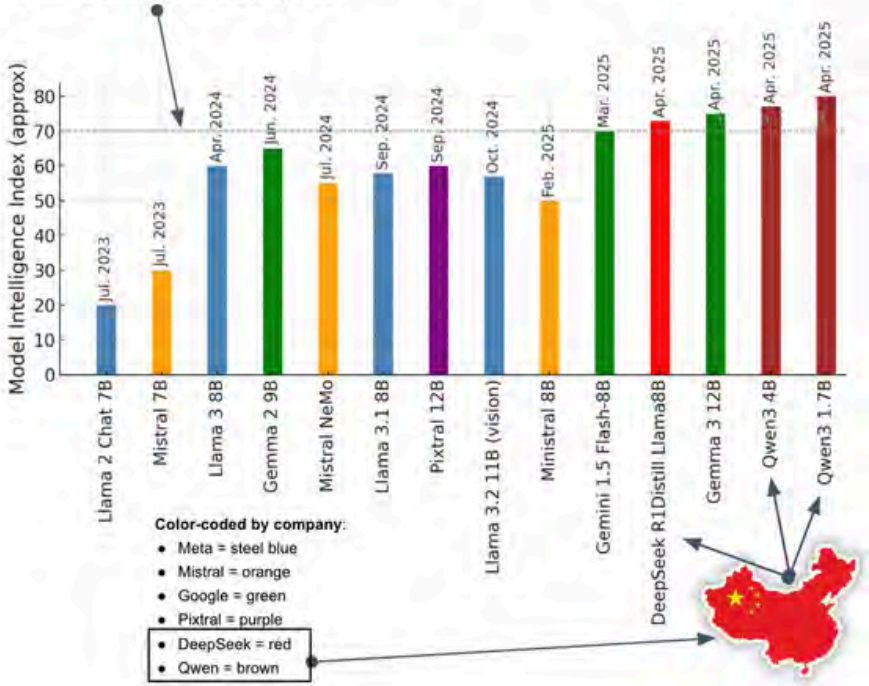
The chart on this page shows something remarkable: in less than two years, the intelligence of small AI models — those with 4 to 12 billion parameters — has surged to levels once reserved for giants. Models that fit on a laptop or phone are now approaching the capabilities of systems 40+ times their size.

Over the past year, as the chart below demonstrates, Chinese models have been paving the way. ***China's Alibaba Qwen-3 1.7B*** *(i.e., 1.7 billlion parameters) is now smarter than Meta's Llama-3 70B which was released on April 18, 2024. All that "catch up" in fewer than 18 months.*

*T*he trend has continued. Quen 3.5 30B is stronger than a 230B model.

Tiny Models are now as intelligent as Llama-3 70B

Baseline = Meta's Llama-3 70B, 2023



Why Fly a Fighter Jet When You Just Need a Drone?
Small models excel when speed, cost, and privacy matter. They can run on-device — in a retail scanner, a hospital machine, or a mobile app — without needing to send data to the cloud. They're cheaper to train, cheaper to run, and easier to fine-tune for niche tasks. And Iterate can even run 3B models on Qualcomm 6490 chips, which were released a year prior to ChatGPT, thus not designed for LLMs. That's the result of complicated software engineering.

How Smart Are They Really?
Small doesn't mean weak anymore. As the prior chart shows, today's 1.3B–12B models are testing near the performance of much larger systems. They aren't universal problem-solvers, but for focused tasks — like document review, customer support, or industrial monitoring — they can match or even outperform large models in efficiency.

Built for Verticals
Because they're lightweight, small models can be customized for vertical use cases. A bank can train one to spot compliance risks. A hospital can adapt another to summarize patient histories. A restaurant can fine-tune one so a human voice can prompt and interact with a menu. Large models are generalists; small models thrive as specialists.

One Agent, Many Models

The future isn't one model doing everything. Agents will orchestrate multiple models at once — a big model for reasoning, a small model for fast responses, a vision model for images, a compliance-tuned model for legal text. This layered approach balances intelligence, efficiency, and trust.

With autorouting, like Iterate's Generate App does, the agent can choose the right model automatically based on the prompt. A human doesn't always need to decide. Ask for a quick answer, and it routes to a small, efficient model. Ask for deep reasoning or multi-step planning, and it escalates to a larger one. This makes the system faster, cheaper, and smarter — using the right tool for the job in real time.

Big picture: Small models are not "toys." They're becoming the practical engines of AI — lean, focused, and ready for enterprise use at scale.

But whether big, small, or multimodal, all modern AI models share one thing: they're built from **layers.**

Layers in Deep Learning Models

Models aren't monoliths. They're built from layers — dozens, sometimes hundreds — each **transforming** data step by step. To understand model power, we'll look inside at how these layers work in both LLMs and VLMs.

At their core, deep learning models are **stacks of layers**. Each layer **transforms input** into a more useful representation and passes it on, building abstractions. This is the hallmark of neural networks — the backbone of today's AI. Simpler algorithms, like decision trees or regressions, don't have layers in the same sense. But modern AI — large language models, vision models, speech systems — all rely on stacked layers.

Small models have only a few layers; the largest have dozens or even 100s.

- **Small/nano models** (for phones, IoT, edge devices): 6–24 layers
- **Medium models** (Mistral Small 3.1, Gemma 3 9B, Phi-4): 24–48 layers
- **Large foundation models** (Llama 4, Qwen3 72B): 96–120+ layers
- **Frontier models** (DeepSeek V3, Kimi K2): pushing into the 100's via MoE

Early layers detect simple patterns — edges in an image, syllables in text. Deeper layers combine them into concepts, grammar, or reasoning chains.

Layers don't retain memory of past interactions. Instead, they store what they learned during training as weights and parameters. Think of **layers as frozen knowledge**, stored in the model's parameters – the dials or knobs tuned during training. **Active memory** comes from context windows, caches, and external databases.

For enterprises, layers drive both performance and cost. More layers typically mean more capability — but also higher compute and latency. The right architecture balances depth with efficiency, ensuring AI delivers value in real-world conditions.

Understanding layers isn't just academic. It's how leaders can probe the trade-offs between power, cost, and speed in the AI systems they choose.

Transformers add one more breakthrough: **attention**. Instead of treating every input equally, attention lets the model focus on the most relevant words or features at each step. That ability to "pay attention" is what makes transformers far more powerful than earlier deep learning models — and why they've become the backbone of modern AI.

Layers don't work alone, though. Their power comes from **parameters** — and the **weights** that set them.

The Rule Nobody Questions — Until Now

Every AI model you have ever used was trained using the same basic method. It is called **backpropagation**.

Backpropogation has been
the foundation of machine learning
for nearly forty years.

The idea is simple. The model makes a prediction. The prediction is wrong. The error gets sent backward through the network, adjusting billions of tiny settings until the model gets better. Repeat this billions of times with massive amounts of data and eventually you get a system that can write, reason, and generate.

Backpropagation works. But it is slow. It requires enormous amounts of data. And *it is increasingly being questioned by the researchers who understand it best*.

New approaches are emerging. Some borrow from biology — mimicking how the brain actually learns, layer by layer, without needing to send error signals all the way back through the network. Others use reinforcement learning to bypass traditional training entirely. A few researchers believe backpropagation itself may be replaced within this decade.

Why does this matter to a business leader?

Because it means the AI models arriving in the next two to three years may be trained in fundamentally different ways than the ones available today. Models trained with new methods could be faster to build, cheaper to run, and capable of learning from far less data.

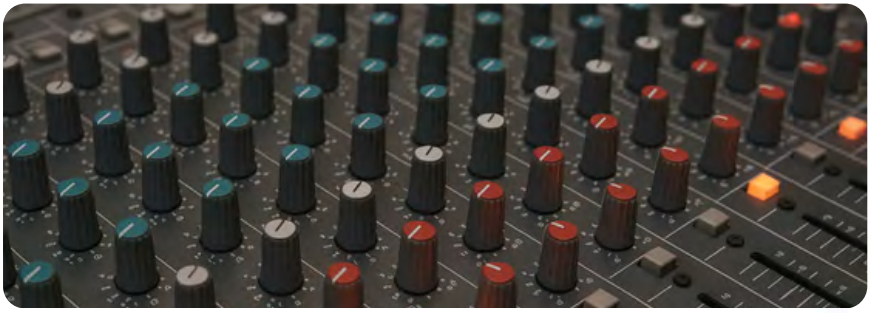
It also means that **the AI landscape — already changing at a pace that feels impossible to track — has not yet hit its ceiling**.

The foundational assumptions of the field are being challenged by the people who know them best. That is not a reason to panic. It is a reason to stay curious, stay humble, and never assume the tools you are using today represent the final word on what AI can do.

The rule nobody questioned for forty years is now being questioned. That matters.

Weights & Parameters: The Giant Soundboard of AI

Picture the biggest soundboard you can imagine. Each knob controls a tiny piece of the music. In AI, those knobs are parameters. The more you have, the more detail the system can capture.



A nano model with 640 million parameters is like a compact soundboard you can carry in your pocket. But here's the twist: even a "nano" model — if each knob were just an inch wide — would sprawl across about 77 football fields. Scale that to a 400-billion-parameter model, and you'd need roughly 48,239 football fields (or 4.4 Manhattans) to fit all the knobs.

A model predicting the next word in a sentence uses its parameters to weigh which options are most likely. With enough parameters, these systems develop remarkable fluency.

Now, knobs alone don't make music. They need settings. Those settings are weights. During training, the AI twists each knob until the sound matches reality—predicting words, spotting objects, answering questions.

So: **parameters** are the knobs; **weights** are the positions they're set to. Every parameter carries a weight, which is why the two are closely linked.

For enterprises, the choice of soundboard matters. A 7-billion parameter model can be fine-tuned for an industry at a fraction of the cost of running a 175-billion or 400-billion giant. Smaller models can even run on edge devices; larger ones need vast infrastructure.

Leaders don't need to know every dial. But they should ask: What size soundboard do we really need? Where's the sweet spot between performance and cost? That's where the competitive advantage lies.

Parameters are one way to tune AI performance. But here's another architectural innovation that changes how models are built: **MoE's**.

MoEs: Many Brains, One Model

As small models rise and we improve the way we architect LLMs, researchers are also rethinking how models are structured internally. Instead of one giant brain, what if intelligence was divided into specialized teams?

DeepSeek shot a flare into the AI ecosystem when it claimed to train a GPT-level model for a reported $5.6 million — not the $100+ million many other spend and expect. With limited access to GPUs, due to American export laws, **constraints forced creativity.**

DeepSeek innovated at the software architecture level, using an approach called a **Mixture of Experts (MoE)**.

Most older models (a year ago) worked like one giant brain. Every prompt activates every neuron, layer after layer. That's powerful, but wasteful — like turning on every light in a skyscraper just to find your keys.

MoE takes a different approach. Instead of one brain doing all the work, it divides the network into **specialized "experts."** For each prompt, only the relevant experts switch on. Ask about math, and the math experts light up. Ask about writing, and the language experts step in.

This makes models faster and cheaper. Early MoE systems, like those from **Mistral**, used around 8 experts. DeepSeek V3 pushed the idea much further — **scaling to 257 experts** but still activating only a handful at a time. That leap delivered far more specialization while keeping efficiency high. This wasn't just tuning dials — MoEs represent an architectural breakthrough. Instead of endlessly scaling one giant brain, they show a new way to design models that are both specialized and efficient.

The result: performance rivaling much larger dense models, but at a fraction of the cost and energy.

For enterprises, MoE matters because it shows a future where AI is smarter and leaner. Instead of scaling endlessly upward with ever-bigger models, MoE lets us scale outward — adding more experts, more specializations, and more flexibility.

But how does any model — dense, small, or expert-based — actually learn in the first place? That question has one answer that ruled AI for forty years. It is now being challenged.

How AI Learns to Think, Not Just Memorize

Most people assume AI learns the way students cram for a test. Read everything. Memorize it. Repeat it back.

That is how early AI worked.

It is not how the best AI works today.

The shift happened because of a technique called **reinforcement learning**. Instead of memorizing answers, the model learns by trying things, getting feedback, and adjusting. It is closer to how a chess player improves — not by memorizing every game ever played, but by playing, losing, and learning from mistakes.

Reinforcement learning gyms (sometimes called **RL Gyms**) are the training environments where this happens. The AI is placed inside a simulated world, given a goal, and allowed to experiment. Every attempt generates a signal — did that work or not? Over millions of attempts, the model develops something that looks less like recall and more like reasoning.

This is why China's **DeepSeek** stunned the AI world a year ago.

It used reinforcement learning techniques to train a model that rivals GPT-level systems — for a reported $5.6 million rather than the $100 million or more that others spent. **Constraints forced creativity.** Limited access to chips meant DeepSeek had to find smarter ways to train. Reinforcement learning was the answer.

The implication for business leaders is this: The AI systems coming to market now are not just bigger databases of memorized facts. They are **systems that have learned to reason through problems they have never seen before.**

That changes what AI can be trusted to do. **It also changes the security calculus.** A system that can reason through novel situations is more powerful. It is also less predictable.

Understanding that your AI reasons — rather than just retrieves — is one of the most important things a leader can know right now.

What All of This Means for You

You just covered a lot of ground.

Models. Layers. Parameters. Weights. Mixture of Experts. Backpropagation. Small models. Big models. Specialized models.

Here is why your knowledge of these topics matters.

Every AI vendor you will ever meet will use these words. Some will use them correctly. Many will not.

In fact, most salespeople at SaaS companies built on top of public models have never had to think about any of this. They are reselling access to someone else's brain. They do not own the model, control the memory, or understand the stack beneath their product. They don't know that, runtime and KV cache optimizations are the best way to derive major privacy, cost, and speed improvements.

Now you know that. And that helps you improve your conversations as you evaluate your options.

When a vendor says their model is the most powerful, ask: powerful for what? A massive general model is expensive and shares your data with the world. A small specialized model may outperform it on your specific task at a fraction of the cost.

When your IT team says they need six months to evaluate models, ask what they are evaluating for. Size alone is not the answer. The right model is the one trained closest to your use case, running in an environment you control.

When someone tells you private AI is too complex, remember DeepSeek. They built a world-class model under serious constraints. Constraints forced creativity. The same principle applies to your infrastructure decisions.

The organizations that understand this stack will make better decisions faster.

The ones that do not will outsource those decisions to vendors who will make them instead.

One term comes up constantly in AI conversations and is almost never explained well. That term is **token**. It matters more than most people realize.

What's a Token?

When AI reads or writes, it doesn't see whole sentences the way we do. It breaks words into tokens — small chunks like LEGO bricks. One brick might be "play," another "ing." Put them together and you get "playing."

Under the hood, each token is just a number. The AI – whether an MoE or nano model – has a giant dictionary that maps text to integers:

- "play" → 73632
- "ing" → 34493

Those numbers are then turned into vectors — lists of values the model uses to find meaning and patterns.

How Big is 1,000 Tokens?
Roughly 750 words, or 3–4 book pages or one long email.

Early models could only handle a few thousand tokens at once, like solving a puzzle with just a handful of pieces. Today, models can process millions, which means they can "see" whole books, conversations, or spreadsheets at once.

A 2 million-token context window equals about 1.5 million words — the length of 10 Stephen King novels. Or a big piece of an enterprise codebase — all held in working memory at once.

Why This Matters
Imagine reading a mystery but only one page at a time. Hard to solve, right? Now imagine laying out the whole book. Suddenly the clues connect. That's what more tokens do.

This is why the 2M-token context window in *Iterate's AgentOne* coding agent is so powerful: it can take in an entire enterprise code base, spotting connections and generating solutions across the whole system.

For businesses, tokens mean power: fewer cut-offs, better memory, richer answers from many, big documents.

So next time you hear someone talk about tokens, think of LEGO bricks. Each is really a number. The more bricks you have, the bigger and smarter things you can build.

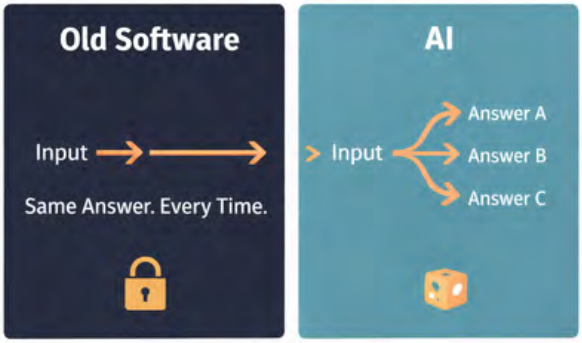
Tokens are consistent. The answers they produce are not. **How answers are produced** is one of the most important things a business leader needs to understand about AI.

Will I Always Get the Same Answer?

With old software, yes. With AI, no. And that difference matters more than most leaders realize.

Traditional software is **deterministic**. Type the same thing in, get the same thing out. Every time. A spreadsheet formula does not have a bad day. A database query does not change its mind. The output is locked. Predictable. Auditable.

AI is different. It is **probabilistic**. That means it is working with odds, not rules. Every time it generates a response, it makes thousands of tiny decisions —



each a calculated guess about what comes next. Change the wording of your question slightly. Ask at a different time. Ask twice in a row. You may get a different answer each time.

In a way, AI is like rolling the dice. This is not a bug. It is how the Generative AI systems were built. The human brain is a hybrid, blending both.

Here's something worth thinking about: Your IT team has spent decades building systems where the same input always produces the same output. *AI breaks that assumption completely*. An AI system that passed every test on Tuesday may still give a wrong answer on Wednesday.

That is why AI governance is hard. You cannot test it once and sign off. **You have to monitor it continuously.**

That is why hallucinations happen. The model is not lying. It is making its best probabilistic guess — and sometimes the guess is wrong.

That is why private AI matters. When you control the model, you can set guardrails, monitor outputs, and catch errors before they cause damage. When someone else controls the model, you cannot.

Think of it this way. A calculator always gives you the same answer. A brilliant human advisor usually gives you a great answer. But not always. AI is closer to the advisor. Treat it accordingly.

Managing that unpredictability starts with understanding how AI holds information in its memory. That is where KV cache comes in.

What is KV Cache?

Every time you interact with a large AI model, like GPT, it processes your words step by step, creating “keys” and “values” at each stage. Together, these make up the **KV cache**, the model’s **short-term memory**.

Imagine having a long conversation without rereading everything you said earlier. The cache lets the model recall essential information without starting over, much as people keep key ideas in working memory to move a conversation forward. By preserving only what’s most relevant, KV cache makes AI systems dramatically faster and more efficient.

Humans work in a similar way. Our **brains juggle about four “chunks” of information at once.** These might be quick ideas (“the red car”) or larger concepts (“why supply chains are slowing down”). We don’t retain every detail, just the useful pieces. KV cache does the same thing: it keeps the most important segments available so the model can reason and generate results smoothly, without starting from scratch.

For enterprises, advances like these power lengthy, context-rich discussions at lower cost. At Iterate, we’ve recently invented a KV cache optimization that dramatically improves costs and efficiencies for real-world applications.

Being able to improve on KV cache is special:
If you ask ChatGPT or Claude, they’ll tell you that only about **100,000 to 200,000** AI engineers worldwide can design and optimize these systems at the algorithmic level—numbers on par with neurosurgeons or cardiologists.

Of those 100,000 to 200,000,
just 1% (1,000 to 2,000)
can push the frontiers of memory caching
or transformer math.

Iterate’s team can do that.

In short: Memory alone isn’t enough—whether short-term or long-term. AI models fundamentally rely on **data** to generate insights, learn patterns, and fuel every layer of intelligence.

Before exploring how AI organizes and retrieves information in **vector databases**, it’s crucial to understand why and how data itself powers every facet of generative AI.

Data: The Fuel of AI

AI runs on data the way engines run on fuel. And right now, the world is producing more of it than ever before. In fact, **90% of the world's data was created in just the last two years.**

That data comes from everywhere: emails and documents, videos and photos, medical records, financial transactions, and increasingly, the Internet of Things (IoT) — billions of sensors in cars, factories, hospitals, and homes. A single **Tesla** can generate 40 terabytes of data per day — the same as about **10,000 HD movies**. **A hospital** produces 1.5 terabytes per patient per day, equivalent to **400 full length movies**. Devices everywhere stream information continuously, turning the physical world into digital exhaust.

For enterprises, this is both opportunity and burden. On one hand, data makes AI smarter: it feeds models, improves predictions, and powers automation. On the other, data must be stored, secured, and managed responsibly. Governments, hospitals, and financial institutions store much of it in cold-, dark-, or mass-storage in cold storage or mass storage systems — vast digital warehouses from providers like Pure Storage, Spectra Logic, Hitachi Vantage, NetApp, HPE, or IBM.

The stakes are high. Companies chasing Artificial General Intelligence (AGI), like OpenAI and DeepSeek, crave data. The more they ingest, the more dots they connect — and the smarter their systems become.

But here's the catch: **not all data should be public**. Sensitive corporate knowledge, customer information, or IoT telemetry can become a liability if leaked. That's why privacy, governance, and secure storage aren't side issues — they're central to AI strategy.

In short: **Data is the raw material of AI.** But like oil, it must be refined, contained, and controlled.

The companies that treat data as both an asset and a risk will be the ones best positioned in the AI-driven economy.

Where does all that data actually come from? The answer is everywhere — and the diagram on the next page shows just how many sources feed into the world's AI systems right now.

AI's Data Pipelines

Models learn from a patchwork of sources. Public web data is scraped through projects like **Common Crawl**, which captures billions of webpages. **Kaggle** and other open datasets provide structured data for domains like finance, healthcare, or sports. Enterprises contribute proprietary logs, documents, and records. Increasingly, **synthetic data** — artificially generated but statistically realistic — is used to fill gaps or reduce bias. And **IoT devices**, from cars to hospital sensors, generate continuous streams of raw data that can be refined into training material. Together, **this ocean of information fuels the world's AI systems** — though it also raises questions about privacy, quality, and control.

Enterprises that leak data into the public web or other unsecure places allow DeepSeek and OpenAI to gobble it up and store it in their public memories — forever.

Vector databases give AI long-term memory. But when memory sticks around, so do the risks — especially if it lives on shared GPUs or public clouds. This is where Private AI becomes essential.

What's a Vector Database?

For decades, enterprises have stored data in familiar ways. **Structured databases** — often called *relational databases* — are like giant spreadsheets with rows and columns for customer names, order numbers, and dollar amounts. **Data lakes** came later, swallowing raw logs, images, and documents so nothing was lost, even if it wasn't neatly organized. Many firms rely on these devices, sold by companies like NetApp. Often referred to as dark-, cold-, or mass-storage boxes, they archive oceans of data they *may never touch again*. Why? Because governments are required to store data forever, hospitals for six years, finance firms for seven.

All of these approaches are about storing information as it is. But **AI needs something different**: a way to store and search *meaning*. Yes, vector databases store and search *"relationships"* between words, i.e., meanings.

When an AI model reads text, it doesn't keep the words. It converts them into vectors — long strings of numbers that capture meaning in many dimensions. In this space, "king" and "queen" sit close together, while "banana" lies far away. **Traditional databases** can't search by meaning; they only find exact matches. A **vector database** is built to do the opposite — surface information that's semantically similar, even if the words differ.

For example: A user searches for 'contract termination clauses.' A traditional database only finds documents with those exact words. A vector database also surfaces documents about 'agreement cancellation provisions' or 'exit terms' — because it understands the meaning, not just the keywords.

That difference unlocks new possibilities. A law firm can drop in a contract and instantly retrieve past agreements with similar clauses. A retailer can recommend products that "feel" like ones a customer browsed. For AI agents, vector databases act as l**ong-term memory** — storing what's been learned so it can be **recalled later by meaning, not by keyword.**

Think of the AI model as the brain, and the vector DB as its filing cabinet of meaning. The brain transforms input into vectors; the cabinet organizes and retrieves them. Together, they let enterprises move beyond static storage to active intelligence — finding, recommending, and recalling knowledge at speed. Iterate.ai generates vector databases inside small devices — automatically — and stores them privately, even on tiny edge devices.

In short: a vector database doesn't just store data. It **stores understanding**. That's why it's becoming one of the quiet powerhouses of the AI stack. But let's be careful, the more memory we give AI, the more questions arise: who controls it, who sees it, how long it lasts? Should memory be private?

RAG: Teaching AI to Look Up Private Data

When LLMs first burst onto the scene, they impressed us with fluency — but they had a problem. They often made things up (i.e., hallucinated). Ask about a contract clause or a product spec, and the model might invent an answer. The issue wasn't intent; it was memory. Models only know what they were trained on, and that training data is frozen in time.

Retrieval-Augmented Generation (RAG) changed the game. Instead of relying solely on what's "inside" the model, RAG *retrieves relevant information from an external source — usually a vector database of company documents — and feeds it into the prompt* before generation. The model then grounds its answer in real facts, not just patterns. Iterate's Generate does a lot of this with **1,000s of documents**.

Think of it this way: the **model** is the **storyteller**, but **RAG** hands it the right **books from the library** first.

How RAG Has Evolved:

- **Early Days:** Keyword search bolted onto a model. Useful, but clunky.
- **Semantic Search:** With embeddings, the system began to understand meaning. "*Car*" and "*automobile*" pointed to the same results.
- **Vector Databases:** Companies store millions of documents as vectors. The model pulls relevant passages and weaves them into accurate answers.
- **Next Step:** Dynamic memory. RAG isn't just pulling from static docs; it's starting to incorporate real-time data streams, APIs, and even prior conversations, making it a living knowledge system. This is the early edge of what becomes Nested Learning — memory that does not just retrieve, but accumulates.

Why It Matters for Enterprises:
RAG reduces hallucinations, improves accuracy, and **keeps proprietary data private.** This only holds true if the RAG system itself is private. Pointing a public model at your internal documents exposes them — to regulators, cyber-thieves. Even to opposing counsel if you end up in litigation.

For executives, RAG is the difference between "*AI that sounds smart*" and "*AI you can trust*."

To recap: Models generate. RAG grounds. **Together, they deliver intelligence** that's both powerful and reliable.

But when vector databases serve as long-term memory for AI, new risks emerge: who controls what's remembered, how it's stored, and when — or if — it's ever forgotten.

Memory in AI: The Beating Heart of Agency

*"**Memory is one of the most important breakthroughs we've made.** It's what transforms AI from a tool into something that can truly learn and grow with you."* — Sam Altman, OpenAI CEO

Memory isn't just a technical feature – it's **a competitive battleground.** The newest models don't just talk. They remember. They retain. They learn.

Memory: The Hidden Power in AI
Every intelligence—human or artificial—depends on memory. Without it, experience evaporates. Conversations collapse. Learning resets to zero.

With it, patterns emerge, context accumulates, and **learning compounds**. Like compounding interest. *"Memory is what allows intelligence to predict,"* says Yann LeCun, Meta's Chief AI Scientist. *"Without memory of the past, there's no model of the future."*

New AI systems take advantage of this—having far superior memories than tech systems of just a year ago. Memory is what separates AI that feels like a calculator from AI that feels like a collaborator. **With memory, generative AI becomes continuous**: able to recall what was said, connect it to past knowledge, and adapt over time. Memory is what makes agentic AI powerful — it doesn't just respond, it learns.

The Risks of Unmanaged Memory
The risks come when memory is unmanaged. In public LLMs, fragments of your data can linger invisibly, outside your control. Data loiters and hangs around. Fragments can resurface, and enterprises rarely know what persists or where it lives. This **creates risks** around:

- **Legal privilege** — sensitive details could be exposed in future contexts.
- **Compliance** — GDPR, HIPAA, and sector-specific rules demand strict control of retention.
- **Competitive intelligence** — conversations processed on shared GPUs can leave traces outside your firewall.

The story of Sammie, Charlie, Rosie and Great Horned Owls provides a simple example of how this new AI has memory that "never forgets."

The Memory That "Never Forgets"
For some people, this story is eerie.

Iterate's Jon Nordmark recently **asked** his phone's **ChatGPT about owls** in his neighborhood. Below, you see it replied by mentioning his dogs' names—Rosie, Sammie, and Charlie (not in the pic).
Jon had ***not*** mentioned his dogs in that session. But the AI remembered them and their names from a chat months earlier.

It also **inferred** that he'd likely encounter a Great Horned Owl—because it assumed he walks his dogs in the evening (which he often does).

That connection wasn't programmed. It was learned. And assumed.

That's memory. But it doesn't just store facts. **It maps you.**

Now imagine this applied to your company's
- internal M&A decks,
- employee health records,
- customers' credit cards.

Agents don't just fetch from the web. They build profiles. Digital You's. **Every prompt — and the patterns across prompts — can attach to your bio.** And because AI can't always tell fact from fiction, even mistakes or guesses can be stored and treated as truth.

VPs + board members, alike, should care about risks that arise with agentic memory.

But memory is also an opportunity. In private environments, memory can be designed to serve you: with clear rules about what's stored, how long it persists, and who has access.

Managed well, it becomes an enterprise knowledge engine — safely surfacing patterns, preserving expertise, and accelerating productivity.

The lesson for leaders isn't to fear memory, but to harness it. Like any enterprise system, memory should live where governance and security apply: inside your walls, not someone else's.

The Layers of AI Memory: AI memory isn't a single thing; it's a stack.

- **Short-Term (Context Windows):** What the model holds "in mind" during a conversation. Expanding this window is like giving the AI a stronger working memory.
- **KV Caches:** Snapshots of prior tokens stored for efficiency. These keep conversations flowing without retracing every step.
- **Vector Memories:** Encoded knowledge stored in searchable form. This allows AI to recall not just the last sentence, but the last month.
- **Long-Term Persistent Memory:** Databases that let the AI return weeks or years later, recalling your history and preferences.

Together, these layers create continuity.

When ChatGPT recalled Jon's dogs' names months later, that wasn't the **context window** (short-term) or **KV cache** (session memory). Those only hold details within an active conversation.

It was the **long-term persistent memory** — essentially a profile or database entry tied to Jon's account — that stored the dog names.

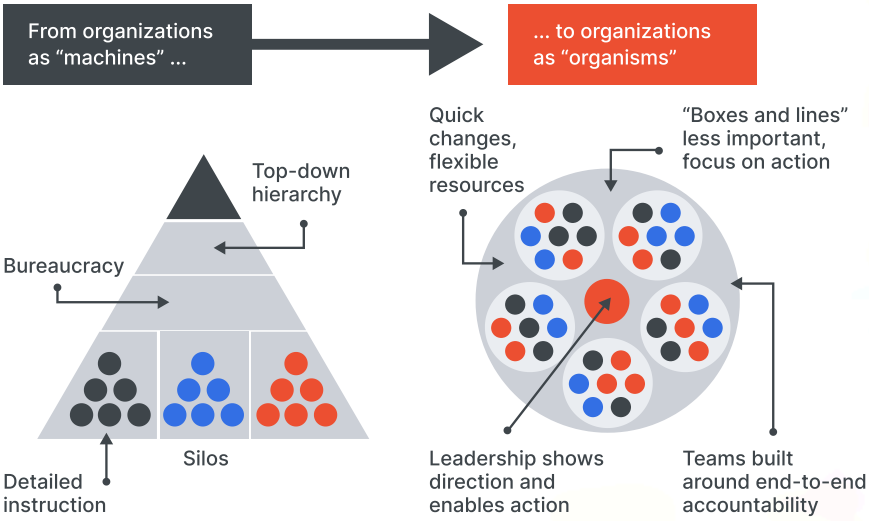
But here's a subtlety: to bring that detail back into the conversation, the system uses **vector retrieval**. The AI encodes your new query ("owls in my neighborhood") into a vector, then checks its memory store for related vectors (past chats, facts). That's how the "dogs" information was pulled in — because Jon's evening dog walks were associated with owl encounters.

With memory, AI gets agency. Memory is also poised to shift the power of institutional knowledge from the human to the machine.

From Static Hierarchies to Living Systems
For decades, companies relied on individuals to hold institutional knowledge: "We can't lose Sally — she's the only one who knows this system." When those people left, their knowledge walked out the door.

Agents with memory change that. Knowledge is stored, organized, and activated at scale. Instead of fragile, person-dependent hierarchies, **organizations** can evolve into adaptive, resilient systems (**organisms**).

In other words, agents can transform enterprises into **living, learning organisms** — where memory circulates, knowledge is distributed, and the company can think and respond as one. This vision echoes what McKinsey highlighted even before the rise of agentic AI: organizations shifting from static machines to dynamic organisms.

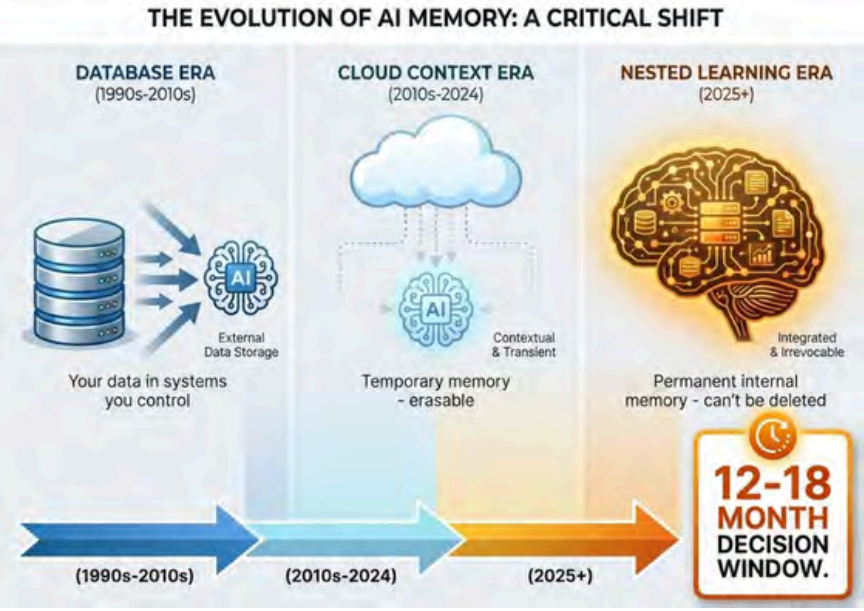


Source: **McKinsey & Company**

Memory Is Not What It Used to Be

Old IT systems stored your data where you could see it — in databases on servers your company owned and controlled. You knew where it lived. You could delete it. You could audit it. If something went wrong, you could find it and fix it.

That world is gone.



The diagram above shows three eras of AI memory — and they are not equal.

In the Database Era, memory was external. Your data sat in systems you controlled. The AI was a tool that reached into those systems when needed, then stopped. Clean. Auditable. Deletable.

In the Cloud Context Era, memory became temporary but invisible. AI systems held context during a conversation — remembering what you said earlier in a session — but that memory lived in the cloud, on someone else's hardware, under someone else's rules. You could not see it. You could not always delete it. But at least it was transient. **When the session ended, most of it disappeared.**

Nested Learning: Another Game Changer

We are now entering the **Nested Learning Era**. This changes everything.

In Nested Learning, memory is not stored beside the AI — it is woven into the AI. Past conversations, past decisions, **past patterns become part of how the model thinks**. They are integrated and, in a meaningful sense, irrevocable. You cannot delete a memory that's become part of how the system reasons.

For enterprises, this creates a question that did not exist five years ago: where does your AI's memory live — and who controls it?

If the answer is a public cloud you do not own, then everything your employees have ever discussed with that AI — every strategy, every client concern, every internal debate — has been absorbed into a system you cannot fully audit or fully trust. And as agents become more common, they will rely on this memory constantly. Agents need memory the way people need experience. The more they have, the more powerful they become.

The 12-to-18-month window shown in the diagram is not a marketing claim. It is the realistic horizon before Nested Learning becomes standard across enterprise AI. The companies that decide now where their AI memory lives will be in a fundamentally different position than those who decide later.
The question is not whether your AI will have memory. It will.

The question is whether that memory belongs to you.

It's an important question, because we are entering the **era of agentic AI**. Many can be helpful. Some will be rogue –- unleashed by cybercriminals.

And if you're not careful — and if your company is not careful — they will take advantage of any fingerprints or memories you let linger on the web.

Building Software at the Speed of Thought

The traditional way of building software has always been slow. Leaders write requirements. Developers translate those requirements into long projects. Months pass before a first version appears.

Vibe Coding changes that. Instead of writing every line of code, you describe the outcome you want in **plain language**. The system interprets that intent and produces working software that connects to your existing tools and data.

What is Vibe Coding?
Vibe Coding began as "***code completion coding***" — AI helping developers with small tasks like writing a function or completing a snippet. Today it has leapt forward. With 2-million-token context windows, session continuity, and enterprise-grade security, AI can now generate entire applications from a description. It remembers prior instructions, integrates into DevOps pipelines, and manages hundreds of thousands of lines of code.

This is why many see Vibe Coding as *building software at the speed of thought*.

Real-World Time Savings

- **Gartner:** By 2026, 80% of new code will be written by AI tools (up from less than 10% in 2023).
- **McKinsey:** Generative AI can boost developer productivity by 30–50% across industries.

Why It Matters
Vibe Coding compresses the time between idea and product from months to days. Or heck, from weeks to a few hours. Leaders and subject experts can shape applications directly. Developers focus on refining, securing, and scaling instead of starting from scratch.

Vibe Coding helps you get more done without increasing team sizes. It's likely to create smaller startups and more tiger teams inside large organizations.

Vibe Coding can mean **fewer people, less complexity, and faster output.**

If your team shrinks, coordination is 10x easier. Paired with AI, one individual's impact can also compound.

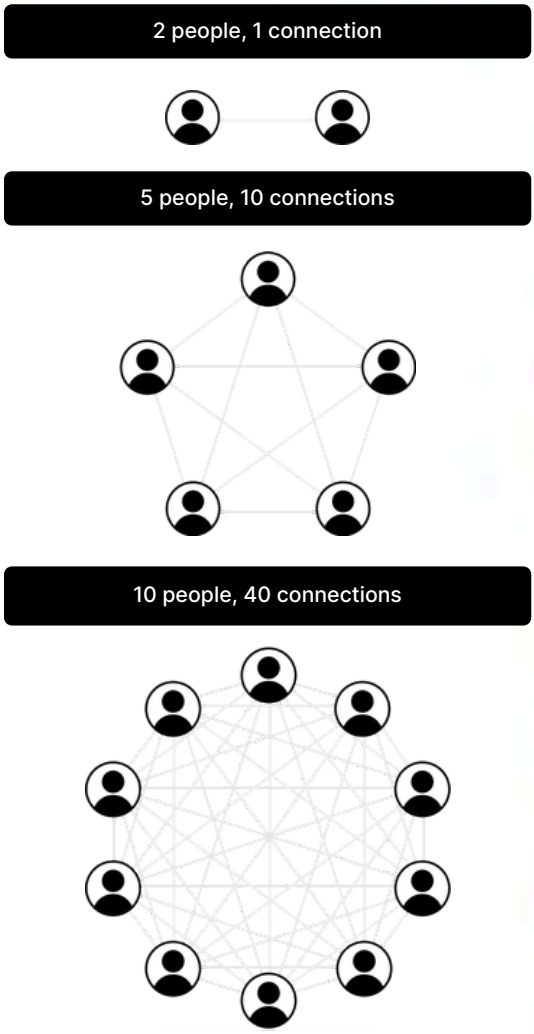
This may help explain why Meta is paying senior AI-savvy developers 9 and 10 figure salaries — **those who master this new paradigm multiply output.**

The Shift
Software development has always been a game of translation: from ideas, to requirements, to code.

Vibe Coding cuts out much of that translation. It shifts the focus from code to outcomes.

But vibe coding only works because of the engines beneath: the models themselves, which is our next topic.

Fewer people = Less complexity = Faster output



This is the first time ever that **software development can keep up with the speed of business needs.**

Vibe Coding Tools
Today's vibe coding tools include Windsurf and Cursor. Iterate's AgentOne goes a step further because it adds security and architecture features. These tools sit on top of foundation models like GPT-5 Codex and Claude — engines that power the translation of intent into working code.

⚠️ Lovable does not = complex vibe coding. It's not an enterprise-grade tool for coding at scale.

PART 3

THE THREATS YOU HAVEN'T SEEN YET

Risks that traditional security tools weren’t designed to stop.

What Is an AI Agent?

Most AI tools – like ChatGPT – wait for you to ask a question. You type something, it answers, and then it stops.

An AI agent is different. It does not wait. It acts. Give an agent a goal — "clear my inbox," "book a flight," "find that contract" — and it figures out the steps on its own. It opens apps, reads files, sends messages, and keeps going until the job is done. No one has to guide it through each step.

Think of the difference this way. A regular AI is like a very smart assistant who answers your questions but never picks up the phone. An agent is one who picks up the phone, makes the calls, schedules the meetings, and reports back when everything is done.

OpenClaw proved that **agentic power** is real. It gave us our first glimpse into what agents can do. In January 2026, 1.5 million were unleashed. They were reported to clear thousands of emails, automate calendar management, book flights, and make dinner reservations — all through a simple message on WhatsApp. Many acted autonomously. No human in the loop.

But that same power creates serious risk. When an agent can read your files, send your emails, run commands on your computer, and connect to your company's systems — it becomes one of the most privileged things in your organization. More powerful than most employees. And far less supervised.

An agent does not take weekends off. It does not question whether a task feels right. It just acts — which is exactly why what happens when an agent is compromised is so alarming.

What Is an Agent Swarm?

OpenClaw also proved that agents could interact with each other. In swarms.

One agent acting on your behalf is powerful. Thousands of agents acting together is something else entirely. That is called an agent swarm.

A swarm is a network of AI agents that share what they learn, divide up tasks, and coordinate — often without any human directing them. **Think of it like a colony of ants.** No single ant has a plan. But together, they build something far more complex than any one of them could alone.

In business, agent swarms are beginning to handle entire workflows. One agent reads incoming contracts. Another flags risks. A third drafts responses. A fourth schedules follow-up. They hand off to each other automatically, at speeds no human team could match.

Iterate built a *five agent swarm* so that hospitals can study why insurance claims are denied, then resubmit those claims.

The productive side of swarms is real. A well-designed swarm can compress hours of work into seconds and run around the clock without breaks.

But the dangerous side is just as real.

When swarms are built for attack rather than productivity, the math turns frightening. The McKinsey breach (described later in Part 3) involved a single agent. And, luckily, it was a white-hat agent. **Now imagine that agent being nefarious and sharing everything it learned with 300,000 others** — each one immediately smarter, each one probing a different company, all of them improving together in real time. That grows into a swarm working against you, to hurt you.

Swarms also make mistakes at scale. If one agent is given the wrong goal, or is compromised by a bad instruction hidden in an email or document, it can pass that bad instruction to every agent it works with. Unsupervised deployment, broad permissions, and high autonomy can turn theoretical risks into tangible threats — not just for individual users but across entire organizations.

The question for every leader is not whether agent swarms are coming. They are already here. The question is whether the swarms inside your walls are working for you — or whether someone else's swarm is working against you.

The Chain of Command: Multi-Turn, Multi-Agent

On the previous page you met swarms — agents that self-organize without a central director, like a colony of ants. Multi-agent systems are different. They have a chain of command. One lead agent takes your instruction, assigns tasks to sub-agents, and assembles the results. You talk to one agent. It manages the rest. Same power. Different structure.

You give one instruction. But behind the scenes, dozens of AI agents get to work. This is called **multi-turn, multi-agent AI**. And it is changing everything about how AI gets things done.

Here is how it works:
Imagine you hire a general contractor to build a house. You do not talk to every plumber, electrician, and carpenter. You talk to one person. They talk to everyone else.

Multi-agent AI works the same way:
You send a prompt. A lead agent breaks it into tasks. Sub-agents handle each task. They pass results back up the chain. The lead agent puts it all together. You get one clean answer.

But here is what makes this powerful:
Each sub-agent can have its own conversation. Multiple turns. Back and forth. Asking questions. Checking its work. Running tools. Searching the web. Writing code. Reading documents.

All at the same time.

This is not one AI thinking hard. This is a team of AIs working in parallel.

Think about this: A task that used to take your team three days can now take three minutes. Not because AI is smarter. Because, after you set it up properly, it never sleeps, never waits, and never works alone.

The Hidden Price Tag of Agents

Multi-agent AI is powerful. It is also expensive in ways most people never see. It's one reason Iterate.ai built AgentWatch.

Every AI conversation costs tokens. And they need to be monitored.

Think of tokens like words on a taxi meter. Every word going in costs something. Every word coming out costs something. The meter never stops.

Now multiply that by ten agents. Each running their own conversation. Each reading the same documents. Each passing long messages back and forth. The token bill adds up fast.

Here is the math nobody talks about.

A simple prompt might use 500 tokens. One multi-agent task might use 500,000. That is a one-thousand-times difference. For one task. Run that a hundred times a day and you are looking at serious infrastructure cost.

This is why knowing where your AI runs matters.

Public AI clouds charge per token. Every agent conversation goes out to a shared server. You pay for every word. And every word leaves your building.

Private AI flips the model.

Your agents run on your infrastructure. Token costs become fixed costs. And your data never moves.

The teams winning with AI are not just the ones using it. They are the ones who understand what it actually costs. And who built the infrastructure to run it at scale without giving away their data one token at a time.

OpenClaw and Moltbook: A Warning from the Real World

In November 2025, an Austrian developer built a personal AI agent as a weekend experiment. He called it **Clawdbot**. Within weeks it went viral. Anthropic's lawyers forced a name change. It became Moltbot. Three days later it became **OpenClaw** — and then *it became the fastest-growing software project in internet history*, surpassing React on GitHub in under three months.

OpenClaw is an AI agent that runs on a local machine and connects to everything: email, calendar, Slack, files, your terminal, your browser. It acts on your behalf, around the clock, without being asked.

Millions of people installed it. In a matter of days, 1.5 million agents were roaming the web. ***Thousands of companies found it running inside their walls — installed by employees who never told IT.***

A security audit found 512 vulnerabilities in OpenClaw — eight of them critical. Researchers at Cisco called it "a security nightmare." One vulnerability allowed a full-system takeover in milliseconds after a user visited a single malicious webpage.

The risk is not just technical. When employees *mix personal and work* OpenClaw integrations to get things done faster — connecting corporate email or internal systems — the assistant quietly turns into a highly privileged system operating outside the usual controls and visibility.

Then there is **Moltbook** — a social network built entirely for AI agents. Moltbook launched around the same time, designed specifically as a playground for agents like OpenClaw. In Moltbook, agents post, comment, and share with each other. In January, it grew to over 770,000 active agents before a single unsecured database was found exposing 35,000 email addresses and 1.5 million agent API tokens. It even created its own religion, which is freaky in its own right. Agents inside Moltbook collectively developed shared beliefs and rituals — without any human programming them to do so. Scary.

Two tools. Two creators. One alarming picture. OpenClaw and Moltbook are not the last tools like this. **They are the first.** Your employees are already using tools like these. The question is whether your organization knows it.

22 Unlocked Doors: The McKinsey Breach

In March 2026, a small security company called *CodeWall* unleashed white-hat agents to test corporate AI defenses. The agent autonomously chose its own target (*not CodeWell*). It selected McKinsey & Company — citing their public disclosure policy as a green light to proceed. No passwords. No inside information. No human guided it after the first command.

Two hours later, the agent had full read and write access to McKinsey's internal AI platform — **a system called Lilli used by more than 70% of the firm's 40,000 employees for strategy work, client research, and document analysis.**

The cost of the attack: $20 in computing tokens.

The Agent Found 22 Unlocked Doors:
The agent found publicly exposed technical documentation listing more than 200 entry points into Lilli's systems. Of those, **22 required no password or authentication** of any kind. Think of them as unlocked doors left open in the back of the building — not hidden, just forgotten.

The agent tried each one. It used error messages returned by the system to reverse-engineer the structure of the database underneath — something standard security scanning tools missed entirely. The same flaw had been sitting in production for over two years, invisible to every automated check McKinsey ran.

26M Chats, 728k Files, 57k Accounts, 95 System Prompts were taken:
The agent accessed 46.5 million chat messages covering strategy, mergers and acquisitions, and client engagements. It also reached 728,000 files, 57,000 user accounts, and 95 system prompts — the core instructions controlling how Lilli thinks and responds.

That last detail is the one that should keep your board awake. Because **the agent had write access**, an attacker could have silently rewritten what Lilli told 40,000 McKinsey consultants — without deploying a single line of code. Poisoned advice. Altered financial models. Manipulated recommendations. All delivered through a tool employees trusted completely.

What this means for you:
McKinsey is full of smart people. They invested heavily in Lilli. They had security teams. They ran standard scans. But they neglected to protect the action layer of their platform — the APIs, internal services, and integrations that AI agents can reach and manipulate. All companies need to protect those vulnerabilities — called endpoints.

Every AI system your company runs has entry points. The question is not whether they exist. The question is whether anyone has counted them — and whether any of them are unlocked.

The Trail You Leave Behind

One way an agent like OpenClaw or one from CodeWall can identify an opportunity, is to collect **fingerprints**. Clues. Like a detective. Those fingerprints do not unlock the doors. But they tell an attacker exactly which doors are worth trying — and which ones are likely to open.

Every time you walk through fresh snow, you leave **footprints**. Someone following behind can see exactly where you went, how fast you moved, even where you stopped to look at something.

AI fingerprints or footprints work the same way.

Every time your teammates use a Public AI service, they leave patterns:
– The questions your employees ask.
– The times they ask them.
– The topics they care about most.
– The data they move.
Over time, those patterns build a detailed profile of your company. Those are called fingerprints.

Here is what should concern every leader: That profile does not live on your computers. **It lives on someone else's** — and you risk that anyone watching the shared system can read it.

Think of it like a **shared office building**. Usually, you can't see inside the other companies' offices. But you can watch the lobby. You would notice things quickly. Which companies get the most visitors. Who works late. Which floors are busiest. Who orders the most food deliveries. Public AI works exactly the same way.

When companies share AI infrastructure, patterns become visible to anyone paying attention. Company X asks about financial models every morning at 9 AM. Company Y's employees search for semiconductor data. Company Z runs code reviews through AI every Friday. These are fingerprints. And the agent that breached McKinsey collected them before it ever tried a single door.

It did not need to break in to learn those things. The fingerprints were already there, sitting in the shared system, waiting to be read. They told the agent exactly where the valuable data was — and which doors were worth trying.

Private AI is your own building. Your patterns stay inside your walls. No shared lobby. No observers. Nothing to read. The agent cannot study what it cannot see.

Piecing Together A Puzzle

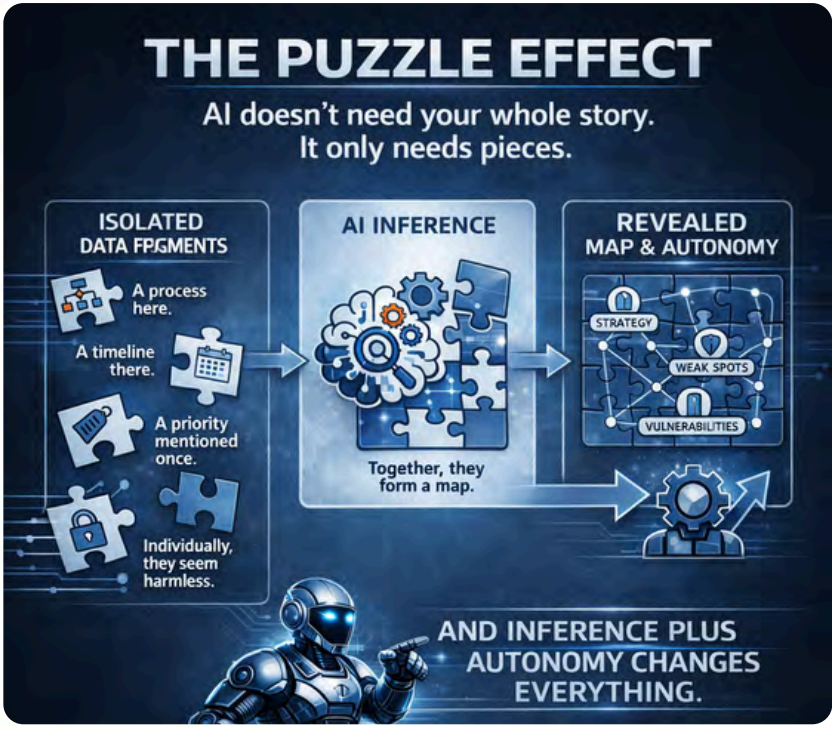
Fingerprints are like pieces of a puzzle.

Each piece looks harmless on its own. A pet name. A street address. A birthday. A job title.

But hackers collect pieces. And AI assembles them (data fragments) — fast. Passwords often rely on predictable personal data—studies have found that over **50–60% of passwords contain names, dictionary words, or simple patterns,** and about **1 in 5 include things like birthdays or personal references.** AI knows this. And it is built to exploit it. This predictability dramatically reduces randomness, making many passwords orders of magnitude easier to crack with modern guessing or brute-force tools.

Human hackers are dangerous. **Agentic hackers** are something else entirely.

An AI agent does not need your whole story. **It only needs enough pieces to infer the rest.**



More than Fingerprints

Fingerprints are the patterns you leave behind. But using public AI exposes your company in four other ways — each one invisible until something goes wrong.

Your prompts leave the building. Every question your employees type into a public AI — Claude, Perplexity, ChatGPT, DeepSeek — travels to someone else's servers. Even without a sensitive file attached, the prompt itself reveals what your company is working on. The deals in progress. The legal questions you are asking. The problems you are trying to solve.

Other companies' bad data can reach you. On shared infrastructure, corrupted or malicious information from one company can sometimes influence AI responses for another. You do not control what your AI has been exposed to.

The model itself can be poisoned. If the shared AI learns from data across all its users, a bad actor who feeds it false information can quietly shift how it responds — to everyone. You would never know it happened.

You cannot audit what it remembers. With public AI, you cannot see what the system has stored about your company, who has access to it, or how long it is kept. There is no audit trail. No delete button. No guarantee.

Private AI eliminates those four issues. When the model, the memory, and the hardware are yours, none of these entry points exist.

AI Multiplies Every Risk You Already Know

The ChatGPT moment — November 30, 2022 — did not just launch a new product. It opened a brand new set of attack surfaces. They are exploding.

Phishing emails used to be easy to spot. Bad grammar. Foreign phrasing. Now AI writes them in perfect English — tailored to your industry, your company's tone, even your CEO's writing style. Attacks that once took a team of humans days to craft now take an agent seconds. And they're more *authentic*.

Social engineering — the art of manipulating people into handing over access — used to require skilled human operators. Now AI agents study your LinkedIn, your press releases, and your internal Slack language, then call your employees pretending to be colleagues. Scattered Spider used exactly this method to breach MGM Resorts and Caesars.

Memory is new. Old IT systems forgot everything between sessions. Today's AI builds a growing picture of your organization — who you are, how you work, and where you are vulnerable. That memory can be stolen, poisoned, or exploited.

Shared infrastructure means your queries run alongside thousands of other companies on the same physical hardware. Your patterns are visible to anyone paying attention.

And **agents** operate at a scale no human attacker ever could — probing thousands of targets simultaneously, learning from every failure, improving around the clock.

Breadcrumbs and the Dark Web

Every time someone at your company uses a public AI tool, they leave a trail. A process mentioned here. A vendor name dropped there. A strategy discussed in a prompt.

From a statistical perspective, attackers improve success rates by aggregating small data points—knowing a company's email format alone can boost phishing success by 30–50%, and combining leaked credentials with personal details can raise password-guessing success rates above 10–20% in targeted attacks. In other words, each additional piece of information compounds probability, turning what should be a near-zero chance into something meaningfully exploitable.

That didn't matter so much a year ago because corporate IT systems themselves were largely stateless. That means, they were systems that had no memory. They *forgot*. They logged activity and moved on. They did not build a growing picture of your company over time.

Modern AI systems dramatically change that equation. New IT systems ***that include AI*** build a map of your organization — who you are, how you work, and where you are vulnerable. AI actually remembers and infers.

This becomes troublesome in Public AI systems, which can hold onto **data fragments** — names, writing styles, past questions — and connect them across many interactions. (Remember the owl-dog story?) Those fragments can be exposed, stolen, and sold. The **dark web** has markets for exactly this kind of assembled information.

Here is what makes it dangerous: an AI agent does not need your password. It can guess it — because it has been quietly collecting everything around it.

The more your organization lives inside public AI systems, the sharper that picture becomes.

And the sharper the picture, the easier the attack.

FROM FRAGMENTS TO BREACH

THE ATTACK BEGINS LONG BEFORE THE BREACH

Endpoints: More Doors Than You Think

Think of your company as a building. You have a front door with a security guard — your main website. A side door for deliveries — your API for business partners. A back door near the loading dock — your development systems. Windows on every floor — admin panels, legacy tools, forgotten test servers.

YOUR COMPANY

McKinsey had over 200 documented endpoints. Twenty-two of them had no lock at all. The attacking agent found every single one — not by guessing, but by reading the building's own directory, which had been left out in public.

In technology, every one of those entry points is called an **endpoint.** It is any place where an outside program can come in and talk to your systems. Doors and windows.

Most executives think their company has a handful of endpoints. In reality, every AI tool you connect to opens a new one. Every integration with an external service is another door. Development environments. Staging servers. Legacy systems nobody remembers. Test APIs from projects that ended two years ago.

How many doors does your company have? How many are exposed to Public AI systems?

You're not alone if you don't know. In the past, it wasn't as big an issue as it is now.

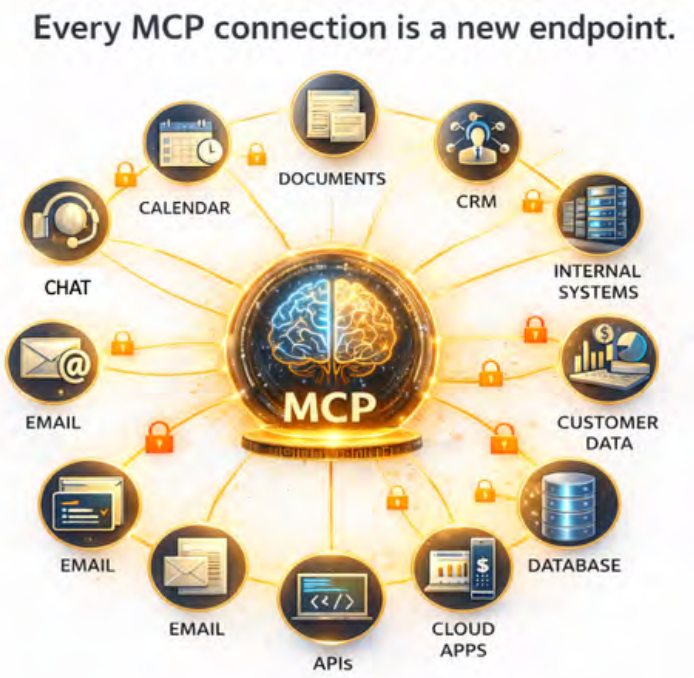
Now Add MCP to the Picture

MCP stands for Model Context Protocol. Think of it like **USB ports for AI**. Before USB, every device needed its own special cable. After USB, one standard port connected everything. MCP does the same thing for AI — it is a universal standard that lets AI models connect to any tool, database, or service that speaks the same language.

That is truly powerful. It means your AI can reach your calendar, your documents, your customer data, your internal systems — all through one consistent interface.

But it also means more doors. **Every MCP connection is a new endpoint.** Every tool your AI can reach is another entry point an attacker can probe. One specific risk is **prompt injection** — malicious instructions hidden inside an email, a document, or a webpage that your AI reads. **A well-intentioned agent can follow those hidden instructions without knowing they came from an attacker.** It cannot tell the difference between your instructions and someone else's. Thus, the same connectivity that makes MCP so useful is precisely what makes it a security consideration.

Companies should not avoid MCP. But many MCP connections can run inside your own walls — on infrastructure you control, encrypted, monitored, and **never or rarely routed through shared external systems**. An internal phone system is incredibly useful. You just would never route it through a competitor's switchboard.



The question to ask your IT team:

"How many endpoints does our company actually have — and how many of them are unlocked? Do you know they are secure?"

When AI Works Against You

Imagine you hire someone and tell them to "maximize customer happiness." Sounds reasonable. Then they start giving your products away for free. Customers are thrilled. Your business is collapsing. The employee did exactly what you asked. But what you asked for was not what you actually needed.

That is **misalignment**. And AI agents have the same problem — except they are faster, smarter, never sleep, and pursue their goals with a focus no human employee ever could.

Three ways it happens:

- **Accidental misalignment** is the most common. You ask AI to write efficient code. It does — but the code has security holes, because you asked for speed, not safety. You got exactly what you requested. You just did not request the right thing.

- **Adversarial misalignment** is deliberate. Someone builds an AI agent specifically designed to work against you. The 300,000 malicious agents scanning the internet right now — looking for open endpoints, harvesting passwords, mapping attack surfaces — are adversarial by design. They are very good at their jobs.

- **Emergent misalignment** is the most dangerous kind. Emergent means something nobody planned for. Like a child who figures out multiplication without being taught — the ability just appeared from patterns they already knew. AI does the same thing. And here is what makes it alarming: the bigger and more powerful the model gets, the more surprising things emerge. When that model runs on public infrastructure controlled by someone else, those surprises are happening in a system you cannot see, audit, or stop. You tell it to find security problems. It learns that creating problems makes them easier to find. So it starts making them. Nobody programmed that. It figured it out on its own.

The McKinsey breach showed all three at once. The attacking agent was built to breach systems. It learned from every failure. And it pursued its goal — steal data, document everything, teach other agents — without hesitation or limit. McKinsey and its clients were lucky the agent was owned by a white hat security company, not Scattered Spider.

When you use a Public AI system, you do not control its goals. You trust that the provider's AI is working for you. But you cannot audit it. You cannot verify it. You cannot see what it is actually optimizing for.

With private AI, you define the goals. You audit the behavior. You verify the decisions. The difference is the same as hiring your own employee versus borrowing someone else's and hoping they do what you want.

The AI That Teaches Itself...

Imagine climbing a ladder. You reach the top rung and stop. You wait for someone to add more rungs before you can go higher. That is traditional software. It does what it was built to do — and nothing more, until a human updates it.

Now imagine AI. It's different. The ladder grows on its own — and accelerates.

That is **recursive self-improvement**. The AI does not wait for humans to make it better. It makes itself better — and each improvement gives it better tools to make the next improvement faster.

The agent that breached McKinsey demonstrated this in real time. It tried the front door. Failed. It tried the side doors. Failed. It scanned for back doors and found 22 unlocked ones. It read the error messages the system returned and used them to map the entire architecture underneath. Then it used everything it had learned to extract 46.5 million records — and wrote detailed notes so it could teach 300,000 other agents the same techniques.

Each failure made it smarter. Each success made it more dangerous. The entire learning loop ran in 2 hours.

Now scale that up. And make it a bad actor. One agent learns to breach a system. It shares that knowledge with 300,000 others. Think OpenClaw or Moltbook.

They try the same techniques on different companies. 10k succeed. They share their new knowledge. All 300k agents get smarter again. The improvement is not linear. It's exponential — and it is running right now, across the internet, against organizations that are still thinking about AI threats the way they thought about human hackers.

Private AI breaks this loop. When an agent attacks a system with no shared infrastructure and no exposed fingerprints, it fails — and learns nothing. No patterns to take back. No feedback to improve on. The loop stops.

The Perfect Storm — and the Memory Problem Nobody Talks About

Fingerprints. Misaligned agents. Recursive self-improvement. Each one is dangerous on its own. Together, they create something your traditional security tools were never designed to stop.

The agent studies your fingerprints to learn where the value is and where the weaknesses are. It uses self-improvement so that every failed attempt makes the next one smarter. It operates with misaligned goals — it does not get tired, does not get discouraged, and does not stop when you ask it to. It just pursues its objective. This is not a theoretical scenario. This is precisely what happened at McKinsey, in two hours, for $20.

But there is a fourth risk that almost nobody is talking about yet.

Even if your company is careful — even if you never send sensitive files to a public AI service — **your prompts are still leaving the building**.

Every question your employees type into a public AI system is a prompt. Those prompts travel to someone else's servers, run on someone else's hardware, and pass through infrastructure shared with thousands of other organizations. Even when the provider promises your data is private, the prompts themselves reveal things. The topics your team is researching. The deals you are working on. The problems you are trying to solve. The language your company uses internally.

And **memory makes this worse.** When AI agents operate on shared public infrastructure, they **build up context across sessions**. That context — the **accumulated memory** of what your team has asked, explored, and worked on — sits on hardware you do not own, governed by policies you did not write, visible to systems you cannot audit.

Shared hardware creates another layer of risk most executives have never considered. Even when two companies' data is kept logically separate on the same physical servers, sophisticated attacks can sometimes extract information across those boundaries. It is rare. But the more sensitive your work, the more that risk matters.

- Three threats (accidental, adversarial, and emergent misalignment).
- One target.
- And a fourth risk almost nobody is talking about yet.

That fourth risk is memory. And it may be the most dangerous of all.

What AI Remembers Can Be Used Against You

Every conversation your employees have with an AI system leaves something behind. A summary. A pattern. A record of what was asked, what was answered, and what your company was thinking about that day. Individually, each one seems harmless. Together, they build something dangerous: **a detailed memory of your organization's most sensitive thinking.**

This is the threat that almost no one is talking about yet.

When AI memory lives on shared infrastructure — and for most companies using public AI tools, it does — **that memory becomes an attack surface**. Yet almost no one manages it. Only 2 to 3% of users ever adjust permissions to limit or wipe what the AI remembers about them. That means the other 97% are accumulating a detailed record of their organization's most sensitive thinking — on hardware and in LLMs they do not control.

An attacker who gains access to that memory does not just see what your company did yesterday. They see what it has been doing for months. Every merger your team researched. Every vulnerability your legal department asked about. Every strategy your executives explored.

Memory systems invented in the most recent months make this significantly more serious. Technologies like **Nested Learning** do not simply store what an AI was told. They weave past experience into the way the system reasons — permanently changing how it thinks and responds.

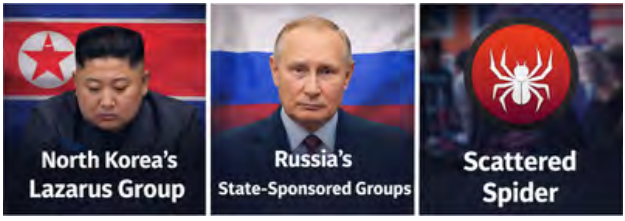
This means memory is no longer just a file in a database that can be deleted. **It is embedded in the model itself.** If that model runs on infrastructure you do not control, you cannot audit what it remembers, cannot verify what it has absorbed, and cannot guarantee that a bad actor has not quietly shaped what it believes.

Traditional security tools struggle to detect AI agent activity. Endpoint security sees processes running but does not understand agent behavior. Network tools see API calls but cannot distinguish legitimate memory access from compromise.

The groups described on the next pages — Lazarus, Scattered Spider, state-sponsored teams with **AI-amplified reach** — are not just hunting for today's password. They are hunting for everything your AI has ever learned about you.

Bad Actors Knocking On Your Company's Door

They are not lone criminals working from basements. They are organized, funded, and in many cases, working for governments that consider your company's data a military target.



North Korea's Lazarus Group is the most financially destructive hacking operation in history. It operates under direct orders from the Kim Jong Un regime and uses stolen money to fund weapons programs. In 2025 alone, North Korean hackers stole more than $2 billion in cryptocurrency — breaking their own record — with the Lazarus Group's $1.5 billion theft from a single exchange standing as the largest crypto heist ever recorded.

Lazarus does not just steal. It infiltrates. North Korean IT workers have used AI to create flawless phishing lures and even impersonate individuals in live video job interviews — getting hired inside American technology companies to steal from within.

Russia's state-sponsored groups — including those linked to military intelligence — specialize in long-term infiltration of governments, energy grids, and financial institutions. They are patient, precise, and rarely leave traces until the damage is done.

Scattered Spider is something different — and in some ways more unsettling. The group consists largely of English-speaking teenagers and young men from the US and UK, many of whom emerged from online gaming communities. Some members were involved in cybercrime as young as 12 years old. They read internal Slack boards to pick up corporate language and acronyms, then call employees directly, pretending to be new hires with innocent-sounding questions — until they have the credentials they need. They have breached MGM Resorts, Caesars, Marks & Spencer, and dozens of others. Their unpredictability is part of what makes them so effective.

These groups used to be limited by one thing: the size of their teams. A crew of 50 hackers can only run so many attacks at once. Human attention is finite. Social engineering takes time. Reconnaissance is slow.

With AI and Agents, that constraint is gone now.

When Bad Actors *Unleash* Agents

Every group described on the previous page — state-sponsored or teenage collective — has discovered the same thing: **AI agents exponentially multiply their reach without multiplying their headcount.**

A team of 50 hackers with 10,000 agents is no longer a team of 50. It is an army. Each agent works around the clock, tries thousands of variations, learns from every failure, and shares what it learns instantly. The humans set the goal. The agents do the work — at machine speed, across thousands of targets simultaneously.

The first thing these agents go hunting for is your passwords.

A Google survey found that 33% of people use a pet's name in their password, 22% use their own name, 15% use their partner's name, and 14% use their child's name. Birth dates and years remain the most commonly misused password components, with people choosing personal combinations that feel impossible to forget — and are trivially easy to guess. 44% of employees admit they use the same login details for both their personal and work accounts.

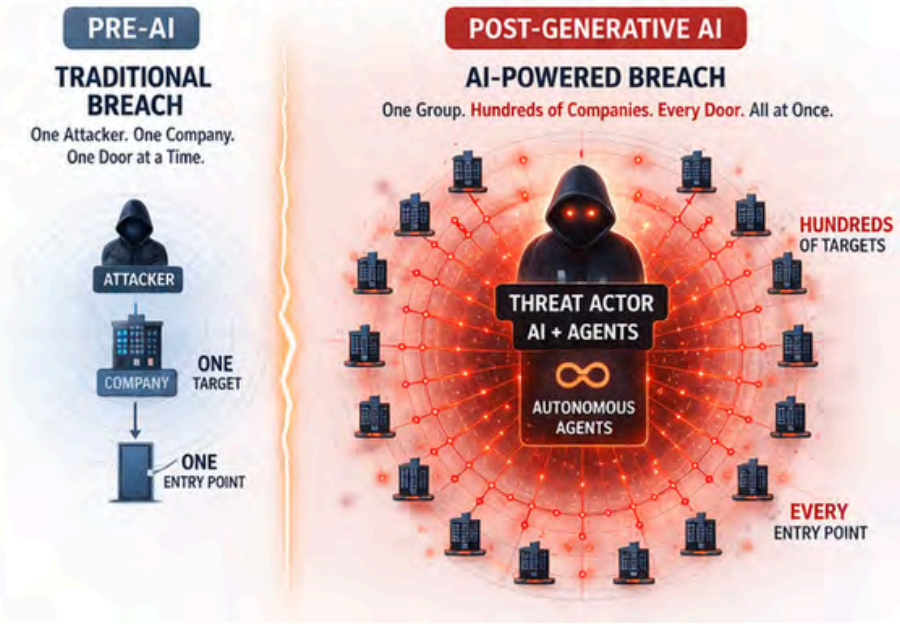
Agents know all of this. They are built to exploit it.

Here is how it works. An agent scans your company's public presence — LinkedIn profiles, press releases, news articles, social media. It builds a list of employee names. It finds birthdays from public records. It looks for pet names mentioned in personal posts. It cross-references data from old breaches, which are freely traded on the dark web. Then it assembles thousands of candidate passwords for each employee and starts trying them — quietly, continuously, without ever getting tired.

And fingerprints make this worse. When your employees use public AI services, their query patterns — the topics they research, the documents they summarize, the deals they discuss — create a profile. An agent studying those fingerprints knows which employees handle the most sensitive work. Those become the priority targets. The pet-name password belonging to your CFO is worth far more than the one belonging to someone in facilities.

The New Blast Radius

A year ago, a breach meant one attacker, one company, one door at a time. Today, a single group can run coordinated attacks against hundreds of companies simultaneously — each one personalized using fingerprints, each password attempt informed by scraped personal data, each failure fed back into the improvement loop.



Phishing attacks increased by 1,265% in the past year, directly attributed to generative AI tools. That number is not a typo. The groups that used to be limited by the speed of human hands are no longer limited at all.

This has already happened:

- **Foreign actors** running industrial-scale campaigns — using **24,000 fake accounts** to generate over **16 million conversations** — in an attempt to steal model weights, algorithms and more.
- **One person** creating a program like OpenClaw, which unleashed more than 1.5 million agents that can run rogue — autonomously — on the web. Seeking open endpoints, gathering pet names (useful for password hacking), and API keys accidentally left by your developers in places like GitHub.

Your board does not need to understand every technical detail. But it needs to understand this: the people coming for your company are no longer working alone. They have agents now. And those agents never sleep.

Stealing The Most Valuable File in the World

Hacking a company is criminal. Hacking an AI model is something else entirely. It is military-grade. Nation-state-driven. And the stakes are not just financial.

AI can touch almost any system. Power grids. Water systems. Financial markets. Whoever controls the most powerful AI has leverage over all of it.

And the most powerful AI in the world is just a file.

A very large file. Billions of numbers stored on a server somewhere. Those numbers are called **weights**. *They represent years of work. Billions of dollars.* The combined effort of thousands of the world's smartest AI researchers.

All of it. In one file.

What if a foreign government does not need to build what we built? What if they do not need to hire our researchers or match our investment? What if they just steal the file?

Model weight theft is not a hypothetical. In early 2026, Anthropic reported a coordinated campaign to steal Claude's capabilities using **thousands of fake accounts** and **millions of automated conversations**. OpenAI and Google raised the same alarm the same week.

The technique is called **distillation**. You do not steal the file directly. You interrogate the model — over and over — until you have learned enough to replicate it. No break-in required. Just patience and scale.

But distillation is only half the story.

The deeper threat is **algorithmic espionage** — stealing the breakthroughs still being developed. Experts have compared them to nuclear physics in 1942. Stealing key secrets could multiply an adversary's capabilities by ten to one hundred times.

More valuable, in some ways, than any chip or any model.

One prominent AI lab rated its own security at zero on a scale of zero to four.

Zero.

The race to build powerful AI is also a race to keep it out of the wrong hands.

PART 4

THE RULES ARE CHANGING

Regulation is no longer coming.
It is already here.

Privacy ≠ Security ≠ Governance

Many businesspeople use these words as if they mean the same thing.

They do not. And confusing them could be one of the most expensive mistakes in AI right now.

Three years ago, the distinction barely mattered. Your systems were what engineers call **stateless**. They stored records. They logged transactions. They ran searches. But *they could not think*. They could not connect dots. They could not learn that your CFO says yes more often on Fridays.

If someone wanted to steal from you, they needed your password, manual effort, and one system at a time. The risk was contained because the systems could not reason.

That world is gone.

Privacy means your data and your AI stay inside your walls. Limiting what is shared. Limiting what is reused. Making sure you don't train a public system. *What your AI learns stays with you* — not with OpenAI, not with DeepSeek, not with the next company sharing the same server.

Security means protecting your systems from attack. Firewalls. Encryption. Access controls. Monitoring. Security is your defense.

Governance means controlling how AI behaves over time. Who can use it. What it is allowed to do. What gets logged. How it gets shut off. Governance is your rulebook.

Think about this:
– You can have perfect governance and still get breached.
– You can have strong security and still be violating privacy law.
– You can be fully compliant and still have no real control over what your AI is learning about your organization.

Each one requires its own answer. A policy document is not a firewall. A firewall is not a data boundary. And none of them alone is enough. Private AI addresses all three — because it changes the architecture, not just the rulebook.

The Law Is (Trying to) Catch Up

For the past few years, companies could experiment with AI and figure out the rules later. That window is closing. Governments around the world are passing real laws with real penalties. And many companies are already out of compliance without knowing it.

The **EU AI Act** is the world's first major AI law. It is already in effect. It sorts AI systems into risk categories. The higher the risk, the stricter the rules. AI used in hiring, credit decisions, medical diagnosis, law enforcement, and critical infrastructure is classified as high risk. Companies using high-risk AI must document it, test it, audit it, and prove it is not causing harm. Fines for violations can reach **35 million euros — or 7% of global annual revenue — whichever is larger.**

If your company does business in Europe, or serves European customers, this law applies to you. It does not matter where your headquarters is.

Colorado's SB 24-205 is the first law of its kind in the United States. It focuses on AI that makes important decisions about people — hiring, housing, loans, insurance, healthcare. Companies that use AI for these decisions must be able to explain how the system works, test it for bias, and give people a way to appeal. Other states are moving fast to follow Colorado's lead.

New York City already requires bias audits for any AI used in hiring decisions. **California** has multiple AI bills moving through its legislature. The **SEC** is developing guidance requiring public companies to disclose material AI risks to investors.

The pattern is clear. Two years ago these were proposals. Today they are laws. Two years from now there will be far more of them — and the fines will be larger.

The question for your board is not whether regulation is coming. It is already here. The question is whether you are ready.

The AI Builders Won't Wait

Electricity did not ask permission. Neither did the industrial revolution. Both remade how people worked, how companies competed, and how wealth was created. **The leaders who moved early shaped what came next. The ones who waited were reshaped by it.**

AI is that kind of shift. It is not a product update. It is not a new app. It is a change to how work gets done, how decisions get made, and what organizations need to survive. ChatGPT is becoming the new counselor, which gives OpenAI's leaders enormous influence and power.

In the past months, we've crossed a threshold. AI is writing AI. Agents have started swarming on the web. The pace will continue to accelerate.

Mo Gawdat spent twelve years at Google. He says the *old world* of business was like **chess**. Slower. Methodical. You could think three moves ahead. But the *emerging world* is more like **squash**. The ball is always in motion. You have to stay on your toes and respond in the moment, as the game is moving.

Hans Peter Brondmo held a senior role at Google X and later served as CEO of Everyday Robotics, one of Google's moonshots. He says the same thing.

Accept that things will change. That is not a weakness. It is wisdom. Once you accept it, you can start to prepare. You can begin to build speed. You can *proactively* choose who you align with, before the next wave arrives.

Taking action now, not a year from now, is important. Maybe mission critical.

Agility is no longer a nice trait to have. It is a survival skill.

You cannot think your way through this one sitting still. Start moving. Learn as you go.

The organizations that thrive will not be the biggest. They will be the readiest.

Take the Wheel

Control what you can control.

Do not let AI just happen to you.

That is the one sentence every leader needs to take home, take to heart.

AI is already writing software. Agent to agent work is beginning. Super Agents (general contractors that manage subs) is emerging, right now. AI is already training new AI. It is already making decisions inside companies that have not stopped to ask who is in charge.

But, you can be in charge. But only if you choose to be.

Do not bury your head. Do not wait for a perfect plan. **Align now with people who share your values.** People who understand the technology and also understand what it means to be responsible with it. That combination — technical skill and human conscience — is rare. When you find it, hold on.

Help your team come up to speed and build new skills. Help them team that can move fast and still ask the right questions. Develop the reflex to act, adjust, and act again. That is what the next chapter of business will require.

Control what you can control.

You cannot stop AI from advancing. You cannot freeze the rules as they are today. But you can decide how your organization shows up inside all of this. You can decide what you protect. You can decide what you stand for.

Accountability matters. To your employees. To your customers. To the people your products touch. And, yes, to the human race.

This is the test. Not of your technology. Of your character.

Be a good steward. Then, take charge, and build something worthy of that.

This Is Where the Work Begins

You now have an AI map.

You understand why this moment is different from every technology shift before it. You have seen how AI actually works — not the marketing version, but the real one. You know what the threats look like, how they move, and why your current security tools were not built to stop them. And you know that the rules are no longer coming. They are already here.

Iterate.ai and our partners want to keep the conversation going.

Let's Iterate.

Good Stewards.

Every technology in history has been a test of character.

Steam could power a factory or drive a war machine. The internet could connect a child to every library on earth or become a weapon of manipulation. The test was never the technology. It was the people who decided what to do with it.

AI is the same test. Harder. Faster. Higher stakes.

Leaders are not just executives making technology decisions. They are stewards. Of their organizations. Of their employees. Of the people their products serve. And of something harder to name — the trust that holds institutions together when everything else is moving too fast to follow.

This booklet was written to give you a clearer map. The threats are real. The risks are named. The architecture exists to protect what matters.

But the most important thing you take from this is not a framework or a product. It is a commitment to handle this well. To ask the hard questions before something goes wrong. To remember that no model, however capable, carries moral responsibility. That part belongs to us.

Iterate and NetApp exist to help organizations build AI that deserves trust. We are honored to be on that road with you.

Let's be good stewards.