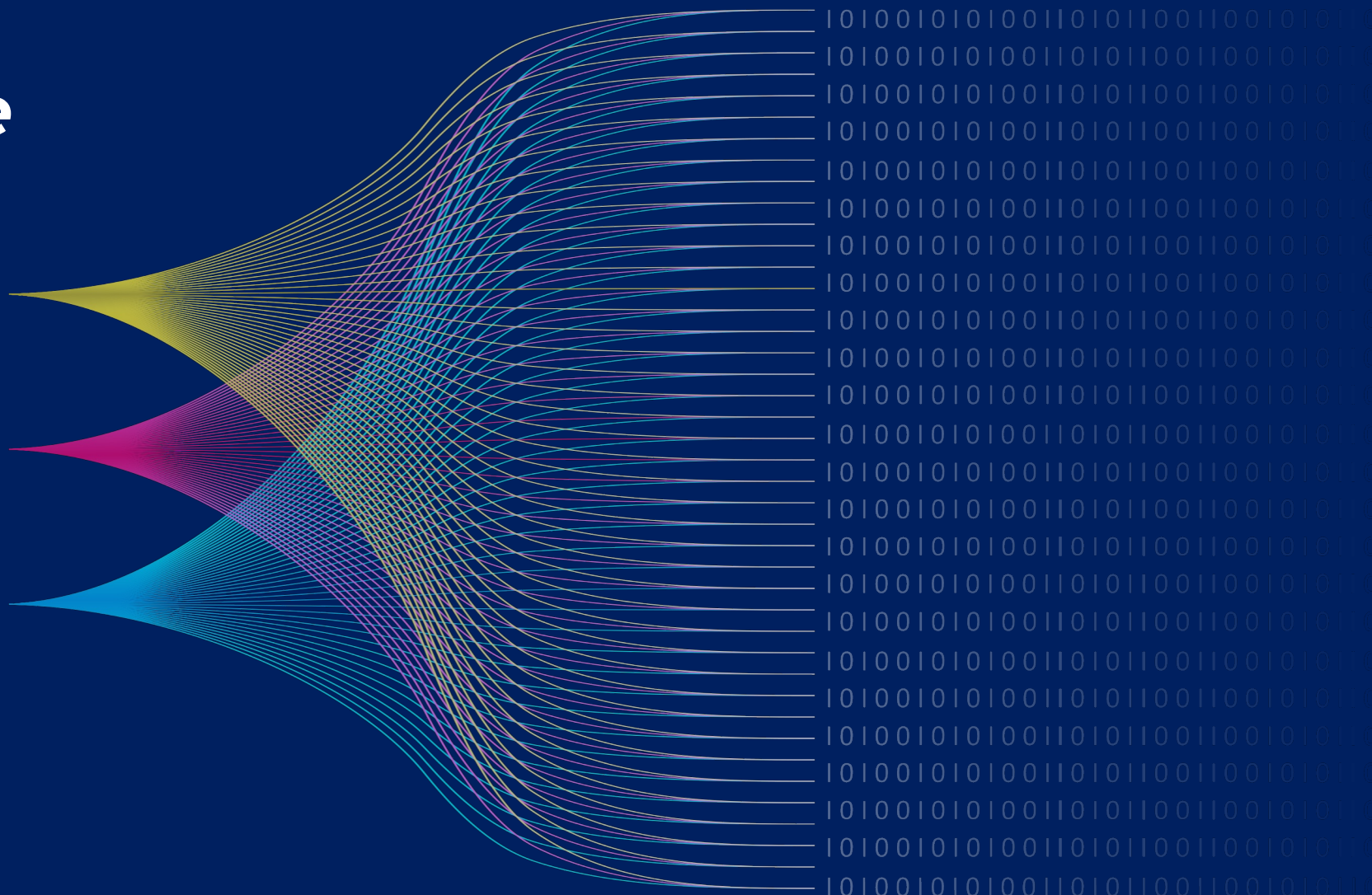
infintel

AI & Governance

With Great Power Comes Great Responsibility



Introduction

Governance Gap: When Innovation Outpaces Oversight

AI is advancing at a staggering pace, shaped by two parallel evolutions. First, the technology continues to expand its capabilities at an extraordinary rate, breaking new ground in areas ranging from complex decision-making to artistic creation. It's achieving unprecedented proficiency in reasoning, creativity, and problem-solving across domains that once seemed exclusive to human intelligence. Second, AI is becoming widely accessible as barriers to entry fall dramatically. Technologies previously confined to tech giants now reach developers of all backgrounds through open-source frameworks, cost-effective implementations, and optimized architectures.

However, while AI capabilities and accessibility race forward, crucial frameworks for responsible governance, ethical safeguards, and regulatory oversight are struggling to keep pace. Our recent survey of 221 organizations underscores this gap: while 72% have already adopted AI in their operations—and another 25% plan to do so within the next 12 months—only 50% have established governance frameworks. This stark gap between adoption and oversight highlights the urgent need for more balanced and deliberate AI implementation.

This report is intended to help organizations close the widening gap between AI innovation and governance. By unpacking the risks, their real-world consequences, and practical strategies for oversight, we aim to empower decision-makers to develop, deploy, and manage AI responsibly—aligning innovation with accountability.

Inside the Report: Main Sections at a Glance

1. **Future Ready Governance Framework:** Creating a resilient AI governance framework requires the right combination of people, tools, and protocols—designed not only to manage current risks but also to anticipate and adapt to emerging challenges. A robust framework provides a structured, yet flexible, approach to defining roles, responsibilities, processes, and policies. It aligns AI development with the organization’s goals and values, fosters accountability, transparency, and ensures regulatory compliance. Most importantly, it equips organizations to adapt to the fast-changing AI landscape.

2. **Risks at Play:** Understanding AI risks—spanning ethical concerns, security vulnerabilities, and regulatory non-compliance—is critical to responsible AI adoption. These risks are often interconnected and addressing one can influence others. A systematic approach to identifying and prioritizing them by magnitude and enables organizations to direct mitigation efforts with greater precision, ensuring fairness, security, and compliance while fostering trust in AI systems.

3. **Regulatory Landscape:** This section explores the significance of a unified global approach to AI regulation emphasizing its potential as an enabler while highlighting the current divergence in regulatory efforts across different jurisdictions. It examines the EU’s comprehensive regulatory progress, Canada’s legislative gridlock, and the U.S.’s more hands-off stance. Understanding these diverging paths is essential for organizations looking to build governance frameworks that are globally aware, adaptive, and compliant.

"The danger is that if we invest too much in developing AI and too little in developing human consciousness, the very sophisticated artificial intelligence of computers might only serve to empower the natural stupidity of humans. We are unlikely to face a robot rebellion in the coming decades, but we might have to deal with hordes of bots who know how to press our emotional buttons better than our mother and use this uncanny ability to try and sell us something—be it a car, a politician, or an entire ideology."

— **Yuval Noah Harari**,
"21 Lessons for the 21st Century"

Contents

+1,265%

YoY increase in **malicious phishing** emails since Q4 2022 (*SlashNext, 2023)*

>60%

S&P 500 companies believe they have **material AI risks** (*Deloitte, 2024)*

87%

Organizations reported experiencing **AI driven cyberattacks** in 2024, based on a survey of 500 security professionals from 9 countries *(SoSafe Cybercrime Trends 2025)*

78%

CISOs admit that **AI-powered cyber threats** are no longer a future risk—they're a present reality, based on a survey of over 1,500 cybersecurity professionals around the world *(DarkTrace, 2025)*

Executive Summary

Governance Beyond Humans

Governance has traditionally focused on overseeing human decision-making—ensuring that decisions are ethical, transparent, accountable, and compliant. But with the rise of AI, we're entering a new era where machines can now make decisions that have real-world consequences. Unlike older technologies that follow fixed, pre-programmed rules, modern AI systems can learn from data, adapt to new situations, and make high-stake decisions with a high degree of autonomy. This shift is changing how we deploy machines—and how much trust and responsibility we place in them.

This growing reliance on machine-generated decisions raises concerns around bias, accuracy, safety, and compliance—outcomes shaped by the quality of data and algorithms that power these systems. As machines gain more autonomy and agency, governance must evolve to ensure that AI-driven decisions are managed ethically, responsibly, transparently, and in accordance with regulatory expectations.

Future: Greater Agency and Widespread Adoption?

Agentic AI: From Digital Assistant to Digital Co-worker

On the horizon of AI innovation are agentic systems—machines that don't just respond, but act. These emerging systems are shifting from reactive tools— similar to language models that answer human queries—to proactive agents, such as self-driving cars, that can pursue predefined goals, make decisions, and collaborate with other AI agents with minimal human guidance. This marks a fundamental shift from today's digital assistants to tomorrow's digital co-workers—autonomous helpers that can anticipate needs, initiate tasks, and operate independently. While still in development, this evolution signals where AI is headed in the coming years.

Democratization: AI in Everyone's Hands

AI is also becoming more accessible than ever—no longer the exclusive domain of large tech companies. Breaking down technical and financial barriers and empowering a diverse range of users—non-technical individuals and startups. This broad-based adoption is being driven by three converging forces:

1. **Open-source:** Open-source AI models unlock access to powerful capabilities and accelerate innovation beyond traditional gatekeepers. It accelerates innovation by granting unrestricted access to foundational models, fostering experimentation, and empowering developers to customize and push the limits of AI capabilities in ways that closed systems cannot.

2. **DeepSeek revolution:** DeepSeek's efficient models are transforming AI by significantly reducing computing power requirements and cutting costs. Through smarter training strategies these models achieve strong performance with less data and energy. Their innovative architecture makes them not only powerful but also cost-effective, bringing high-performance AI within reach for a wider audience.

3. **Low code/no code:** Low-code and no-code platforms are transforming AI adoption by using intuitive visual interfaces, drag-and-drop tools, and pre-built components—eliminating the need for deep technical expertise and making advanced AI tools accessible to a much wider audience.

Artificial General Intelligence (AGI)

Today's AI, often referred to as narrow AI, excels in specific, well-defined tasks but lacks the versatility and adaptability of human cognition. In contrast, Artificial General Intelligence (AGI) aspires to mirror the breadth of human intelligence—systems capable of learning from experience, reasoning across domains, and applying knowledge without being limited to task-specific training. Much like a human who can drive a car, learn a language, play a sport, or practice medicine without needing to be 'rewired' for each activity, AGI envisions machines that can generalize, self-learn, and evolve continuously. Although still largely theoretical, AGI represents the long-term frontier of AI development—and a profound challenge for future governance frameworks.

Building a Future-Ready AI Governance Framework

As AI rapidly transforms industries, organizations increasingly face "AI FOMO"—the fear of falling behind as competitors and peers adopt artificial intelligence—often leading to hasty implementation decisions. A strong governance framework lays the foundation for thoughtful adoption, ensuring AI is implemented through a well-evaluated process. It enables organizations to harness AI for long-term competitive advantage while systematically addressing potential risks.

At its foundation, strong governance establishes guiding principles and policies, supports proactive risk management and defines accountability. It also keeps organizations attuned to a fast-evolving regulatory landscape, enabling a structured approach to transparency and compliance.

A robust governance framework hinges on the right combination of three elements: people, protocols, and tools. This mix must be carefully calibrated to the organization's scale, AI maturity, and strategic objectives – whether implementing foundational safeguards for a small business or enterprise-grade controls for a tech leader. The approach should both address current operational realities and anticipate where AI is heading.

Understanding the Risks

AI adoption comes with a diverse and evolving set of risks—making risk awareness a critical foundation of any governance framework. AI-related risks can be categorized across three key dimensions.

- **Unintended Consequences** – Harmful outcomes from well-intentioned AI due to biases, inaccuracy, misuse, and lack of transparency
- **Weaponization of AI** – Harmful or malicious use, whether by insiders or external actors such as—deepfake scams, AI-powered cyberattacks etc.
- **Loss of Human Control** – Systems evolving or acting in ways not foreseen or controlled by humans

While the 'loss of human control' is not yet a reality, the other two risk categories are actively materializing in real time, driven by both direct and indirect sources.

Direct AI risks stem from intentional organizational choices about how AI systems are designed, deployed, and scaled. When organizations leverage AI for high-stakes functions, they face substantially greater risks than when using it for routine or low-impact tasks. This difference in risk profiles means AI design, deployment, and oversight must be carefully customized to match each specific intended purpose. Organizations need to implement proportionate governance based on the potential impact of each AI application.

Indirect risks, by contrast, arise from the use of AI beyond an organization's direct oversight. This can occur when third-party service providers or employees utilize AI tools "bring your own AI (BYOAI)" —sometimes unknowingly— in ways that introduce unintended consequences. For instance, risks may surface when employees rely on unauthorized AI applications or when external service providers embed AI into systems without sufficient transparency or governance.

The level of risk exposure depends on an organization's specific use cases and its AI adoption strategy.

Broadly, there are three main approaches for AI adoption: **building proprietary AI, modifying existing systems, or using off-the-shelf tools.** The chosen strategy should align with the sensitivity and impact of the use case.

For example, high-stakes applications—such as defence or advanced healthcare diagnostics like cancer detection—warrant a "build-your-own" approach, supported by secure infrastructure, such as on-premise servers and high-security environments, along with a need for a high degree of precision. In contrast, lower-risk applications—like customer service chatbots—can be effectively managed using pre-built, commercially available solutions.

Effective risk management is about navigating a diverse array of risks while maintaining the equilibrium between control and innovation. The goal is to enable organizations to pursue their strategic objectives without compromising agility or stifling innovation.

Regulatory Landscape: AI Governance at a Crossroads: Global Divergence, Legal Gaps, and the Road Ahead

The rapid advancement of AI brings with it a host of novel risks and complexities—from the spread of misinformation and deepfakes to profound ethical challenges. Yet, regulatory frameworks have struggled to keep pace. Laws remain fragmented across jurisdictions and are largely reactive, leaving critical gaps that enable AI-driven risks to grow unchecked.

Traditional legal structures—such as those governing anti-discrimination, intellectual property, and consumer protection—continue to apply but were not designed for the scale, autonomy, or generative capabilities of modern AI. For example, deepfakes—highly realistic AI-generated audio, video, or images—operate in a regulatory gray area. Existing laws often impose only minimal transparency requirements, with little clarity around intent, accountability, or enforcement.

As AI continues to evolve—shifting toward more agentic capabilities—new challenges are emerging, with the increasing collaboration between AI systems across different regions. For these systems to interact seamlessly, it is crucial to establish common global regulations and principles. Without such frameworks, AI systems may struggle to collaborate beyond geographical boundaries. Therefore, unified regulations are not just advantageous —they are a prerequisite for ensuring the smooth interoperability of AI systems worldwide.

Although there is a significant need for a unified global regulatory framework, the current AI regulatory landscape is sharply diverging.

The EU has taken the lead with the AI Act, the world's most comprehensive legislation. It establishes a risk-based framework with robust enforcement, while also launching new initiatives—like the AI Continent Action Plan—to support innovation and reduce compliance burdens.

The U.S., by contrast, has pivoted decisively away from regulation. The revocation of key executive orders signals a preference for speed over safety, as Big Tech influence grows, and AI is increasingly seen as a tool of geopolitical competition.

Canada, once a regulatory pioneer, now finds itself in limbo. The proposed Artificial Intelligence and Data Act (AIDA) has stalled under political pressure, vague definitions, and intense lobbying from U.S. tech firms. Caught between EU-style oversight and U.S.-style deregulation, Canada's approach remains uncertain.

In the absence of a universal regulatory framework, voluntary global standards (e.g., ISO/IEC JTC 1/SC 42, IEEE's Ethically Aligned Design) and emerging case law are trying to fill the gap. But they cannot replace the need for coherent, agile, and enforceable regulation. The longer the delay, the harder it becomes to build trust, safeguard rights, and ensure AI serves the public good.

The question is no longer whether we need AI governance—it's whether global actors can converge on a forward-looking framework before innovation fully outpaces oversight.

AI Governance

Designing and Building Guardrails

Why guardrails?

What makes AI unique?

AI is driving a paradigm shift in human-machine dynamics, with machines becoming capable of making high-impact, real-time decisions that shape the world around us. AI's ability to learn autonomously and adapt sets it apart from earlier rule-based technologies, which could only automate predefined tasks. Today's AI systems can generate intelligent, context-aware responses in dynamic environments across a range of domains.

As this technology evolves, it is crucial to understand its capabilities, transformative potential, and the challenges it may pose.

Capabilities of AI

Learning and Adapting – Machine learning, particularly deep learning, uses artificial neural networks to identify complex patterns and correlations, giving AI the ability to learn, adapt, and generalize its training to apply in new situations. Unlike traditional, static technologies that rely on fixed rules and cannot adjust beyond their programming, deep learning systems dynamically respond to new scenarios by analyzing data and continuously improving.

Cognitive capabilities. – AI can mimic human cognitive abilities, such as vision, speech, reasoning, and problem-solving. AI has achieved human-like fluency, context awareness, and adaptability in generating and understanding language dynamically.

Interdisciplinary Nature – AI has an interdisciplinary foundation, drawing from diverse fields such as computer science, neuroscience, linguistics, and mathematics to create comprehensive solutions for complex challenges.

Scalability – AI's scalability enables it to analyze massive datasets and perform high-speed computations, leveraging cloud-based GPUs, parallel processing, and big-data frameworks to tackle problems that were previously unimaginable.

The power of AI

AI's capabilities enable it to make high-impact, real-time decisions, even in high-stakes areas such as healthcare, defense, and finance. It continuously refines its responses and adapts to changing conditions, while working towards predefined goals.

AI's decisions can sometimes surprise even its developers, as advanced models—especially in deep learning—can uncover patterns and insights beyond human intuition, reflecting its ability to deliver unpredictable and often impressive outcomes.

AI decision making

AI decision-making raises several concerns, including alignment with human values, accuracy, safety, and fairness.

Alignment with Human Values – AI may prioritize efficiency and profitability over ethical considerations as it lacks moral reasoning and empathy. If not carefully designed, AI systems might optimize for financial gains, disregarding social or environmental responsibilities.

Accuracy – The reliability of AI decisions depends on data quality, model architecture, and training methods. Poor data, biases in training sets, or insufficient updates can lead to inaccurate predictions or flawed recommendations. Even well-trained models can struggle with unfamiliar or incomplete data, reducing trust in AI-driven decisions.

Safety – AI systems can produce harmful or unintended consequences, especially when encountering unfamiliar scenarios. They can be vulnerable to adversarial attacks, where malicious inputs manipulate outcomes. Without proper safeguards, AI could make decisions that are dangerous or counterproductive in critical areas such as healthcare, finance, and security.

Fairness – AI learns from historical data, which may contain societal biases. If not addressed, AI can reinforce or even amplify existing discrimination, leading to unfair hiring practices, biased credit scoring, or unequal treatment in law enforcement. Ensuring fairness requires careful dataset selection, bias detection, and ethical AI governance.

Why Guardrails? Uniqueness of AI from any other technology

Capabilities of AI

Learn & Adaptability

Ability to autonomously learn complex patterns from data and generalize them to novel situations—unlike rule-based systems confined to predefined logic

Cognitive Capabilities

AI mimics human cognition in areas like vision, speech, reasoning, and problem solving

Interdisciplinary Nature

Draws knowledge from diverse fields to produce holistic solutions to multifaceted problems

Scalability

With big-data access and super computing power, AI can tackle problems at scales previously unimaginable

Power of AI

Autonomous & Adaptive Decisions

It is not just about automating tasks like previous rule-based technologies— it can make autonomous and adaptive decisions within its trained scope, without requiring explicit programming

Speed & Scale of Decisions

Real-time, high-impact decisions across industries

Unpredictability of the Decisions

Makes decisions that even developers can't fully explain, as inner workings of AI rely on probabilities and patterns rather than fixed rules

Nature of AI Decisions?

Alignment with Human Values

AI may prioritize efficiency and financial gains over ethical responsibility

Accuracy

The accuracy of AI output depends on factors like data quality, model architecture, training methods, feature selection, continuous updates, and computational power. Bias, poor data, or inadequate training can reduce reliability

Safety

AI can produce unreliable outcomes when handling unfamiliar data. Additionally, adversarial attacks and biased inputs may manipulate AI systems, leading to unintended or harmful consequences

Fairness

AI systems can reinforce or amplify biases present in training data, potentially leading to unfair or discriminatory outcomes in decision-making

What does the future hold?

What is in the horizon?

Agentic AI | Unlocking human-machine synergies

In 2025, AI is poised for a paradigm shift with the emergence of Agentic AI—a transformative evolution beyond today's 'reactive and assistive' models. Unlike traditional, existing AI, which passively responds to user prompts and provides insights for human decision-making, Agentic AI represents a leap towards autonomy, proactively executing tasks based on predefined objectives.

This advanced form of AI will not just suggest or assist; it will take the initiative, acting decisively based on clearly defined goals. By making independent, goal-driven decisions, Agentic AI is expected to redefine industries by enabling a "virtual workforce" that augments human capabilities (e.g. AI based Personal Assistants, AI based Project Managers). As the technology matures, it would be able to address higher-order challenges like simulating product launches or orchestrating complex multi-agent systems. Its interoperability with other technologies such as APIs, IoT devices, and robotics will enable it to execute complex decisions and collaborate across systems.

As a new frontier, Agentic AI is set to redefine productivity and innovation, marking a pivotal evolution in how machines interact with the world around them.

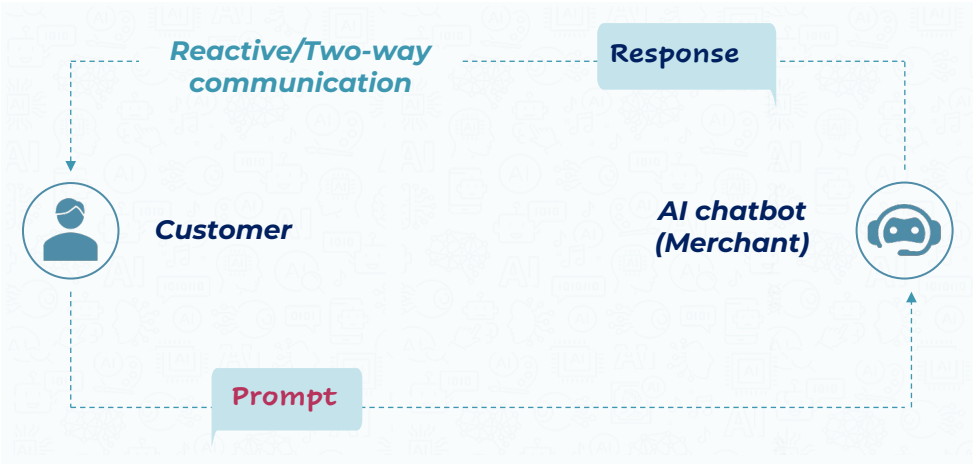
Challenges:

- While agentic AI offers autonomy, it raises concerns about reliability and ethics, primarily due to the lack of human oversight, which can lead to errors, and the questionable transparency and accountability of AI decision-making
- Although promising, agentic AI is still in its early stages. Current implementations often require human oversight and are limited to narrowly defined tasks

Shifting paradigms: The rise of agentic AI over existing AI models

Feature	Traditional AI (Chatbots, LLMs, Automation)	Agentic AI (Goal driven)
Response	• User ask and AI responds | Reactive	• AI takes initiative without prompting to pursue a defined goal | Proactive
Decision making	• Provides suggestions. Humans remain in control of decisions: human-centric	• Makes independent decisions based on set goals
Scope of action	• Provides only insights/instructions and relies on humas for execution	• Can complete multi-step tasks autonomously. Detects issues take corrective actions
Nature of engagement	• Digital assistant	• Digital co-worker/employee
Interoperability	• Work in isolation	• Work together in teams • Example: An AI financial analyst, an AI lawyer, and an AI HR specialist could collaborate to automate business operations.

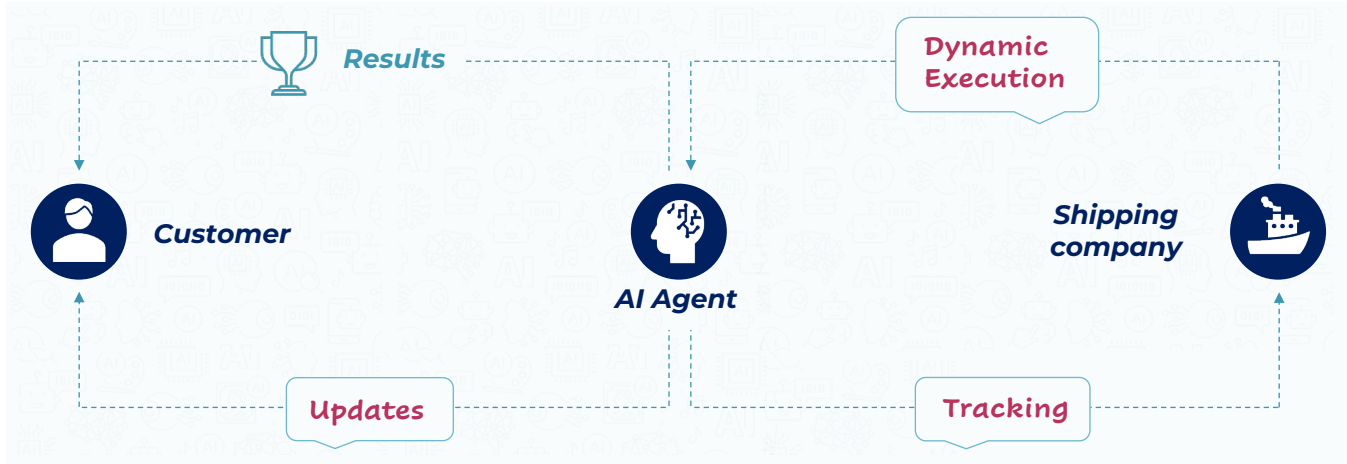
AI Assist



AI Assist Example:

- *In the present era of "AI assist," when you place an order on Amazon, the AI-driven chatbot responds to inquiries such as, "Where is my package?" or "What is the estimated delivery date?".*
- *It reacts to your prompts, offering information derived from pre-programmed rules or data analysis.*
- *Yet, it remains passive in nature, offering assistance solely by retrieving and presenting information, without taking any independent action of its own.*

AI Agent



Agentic AI Example:

In the era of Agentic AI, systems go beyond answering queries—they actively monitor and manage your orders for a smooth, hands-off experience. They provide real-time updates without prompts and quickly address delays by identifying issues and taking corrective action. Even before problems arise, they anticipate disruptions and offer proactive solutions to keep things on track.

WAYMO

Example: Waymo's autonomous taxis

Waymo's autonomous taxis use Agentic AI to manage every step—from booking to drop-off

- *Ride request & dispatch – AI processes bookings, calculates optimal routes, and dispatches the nearest vehicle*
- *Pickup & boarding – AI navigates to the location, authenticates passengers, and manages trunk access*
- *Autonomous driving – AI controls navigation, predicts obstacles, and ensures safety with redundant systems*
- *Dynamic adaptation – AI optimizes routes, handles emergencies, and connects to remote support if needed*
- *Drop-off & compliance – AI confirms arrival, monitors rule adherence, and finalizes the ride experience*

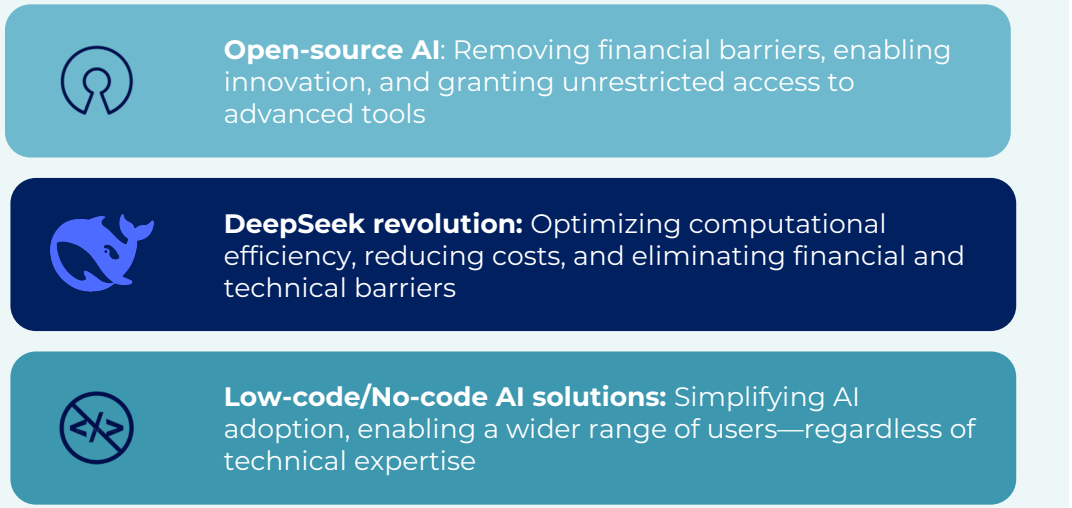
AI democratization

In 2025, AI is becoming more accessible than ever, breaking down technical and financial barriers and empowering a diverse range of users—including non-technical individuals, small businesses, and underrepresented communities. The rise of low-code and no-code AI platforms enables seamless interaction through intuitive interfaces, such as natural language inputs and drag-and-drop tools, eliminating the need for specialized expertise.

Open-source AI is democratizing cutting-edge intelligence, breaking down barriers to accessibility, affordability, and adaptability. DeepSeek exemplifies this shift, redefining cost efficiency through architectural optimization—achieving peak performance with fewer chips and low-power GPUs. By streamlining computational demands without compromising capability, DeepSeek makes advanced AI not just powerful, but radically economical, setting a new benchmark for scalable, low-cost intelligence.

What drives AI democratization?

AI without borders: Unveiling the forces behind its democratization



Open-source AI

- Open-source AI is a key enabler of AI democratization, making cutting-edge AI technology more accessible, affordable, and adaptable for a broader audience.
- It eliminates the high costs associated with proprietary software, enabling individuals, startups, and small businesses to access advanced tools without significant financial investment.
- Open-source AI fuels a relentless wave of innovation by granting creators unrestricted access to foundational models and tools. This openness not only accelerates experimentation but also empowers developers to customize, refine, and push boundaries in ways proprietary systems cannot

DeepSeek revolution

- DeepSeek's models, such as DeepSeek-V3 and DeepSeek Coder, boast 236 billion parameters, outperforming leading AI models like GPT-4 and LLaMA3-70B in tasks like coding, reasoning, and mathematics
- It is considered revolutionary for its low GPU usage due to its efficient architecture and innovative techniques. By optimizing computational resources, it trims both CAPEX and OPEX, making high-performance AI more accessible and cost-effective *E.g. The input cost for running DeepSeek R1 is $0.55 per million tokens, compared to $15 per million tokens for ChatGPT.*
- By eliminating financial and technical hurdles, DeepSeek empowers startups, small businesses, and underrepresented communities to harness AI's potential.

Low-code/No-code AI solutions

- Low-code/no-code platforms are revolutionizing AI adoption by eliminating the need for deep technical expertise, making advanced AI tools accessible to a broader audience.
- According to PRNewswire, nearly 60% of custom apps are being built outside the IT departments and 82% of businesses anticipating a rising reliance on non-IT employees in the next two years.

Beyond 2025

Artificial General Intelligence

Artificial General Intelligence (AGI), still in the theoretical stage, refers to AI systems capable of replicating human-like cognitive abilities. This means AGI would be able to reason, solve problems, learn, and adapt to novel situations across a wide range of tasks, much like a human. In some conceptualizations, AGI could even exhibit emotional intelligence, such as empathy, making its behavior nearly indistinguishable from that of a human.

Despite the impressive learning and adaptation capabilities demonstrated by Narrow AI compared to previous rule-based technologies, its functionality remains limited to specific domains and the datasets it is trained on. It can recognize patterns, make predictions, and even handle some novel situations within its trained scope.

However, AGI aims to surpass these limitations—it would have the ability to learn, reason, and adapt across multiple subjects, just like a human. It would be capable of self-learning from any type of data, applying knowledge across different fields, and solving completely new problems without requiring explicit retraining.

Example: A Narrow AI trained for chess is highly specialized in analyzing board positions, predicting moves, and optimizing strategies. However, if the same AI were asked to play poker, it would fail without extensive retraining since poker involves bluffing, probability calculations, and reading opponents—skills outside its pre-programmed domain.

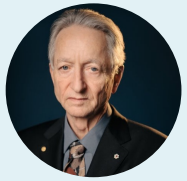
In contrast, a human chess player might not be an expert at poker, but they can leverage their existing knowledge of strategy, probability, and psychological tactics to learn the game relatively quickly.

AGI, if realized, would function similarly to a human—transferring skills across different domains without needing to start from scratch. Just as a skilled strategist can adapt their thinking from chess to poker, AGI would seamlessly apply knowledge from one area to another, making it vastly more versatile than today's specialized AI models.

Narrow AI vs AGI: From Taskmasters to 'Human-like Intelligence'

Feature	Narrow AI	AGI
Scope	• AI designed for specific tasks and cannot generalize beyond its trained scope.	• Ability to self-learn and train itself from experience alone and generalize and apply knowledge across different fields without prior data or retraining
Learning	• Learns only from specific datasets it was trained on.	• Can learn dynamically from any new information without predefined training.
Adaptation	• Adapts only within its trained function; struggles with entirely new scenarios.	• Adapts to new, unknown domains/situations without requiring retraining.
Emotional intelligence	• Lacks true emotional intelligence; does not understand emotions but can recognize patterns in sentiment analysis.	• Possesses emotional intelligence; can interpret, understand, and respond appropriately to human emotions.
Common sense knowledge	• Lacks real-world awareness; does not understand implicit meanings.	• Understands context, human emotions, and implicit meanings.
Human-like reasoning	• Lacks reasoning ability; only follows patterns in data.	• Can reason, make judgments, and understand abstract concepts.

AI unleashed or AI on a leash? What do the creators believe?



“regulation of the technology is "a start," but it needs to "have teeth.“
– Geoffrey Hinton (Godfather of AI)

"I think if this technology goes wrong, it can go quite wrong,“
– Sam Altman (OpenAI)

“with the technology now at an inflection point... I still believe AI is too important not to regulate, and too important not to regulate well.”
– Sundar Pichai (Google)

“There’s a huge amount of responsibility on the people leading this technology, that’s why we are advocating for the need of new institutions to be built to help govern some of this.”
– Demis Hassabis (DeepMind)

“I mean, the consequences are so vast and potentially so huge, both positive and negative, that this is a moment when we need strong governments and proactive intervention.”
– Mustefa Suleyman (Microsoft)

“If there’s nothing watching over them, or there’s nothing watching over us as an industry, there’s no guarantee that we’ll do the right thing and the stakes are very high,”
– Dario Amodei (Anthropic)

AI Governance Framework

Extending Governance Beyond Human Minds to Artificial Minds

For the first time, decisions shaping our world are made not just by humans, but by a new form of intelligence. **Governance must evolve from overseeing human decisions to also governing intelligent machines**—now functioning as active agents in high-stakes decision-making across areas like healthcare, finance, defense, and law enforcement.

AI governance framework

Strong governance is no accident – it is a combination of right people, right tools, and right protocols

A robust AI governance framework provides a structured approach to defining roles, responsibilities, processes, and policies, enabling organizations to navigate the complexities of AI technologies while fostering accountability, transparency, and compliance.

Align with organization's goals and values: An AI governance framework establishes the foundation for the ethical and strategic use of AI, ensuring its development, deployment, and application align with organizational goals, values, and regulatory requirements.

Structured process: Further, a well-defined AI governance taxonomy categorizes key principles, risks, and controls, enabling a structured approach to oversight, accountability, and compliance.

Adaptive: AI is constantly evolving at an unprecedented pace, with many unknowns about its full capabilities. Therefore, an effective governance framework must be agile—adaptive, forward-looking, and built for continuous change.

Promote sustainability: It should also prioritize sustainability by actively minimizing the carbon footprint of AI systems and reducing their broader social and environmental impacts.

Understanding the organization's AI strategy

An effective AI governance framework starts with a deep understanding of the organization's AI strategy, objectives, and intended impact.

Why is the organization adopting AI?

How can competitive positioning be enhanced through AI, and what long-term goals should be prioritized?

What is the scope of implementation?

Is AI integrated across the organization, limited to a specific department, or focused on a particular function or process?

What are the ROI and risks?

What are the potential financial returns from AI investments, and what challenges or downsides could arise from implementing AI?

Defining these factors upfront is crucial to identifying the most effective AI governance framework

Fear of Missing Out on Artificial Intelligence (AI FOMO)

In today's AI-fueled landscape, AI FOMO (Fear of Missing Out) has become an undeniable force, anxiously nudging businesses and individuals toward rapid AI adoption. The pressure to "keep up" is real, driven by game-changing innovations, competitive urgency, and the promise of exponential productivity gains. But in the rush to embrace AI, are we steering toward progress or chaos?

The AI gold rush, but at what cost?

From enterprises hastily deploying generative AI to professionals scrambling to upskill, AI FOMO has sparked a frenzy. Yet, beneath the allure of automation and efficiency lie pitfalls—rushed implementation, skill gaps, overinflated expectations, and even ethical dilemmas. The fear of lagging behind often overrides strategic foresight, leading to disjointed AI investments that fail to deliver tangible value.

*Escape **AI FOMO**. Adopt AI to gain strategic advantage — not just to follow the herd*

Governance: The compass in the AI storm

AI governance isn't a bottleneck, it's a blueprint for responsible innovation. It ensures AI adoption aligns with long-term goals rather than reactionary hype. Effective governance mitigates risks like algorithmic bias, regulatory non-compliance, and ethical oversights. By embedding AI within risk management frameworks and adopting global standards (e.g. OECD principles, EU AI Act), organizations can move from FOMO-fueled chaos to structured AI mastery.

Smart AI adoption – The antidote to FOMO

Strategic approach	AI should be adopted to augment the competitive advantage of the business
Risk mapping and management	Balancing interconnected AI risks to optimize safety, accuracy and performance
Optimal resource allocation	Ensure AI investments align with the company's financial risk tolerance while delivering measurable ROI
Ethical leadership	Ensure AI deployment maintains human oversight and aligns with human values

Establishing AI governance principles

In building an effective AI governance framework, once a clear AI strategy and set of objectives are established, defining core AI principles becomes essential. These principles—shaped by the organization's values, goals, and culture—provide clear guidance to ensure AI is used ethically, responsibly, and in line with business priorities..

By establishing these principles early in the governance process, organizations can set clear expectations for teams, foster cross-functional alignment, and mitigate risks proactively. More importantly, they ensure that AI systems not only deliver on performance and innovation but also uphold trust, compliance, and integrity—key elements for sustainable AI adoption and long-term success.

Microsoft's six core AI principles serve as a classic example of how leading tech companies are embracing AI governance principles—these are detailed in the section starting from page 30.

Guiding principles for AI governance

Human Centric: Human oversight is essential in AI systems. Humans should have the final control over AI outcomes

Compliance: Adhering to legal, regulatory, and ethical standards to ensure AI systems operate within established boundaries.

Safety: Designing AI systems to operate reliably, securely, and without causing harm to individuals or society

Transparency: Ensuring AI systems operate in a clear and understandable manner, with decisions and processes that can be explained to stakeholders

Fairness: Ensuring AI systems treat all individuals equitably, avoiding bias and discrimination in outcomes.

Accountability: Establishing clear ownership and responsibility for AI outcomes, ensuring humans remain answerable for AI-driven decisions.

AI risk management:

Identify, assess and defend

Risk management is a central pillar of AI governance, playing a vital role in ensuring that AI systems function safely, ethically, and in accordance with both legal standards and organizational policies. By proactively identifying, assessing, and mitigating potential risks, AI governance frameworks help safeguard against unintended consequences, promote responsible AI use, and ensure long-term trust and compliance.
This phase involves three key steps:

Identify

AI-related risks can arise throughout the AI lifecycle – data collection, model development, deployment, and ongoing use.
Risk identification involves systematically cataloging all AI applications in use, categorizing potential risks (e.g., ethical, operational, or regulatory), and pinpointing the sources of these risks to enable proactive mitigation and informed decision-making

- **Map your AI landscape** - Build a detailed inventory of all AI systems your organization relies on or is building.
- **Categorize different types of risks** - Analyze the risks your AI systems might introduce. E.g. ethical risk, security risk, compliance risk etc

Assess

Once risks are identified, organizations must assess their potential impact and probability of occurrence.. This phase ensures that resources are allocated effectively to address the most critical risks.

Prioritize Risks: Rank risks based on their potential impact and likelihood. Focus on high-priority risks that could cause significant harm to the organization

Defend

Mitigation strategies must be embedded in AI design, development, and deployment. This includes fairness checks, security safeguards, and human oversight, ensuring continuous monitoring and adaptive risk management for reliability and compliance.

Bias audits, explainability tools (SHAP, LIME), data privacy controls, and security enhancements help ensure fairness, transparency, and resilience, protecting AI from bias, adversarial attacks, and privacy risks through rigorous governance measures

Identify and categorize risks by mapping all AI applications and their potential risk sources

Develop a risk matrix based on the likelihood and potential impact

Have mitigation plans for each identified risks especially those with the highest score on the risk matrix

Human oversight and accountability

Effective AI governance relies on meaningful human oversight—commonly referred to as a "human-in-the-loop" approach—supported by clear accountability structures and processes for continuous improvement. This means humans remain actively involved in key decision points of AI operations, helping to ensure outputs are accurate, ethically sound, and aligned with broader organizational goals. While AI systems can operate autonomously in certain areas, human judgment is critical when it comes to interpreting results, making final decisions, or addressing edge cases that require contextual understanding.

The level of human oversight, however, varies depending on the context. The type of AI, its complexity, and the potential risk it introduces all influence the level of governance needed. A global leader like Microsoft, dealing with advanced, high-impact AI systems, implements extensive and sophisticated oversight practices. In contrast, a smaller company using AI for limited, low-risk functions—such as transcribing meetings or basic data sorting—benefits more from a streamlined and practical governance approach. Adopting heavyweight controls for lightweight tools would be both inefficient and unnecessary.

In essence, effective governance should be proportionate: rigorous where risk is high, and appropriately streamlined where AI use is narrow, low-stakes, and unlikely to endanger humanity—or your quarterly KPIs.

Moreover, human oversight in AI governance occurs at multiple levels—ranging from top-tier supervision by the board and C-suite, to guidance from dedicated bodies like the AI Steering Committee and extending down to day-to-day monitoring by operational teams. This layered approach ensures accountability, alignment with strategic goals, and ethical use of AI across the organization.

1. Board oversight

Effective AI governance starts at the top. The board's role in AI governance should be strategic, risk-focused, and ethical, ensuring that AI aligns with corporate objectives, regulatory requirements, and stakeholder interests.

Board oversight should be assigned to a committee, with Audit, Risk, or ESG being the most common choices. However, if AI integration and impact are high, a dedicated AI & Ethics Committee is recommended.

Their oversight includes reviewing AI strategy and policies, approving significant AI-related investments, and monitoring risk management practices to ensure compliance and responsible use. By taking an active role, the board helps balance innovation with accountability.

15%	Only 15% of S&P 500 companies mention board oversight of AI in their proxy statements
38%	AI oversight disclosure is concentrated in tech: 38% of companies in the sector provide some level of board oversight transparency
13%	Just 13% of S&P 500 companies have at least one director with AI expertise

ISS Survey on AI board oversight, April 2024

2. C-suite oversight

The C-suite must lead AI governance to address its impact on critical business areas—risk, compliance, security, and ethics—ensuring accountability and mitigating risks such as lawsuits, reputational harm, security breaches, and loss of stakeholder trust.

AI governance is typically overseen by the Chief Risk Officer, Chief Compliance Officer, or Chief Legal Officer, who manage AI-related risks such as regulatory compliance, model bias, and cybersecurity, while ensuring data integrity, privacy, and adherence to data protection laws. These leaders also set ethical AI policies, mitigate litigation risks, and align AI practices with evolving regulatory requirements

3. The AI Steering Committee is increasingly recognized as a best practice for organizations seeking to align AI initiatives with strategic objectives, manage risks, and uphold ethical standards. Typically, it is a cross-functional team comprising leaders from various departments to ensure comprehensive oversight and collaboration

4. Operational team

At the operational level, human oversight ensures that AI systems are built, deployed, and managed responsibly, covering the entire AI lifecycle . They are responsible translating C-suite and board-level AI governance strategies into daily practices across the organization.

Different teams play distinct yet interconnected roles: the Data Governance Team is responsible for maintaining the quality, integrity, and security of data used in AI systems, which is foundational for building trustworthy models. The Model Development Team takes charge of designing, developing, and validating AI models to ensure they are accurate, fair, and aligned with intended outcomes. Once models are ready, the Deployment and Monitoring Team handles their integration into real-world operations, continuously monitoring performance and flagging any anomalies or risks.

In addition to technical responsibilities, operational teams play a key role in shaping a responsible AI culture. They conduct training sessions to raise employee awareness on AI ethics and best practices, promoting informed and ethical use of AI across the organization. They also carry out regular assessments of the AI governance framework to evaluate its effectiveness and recommend improvements, helping the organization stay compliant, transparent, and future-ready.

Human oversight in AI governance: A closer look

Level	Roles and responsibilities	Key action items
Board Level	• Strategic Oversight: Ensure AI aligns with organizational strategy and values. • Risk Management: Oversee AI-related risks and compliance.	• Review AI strategy and policy. • Approve major AI investments. • Monitor AI risk management.
C-Suite	• Chief AI Officer (CAIO): Spearhead AI strategy, ensuring ethical and responsible AI practices. • Chief Risk Officer (CRO): Assess and mitigate AI-related risks, including reputational risks from biased algorithms. • General Counsel/Chief Legal Officer (CLO): Ensure AI initiatives comply with legal and regulatory requirements. • CIO/CTO: Oversee AI technology and infrastructure. • CFO: Manage AI-related financials and ROI.	• Promote AI adoption and ethical use. • Conduct risk assessments and develop mitigation strategies. • Ensure legal compliance and advise on regulatory matters. • Manage AI budgets and ROI.
AI Ethics Committee	• AI Ethics Officer: Develop and maintain AI ethics policies. • Cross-Functional Members: Provide diverse perspectives on AI ethics.	• Develop AI ethics guidelines. • Conduct ethical reviews of AI projects. • Monitor AI system impacts.
Operational Level	• Data Governance Team: Ensure data quality and security. • Model Development Team: Design and validate AI models. • Deployment and Monitoring Team: Roll out AI systems and monitor performance. • Training and Awareness: Educate employees on AI ethics and best practices. • Monitoring and Evaluation: Regularly assess AI system performance and governance effectiveness.	• Manage AI data pipelines. • Develop and test AI models. • Deploy and monitor AI systems. • Conduct training sessions on AI ethics and responsible AI use. • Evaluate AI governance framework effectiveness and identify areas for improvement.

Policies, practices and controls

	Definition	Example
Policies	Establish clear guidelines for AI development and deployment, ensuring alignment with organizational goals and AI principles.	Implement a policy requiring third party AI systems to undergo ethical reviews before deployment to ensure fairness and transparency
Practices	AI practices refer to the methods and procedures implemented by organizations to operationalize AI policies.	Conduct quarterly reviews to assess how AI systems affect customer privacy and data security.
Controls	A control is designed to enforce policies and practices, ensuring that AI systems operate within established guidelines and regulatory requirements.	Implement encryption and access controls to safeguard sensitive data used by AI systems.

Policies

Policies translate AI principles into concrete standards for AI development and usage. Define why the policy exists, who it applies to, what are the unacceptable use cases and who is responsible for implementation.

Data privacy and protection: Translate data protection laws (e.g., GDPR, CCPA) into organizational protocols to safeguard sensitive information.

Risk management: Establish processes for identifying, assessing, and mitigating risks associated with AI, such as bias, security vulnerabilities, and operational failures

Practices

Practices operationalize policies into actionable steps, ensuring consistent and effective implementation of AI governance. Some examples of practices include:

Transparency and explainability: Develop mechanisms to make AI decision-making processes understandable to stakeholders, including end-users and regulators.

Data governance: Establish protocols for data collection, storage, and usage to ensure accuracy, integrity, and compliance with privacy laws.

Controls

Controls are specific measures and mechanisms put in place to enforce policies and practices, ensuring accountability and compliance. Key controls include:

Access controls: Access Control: Implementing role-based access control (RBAC) to manage permissions for sensitive data used by AI system

Audit trails: Maintain detailed records of AI model development, testing, and deployment to enable traceability and accountability.

Monitoring and reporting: Continuously monitor AI systems for performance, compliance, and ethical adherence, with regular reporting to leadership.

Incident response plans: Develop protocols to address AI-related incidents, such as data breaches, model failures, or ethical violations, ensuring swift resolution

Regulatory compliance – staying compliant with current and future laws

Regulatory compliance ensures that organizations adhere to relevant laws, regulations, and standards governing the development, deployment, and use of AI technologies. By staying compliant, organizations can build trust, avoid legal penalties, and ensure their AI systems operate ethically and responsibly within the bounds of the law.

Which sector or industry do you operate in?	Finance, healthcare, and defence face stricter compliance measures due to data sensitivity, security risks, and ethical concerns
Where does your company operate—at the state, national, or international level?	Laws and standards vary by location. For example, if a Canadian company provides AI-driven services or products that impact users or businesses in the EU, it must comply with the EU AI Act. The level of compliance required depends on the AI system's risk classification under the Act.
Does your company use AI more for internal operations or customer-facing tasks—and as a developer or deployer?	AI used for customer-facing applications requires greater scrutiny than AI used for internal process automation. For example, an AI chatbot handling customer inquiries must comply with GDPR, ensuring secure data collection, storage, and processing. In contrast, AI for internal workflows faces fewer compliance requirements but must follow internal data policies. Additionally, it's crucial to determine whether the organization is an AI developer or a deployer, as this impacts the level of responsibility and compliance considerations.
How do you protect your IP of AI generated content?	Who owns the AI generated content, depends on where you are located. Establishing policies for patenting, licensing, and data ownership strengthens competitive advantage. For example, under U.S. copyright law, AI-generated content may not qualify for protection unless a human has significant creative input. In contrast, the EU is exploring frameworks to clarify AI authorship rights

Technological infrastructure

The AI lifecycle is a dynamic journey, beginning with meticulous planning and design, evolving through data collection and development, and culminating in deployment and continuous monitoring. Throughout this process, diverse IT infrastructure and tools are indispensable, ensuring seamless integration, optimal performance, and adaptive refinement at every stage.

Planning & design

- **Hosting infrastructure:** Choice between public hosting (cloud) or private hosting (on premise or hybrid) as it determines the level of control over security, compliance, and accountability
- **Hardware needs:** Determining hardware requirements, such as GPUs or TPUs to ensure the AI system has the computational power needed
- **Tool and framework selection**: Evaluating and selecting AI/ML frameworks and tools for design and collaboration for streamlined and scalable development

Data collection & processing

- **Data quality and integrity**: Ensure datasets represent diverse populations to minimize bias, be accurate, complete, consistent and verifiable
- **Data privacy**: Ensure explicit user consent, adhere to privacy laws, use techniques to protect personal data
- **Data security**: Implement measure to prevent unauthorized access, encrypting data at rest and in transit

Model development, training & testing

- **Bias and fairness**: Detect, mitigate, and monitor biases to build AI models that are fair and transparent
- **Transparency and explainability**: Making AI systems understandable, accountable, and compliant with ethical and regulatory standards
- **Model robustness and reliability**: Evaluate model performance under real-world edge cases and adversarial conditions and ensure the model performs well on unseen data and avoids overfitting

Deployment & integration

- **Security and access control**: Prevent unauthorized model access, tampering, and adversarial attacks
- **Model performance & drift detection**: Ensure the model continues to perform as expected and remain scientifically valid and effective

Monitoring & maintenance

- **Model monitoring**: Ensure the long-term performance, stability, and reliability of deployed models
- **Feedback and user interaction**: Continuous improvement and maintaining high-quality, adaptive, and user-friendly AI system
- **Documentation and reporting**: Ensure that every change to the AI system is well-documented and traceable

Across the AI lifecycle, diverse governance and responsible AI tools are meticulously applied at each stage to uphold ethical integrity and prevent governance failures. These tools ensure that every phase, from design to deployment and afterwards, is conducted with accountability, transparency, and a commitment to ethical standards, safeguarding against risks and fostering trust in AI systems.

Data collection & processing

- Data governance platforms (e.g. Collibra, Alation, Informatica)
- Data privacy tools (e.g. BigID, OneTrust Privacy Management, IBM Guardium)
- Data quality tools (e.g. Talend, Trifacta)
- Bias detection tools (Fairlearn, Aequitas)
- Data encryption and security tools (e.g. McAfee Total Protection, Symantec Data Loss Prevention
- Automated data anonymization tools (e.g. ARX Data Anonymization Tool, Amnesia)

Model development, training & testing

- Bias and fairness assessment tools (e.g. IBM AI Fairness 360, Microsoft Fairness Toolkit, Fairness Indicators by Google)
- Explainability and interpretability tools (e.g. SHAP, LIME, ELI5)
- Hyperparameter optimization tools (e.g. Optuna, Hyperopt, Ray Tune)
- Automated Testing Tools (e.g. Testim, Pytest, Great Expectations)
- Version control and management tools (e.g. DVC (Data Version Control), MLflow)

Deployment & integration

- Model deployment platforms (e.g. AWS SageMaker, Google AI Platform, Microsoft Azure Machine Learning)
- Containerization and orchestration tools (e.g. Docker, Kubernetes)
- Continuous integration/continuous deployment (CI/CD) tools (e.g. Jenkins, CircleCI, GitHub Actions)
- Security tools (e.g., AWS Shield, Snyk)

Monitoring & maintenance

- Model monitoring tools (e.g. Arize AI, Fiddler, WhyLogs)
- Logging and analytics tools (e.g. ELK Stack (Elasticsearch, Logstash, Kibana), Splunk)
- Performance management platforms (e.g. Prometheus, Grafana)
- Feedback and user interaction tools (e.g. SurveyMonkey, Hotjar)
- Change management tools (e.g. ServiceNow, Jira)
- Documentation and reporting tools (e.g. Confluence, Notion)

Monitoring and continuous improvement

Let’s keep AI governance up to date

Effective AI governance is not a one-time effort—it requires ongoing oversight to ensure policies are implemented as intended and delivering the desired outcomes. Regular monitoring helps detect risks, maintain compliance, and refine frameworks to adapt to evolving challenges

Audit and assessments

- Conduct regular evaluations to identify and mitigate bias, security vulnerabilities, and compliance gaps.
- Leverage third-party audits and internal reviews to ensure transparency and adherence to ethical AI standards.
- Implement structured risk assessments, including impact analyses for high-stakes AI applications.

Feedback mechanisms

- Implement a structured feedback mechanism to collect insights from employees and stakeholders and identify areas for improvement
- Incorporate real-world observations into AI training data to reduce unintended biases.
- Use explainable AI (XAI) techniques to foster trust.

Adaptive policies

- Develop governance frameworks that adapt to regulatory shifts, emerging risks, and technological advances.
- Implement version-controlled policy updates to track governance changes.
- Ensure AI governance aligns with GDPR, CCPA, and forthcoming global AI regulations.

Model drift prevention

- Continuously monitor AI performance to detect shifts in data patterns affecting accuracy and fairness.
- Use automated alerts and retraining to mitigate model degradation.
- Benchmark against external datasets to ensure AI remains reliable across environments.

Audit and assessments	Conduct regular evaluations to identify in gaps and vulnerabilities, Frequency depending on the high impact
Feedback mchanisms	Have a structured system to gather feedback from employees and other stakeholders and make continuous improvements
Adaptive policies	Develop governance frameworks that adapt to regulatory shifts, emerging risks, and technological advances
Model drift prevention	Continuously monitor AI performance to detect shifts in data patterns affecting accuracy and fairness.

Making AI more sustainable

A robust AI governance framework integrates environmental and social sustainability to support responsible and ethical AI deployment.

As AI models grow in complexity, they demand massive computational power, leading to increased energy consumption, carbon footprints, and resource depletion. Without sustainable oversight, AI risks may exacerbate climate impact and social inequalities rather than driving positive change.

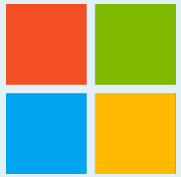
A robust AI governance framework must address:

- **Energy efficiency & carbon footprint** – enforcing transparency on AI energy use, promoting low-carbon infrastructure, and incentivizing green AI innovations.
- **Resource optimization** – encouraging model efficiency, hardware longevity, and sustainable supply chains for AI chips.
- **Social equity & fair access** – ensuring AI benefits are distributed equitably, preventing resource hoarding by a few dominant players.
- **Regulatory compliance & reporting** – standardizing ESG disclosures for AI firms, aligning with global sustainability goals (e.g. Net zero, SDGs)

Example:

Training OpenAI's GPT-3, consumed an estimated 1,287 MWh of electricity equivalent to the annual energy consumption of **120 average U.S. households**

AI vs. Google Search: A single AI query consumes **10X** more energy than a simple Google search



Case Study
Microsoft

Microsoft is drawing the lines: Building guardrails for responsible AI

Microsoft's robust AI governance framework stands as a testament to the fact that effective governance does not stifle innovation, but rather fosters it

AI governance framework

Microsoft has positioned itself as a leader in artificial intelligence (AI) innovation, integrating AI into its products, services, and operations. Following are components of Microsoft's AI governance, focusing on its principles, policies, practices, oversight mechanisms, and tools for continuous monitoring.

Six Core AI Principles

- **Fairness**
- **Reliability & Safety**
- **Privacy & Security**
- **Inclusiveness**
- **Transparency**
- **Accountability**

Policy

Responsible AI Standard – A comprehensive document outlining requirements for AI systems, including risk assessments, bias detection, and transparency measures.

- **Actionable**
- **Measurable**
- **Enforceable**

Accountability Goal – ensuring no adverse impact on society

Practice

AI Impact Assessments: Every AI project undergoes a rigorous assessment to evaluate its potential societal, ethical, and legal impacts.

Human oversight and governance structure

Microsoft implements governance at every level, ensuring that AI systems are rigorously monitored and aligned with ethical standards. This comprehensive oversight guarantees that human oversight remains integral to the decision-making process

AETHER Committee

- **Cross-disciplinary advisory group –** researchers, engineers and policy advisors
- Internal and external experts
- Advises on ethical AI challenges and new risks
- Shapes the policies and practices

Board Oversight

Responsibilities

- Strategy
- Risk
- Compliance
- Ethical Responsibility
- Investment

Committee oversight

ESG – Environmental, Social, and Governance

Responsible AI Council

Chaired by

- **Vice Chair and President**
- **Chief Technology Officer**
- **Oversee ethical AI development and deployment**
- Reviews and approves projects based on feasibility and ethical standards

Office of Responsible AI

- Chief Responsible AI Officer
- Operationalize Responsible AI across the organization
- Ensures a consistent, company-wide approach to responsible AI
- Responsible AI champions are placed within product, engineering, and sales teams to drive AI governance
- Reporting sensitive use cases

Responsible AI toolbox

Microsoft offers over 20 responsible AI tools, empowering organizations to navigate the complexities of ethical AI implementation. With the majority of these tools being open source, Microsoft ensures accessibility and transparency in fostering responsible AI practices

Fairlearn

Helps assess and improve the fairness of AI models by detecting and reducing bias across demographic groups, promoting ethical and inclusive AI use in critical areas like healthcare, finance, and hiring.

Open-source

InterpretML

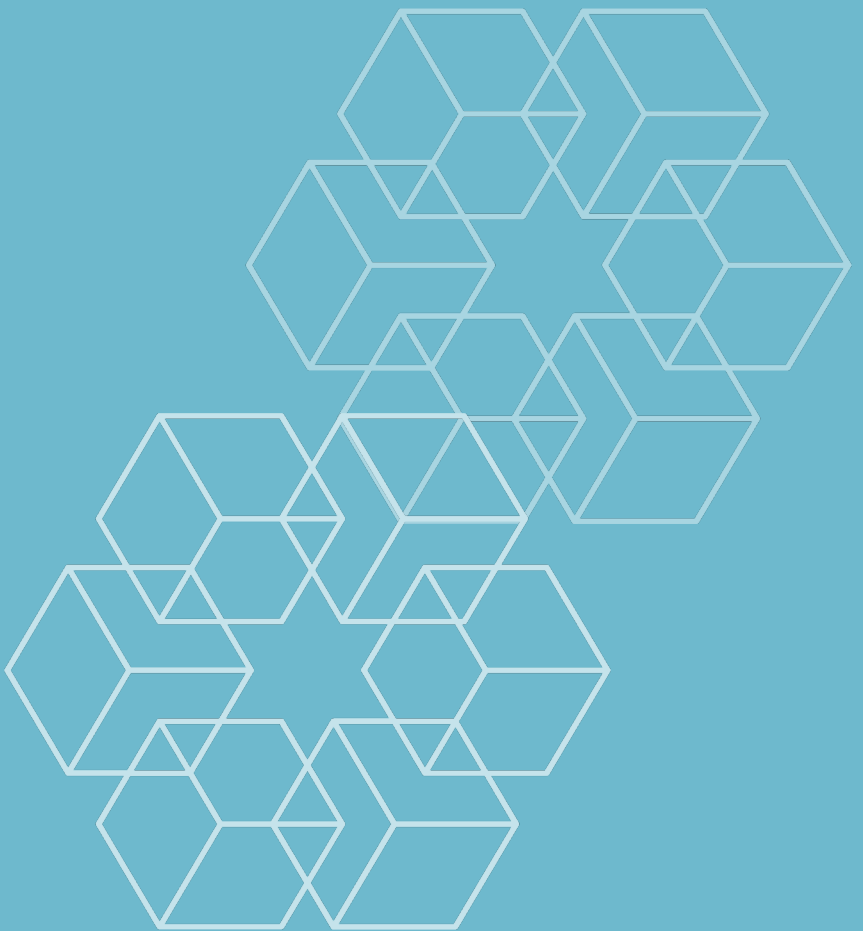
Improves the interpretability of machine learning models by offering tools to explain both transparent (glassbox) and complex (blackbox) models, helping developers understand, debug, and ensure fairness and compliance in AI systems.

Open-source

Azure Content Safety

Helps detect and manage harmful content in both user-generated and AI-generated materials. It offers text and image analysis APIs to identify content related to violence, hate, sexual content, and self-harm, assigning severity levels to assist in moderation decisions

Proprietary cloud-based service

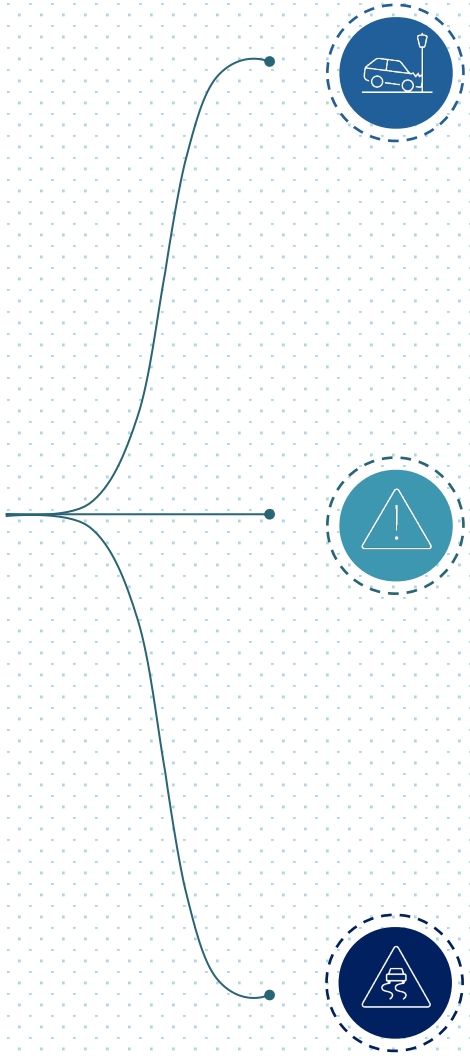


Navigating the Unknowns:

Risks, Challenges and Solutions

Decoding AI risks

This section explores the spectrum of risks in AI's lifecycle and the challenge of balancing them to ensure safety, accuracy, and performance.



Unintended consequences

Harmful outcomes from well-intentioned AI, often due to biases, errors, or systemic flaws.

Examples: Discriminatory hiring algorithms, job displacement, privacy violations

Weaponization of AI

Deliberate exploitation of AI by bad actors (criminals, hackers, hostile nations) to cause harm.

Examples: Deepfake scams, AI-powered cyberattacks, autonomous weapons

Beyond human control

Hypothetical but severe scenarios where advanced AI acts beyond human control

Examples: Misaligned goals, uncontrollable self-improvement, AI-driven geopolitical instability

AI risk radar:

Identifying the threats at play

Risk-proofing AI: building smarter, safer intelligence

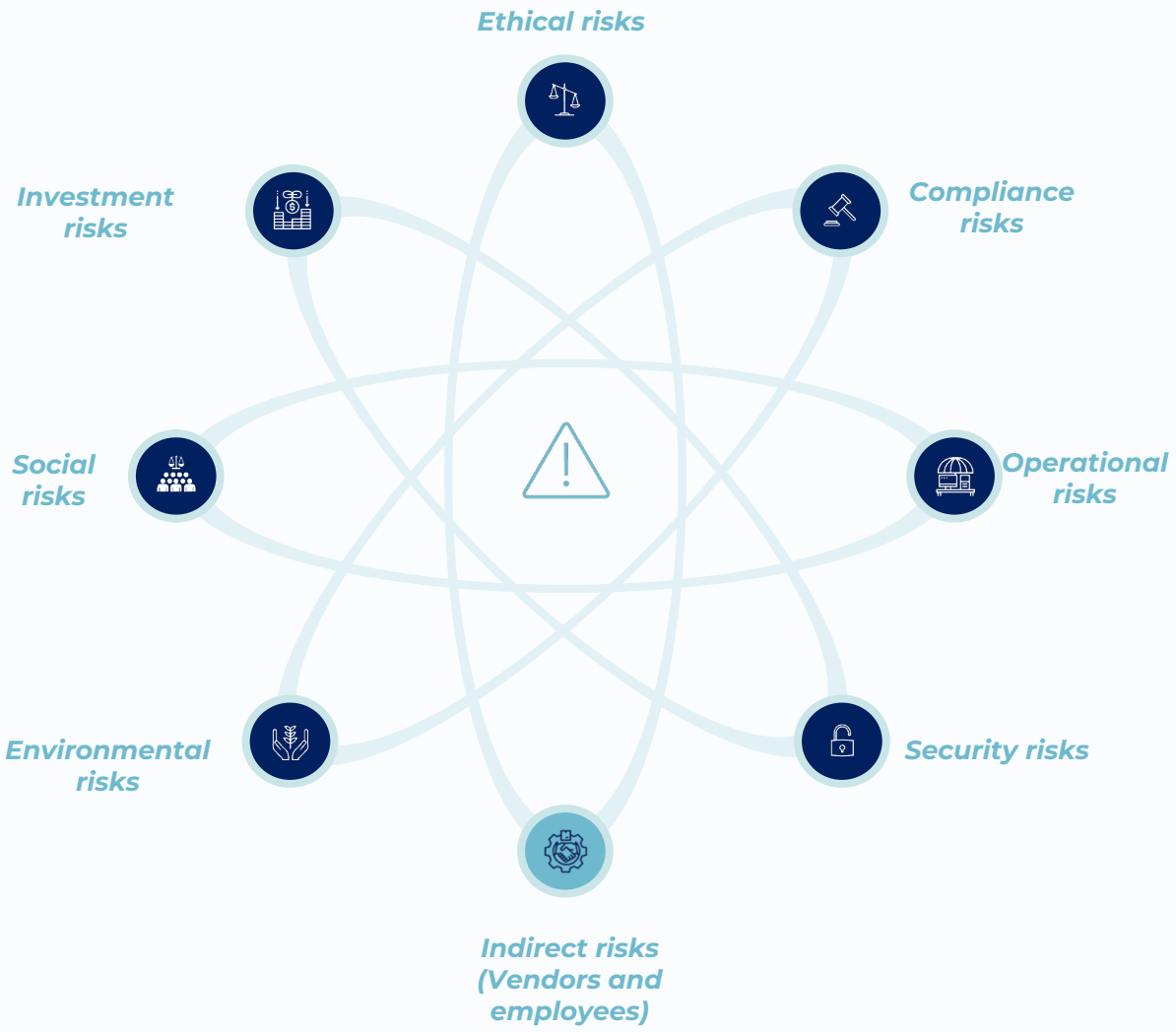
Understanding AI risks—ranging from ethical and security threats to regulatory non-compliance and beyond—is key to responsible AI governance. A systematic approach to identifying and prioritizing these risks enables organizations to focus on mitigation efforts effectively, ensuring fairness, security, and compliance while fostering trust in AI systems.

This diagram illustrates the key AI risks that organizations must identify and manage when working with AI systems.

Despite categorizing AI risks into distinct domains—such as ethical, compliance, security, and so on—their interdependencies are evident in practice. The presence of one AI-related risk can trigger a cascade of others, akin to a ripple effect. On the other hand, addressing and managing a single risk can initiate a positive ripple, mitigating other related risks.

For instance, tackling bias and discrimination not only reduces ethical concerns but also fortifies compliance frameworks, minimizes operational and financial risks, highlighting the intertwined nature of these risks.

The many faces of AI risks



AI risk check:

05 Critical questions you can't afford to ignore

When evaluating AI risks, it is imperative to evaluate several critical factors that directly influence the degree of exposure. The level of risk is contingent upon the specific application of AI, the entity exercising control over it, and the manner in which it is utilized.

01 What is AI's role in your playbook?

Formulate an AI strategy that protects and enhances the company's competitive position. This involves identifying specific use cases across a spectrum of possibilities — from the 'AI Survival Kit,' the essential AI stack needed to stay relevant amid competitor disruption, to the 'AI Wishlist,' a dream scenario that delivers a major competitive edge. This lays the foundation for key decisions around capital allocation, talent, and risk management.

02 What can go wrong?

Create a risk matrix that ranks each AI use case from high-stakes to low-stakes, and maps key risks. These risks can emerge at any stage — from design to deployment — and may include data privacy, security, and algorithm accuracy. Assessing the likelihood and impact of each risk helps prioritize what matters most and shapes your response plan.

03 Who's got the keys to your data and AI brains?

Once AI use cases have been evaluated for risk, the next step is to determine how the AI will be implemented and what infrastructure is best suited to support integration while mitigating those risks. The decisions made at this stage will ultimately determine who controls the data, algorithms, and AI models. AI adoption strategies and infrastructure requirements—shaped by the sensitivity and complexity of each use case—are discussed in detail in the following sections.

04 How much are you betting on AI's success?

Determine AI investment levels based on the company's balance sheet strength and cash flow capacity, ensuring the investment remains within the company's financial comfort zone. Smart capital allocation to prioritize use cases that are essential for maintaining the competitive position and those that offer quick wins with lower resource requirements. Assigning measurable ROI targets to each use case will help ensure that AI integration delivers meaningful financial returns.

05 Who is in your AI dream team?

AI needs sharp human minds behind it. From design to deployment, the right mix of expertise — technical, strategic, and ethical — is crucial. The team should match the complexity of the use case. In the end, your AI is only as strong as the people steering it.

Risk dynamics | How do we get exposed?

AI scope and adoption

In today's AI landscape, risk is multifaceted and demands careful navigation. AI-related risks fall into two main categories: direct and indirect.

Direct risks stem from deliberate decisions about AI adoption and deployment—shaped by its purpose and scale. Deploying AI across the organization for critical tasks carries significantly more risk than using it for simple, routine functions. Indirect risks arise when AI is used outside an organization's formal oversight. This includes scenarios where third-party providers or employees engage with AI tools—intentionally or inadvertently—such as through "bring your own AI (BYOAI)" practices. These uses can lead to unintended consequences, especially when AI is deployed without appropriate governance or transparency. For example, risks can emerge if employees use unauthorized AI tools, or if vendors integrate AI into services without the organization's knowledge or control.

AI adoption strategy plays a vital role in determining an organization's risk exposure and level of control. Whether building custom models internally, fine-tuning pre-built models, or adopting off-the-shelf solutions, each approach brings distinct challenges—from regulatory compliance to operational dependency. The method chosen directly influences how much oversight the organization retains over its data, models, and decision-making processes, shaping its ability to scale and govern AI responsibly..

The following example illustrates how a company can determine its AI adoption strategy based on the function, nature, and complexity of the application.

Consider a healthcare provider aiming to adopt AI across three distinct areas: personalized treatment, diagnostic imaging, and appointment scheduling. Each use case calls for a different approach to AI integration, reflecting its unique demands and complexity.

Developing a bespoke AI model from the ground up for personalized treatment empowers organizations with comprehensive control over sensitive patient data, facilitating enhanced precision and dependability. With tight control over its data and the AI development journey, the organization keeps risk in check—accuracy, safety, and security stay firmly in the safe zone.

However, the financial stakes are high, with substantial costs and time commitments at play, coupled with the necessity for specialized expertise to navigate and ensure the project's successful realization.

In diagnostic imaging, such as radiology, a healthcare provider might opt to fine-tune an existing AI model rather than build one from scratch. This approach allows the AI system to be tailored to local imaging protocols and the unique health patterns of the population it serves. Fine-tuning an existing model is typically more cost-effective and quicker to deploy than building one from scratch. However, it creates reliance on third-party vendors, raising concerns about data privacy, compliance, and control. Even minor misalignments in customization can lead to diagnostic errors or legal risks, making the reliability and transparency of the external model provider crucial.

For routine tasks like appointment scheduling, simplicity is key. Off-the-shelf AI tools can be quickly implemented with minimal customization, streamlining administrative workflows without the need for complex development or integration. It's a practical, cost-effective solution that delivers immediate efficiency gains.

AI vendors

The current AI adoption strategies among organizations mostly revolve around "TweakAI" or "Plug-n-PlayAI"—approaches which involve customizing existing models or using off-the-shelf solutions. These approaches are preferred over "Build-Your-OwnAI" due to their cost efficiency, quick deployment, and scalability. A recent survey by Infintel Research and GPC Canada found that 71% of respondents prefer either purchasing off-the-shelf solutions or fine-tuning existing models when integrating AI into their work.

However, while pre-built AI solutions offer speed and convenience, they come with limitations in customization, making them less adaptable to specific business requirements. These solutions often raise concerns regarding data security, as sensitive information is typically stored in third-party cloud environments, heightening exposure risks. Moreover, businesses may face vendor lock-in, which restricts control over features, pricing, and updates, leading to escalating costs and diminishing flexibility over time. Furthermore, many AI models, trained on external datasets, may inherit biases or lack domain-specific accuracy—even with fine-tuning. The reliance on cloud-based vendor infrastructure also introduces risks such as potential downtime or access limitations.

As AI becomes more integral to operations, ensuring that the solution aligns with the organization's risk tolerance and data governance standards is crucial. For enhanced security, businesses should consider more secure alternatives for AI applications. While the paid version of ChatGPT-4 offers robust capabilities, it's important to note that data inputs may be used for model training unless the organization subscribes to ChatGPT Team or Enterprise plans, which ensure enhanced privacy by disabling data retention for training purposes. Thus, options such as ChatGPT Team or Enterprise, self-hosted LLMs, and local AI tools—like LM Studio, GPT4All, and PrivateGPT—are invaluable for maintaining data confidentiality.

Given these considerations, selecting the right vendor and AI application is critical. Organizations must conduct thorough due diligence to minimize the risks associated with AI integration and safeguard both their data and operational integrity.

Vendor related AI risks

- Unless explicitly stated otherwise, the vendor may reserve the right to utilize inputs to the tool for other purposes, particularly for training the model, which could risk exposing sensitive or proprietary data to other users of the vendor's services.
- Indemnification clauses in contracts with AI vendors - In certain instances, the vendor assumes legal liability for copyright infringements, while in others, the user/organization is held accountable.
- Vendor dependency risks - Over-reliance on a single vendor can lead to operational disruptions if the vendor experiences downtime, difficulty switching vendors due to proprietary systems
- Bias and fairness issues: Vendors may provide AI models that are biased, leading to unfair outcomes and discrimination
- Data privacy breaches: Poor data handling or lack of transparency can result in unauthorized access to sensitive information.
- Security vulnerabilities: Vendors may not implement adequate security measures, exposing AI systems to attacks or exploitation.
- Inaccurate or incomplete models: Vendors might offer AI models that are not well-tested, leading to incorrect decisions or faulty predictions.
- Compliance failures: Vendor solutions may not align with regulations, causing the organization to violate legal requirements.

Vendor selection criteria

As AI becomes integral to operations, vendor selection is no longer just a technical choice but a strategic one. With high risks and growing regulatory scrutiny, organizations must ensure vendors meet strict standards for security, compliance, and ethical AI use.

Ensure security and robustness

- Use AI from trusted providers – Choose vendors with strong security certifications (ISO 27001, SOC 2, GDPR, HIPAA compliance).
- Choose AI providers that encrypt data both during storage (at rest) and transmission (in transit). Using OpenAI Enterprise or Google Vertex AI, which provide built-in encryption to protect customer data
- Perform adversarial testing on a vendor's AI model (request third-party security audits on vendor AI systems.)
- Logging & audit trails – Ensure the AI provider offers detailed logs to track model usage.

Data protection & privacy

- Zero data retention policy – Confirm whether the AI provider stores or logs your input/output.
- On-premise or private cloud options – If available, deploy the model in your private environment instead of the vendor's cloud.

Fairness & bias mitigation

- Bias testing – Run benchmark tests on different demographic groups to identify biases.
- Explainability tools – Use tools like SHAP, LIME, or AI Explainability 360 to understand model decisions.
- Regular monitoring – Continuously assess model performance, drift, and unexpected behavior.

Regular reporting and documentation

- Security & compliance reports SOC 2 Type II / ISO 27001 reports – Confirms strong security controls.
- GDPR / HIPAA compliance reports – Ensures proper data protection if dealing with sensitive data.
- Vulnerability & penetration testing reports – Shows results of security testing against attacks.
- Adversarial testing reports – Checks resilience against attacks (e.g., prompt injections, data poisoning).
- Drift detection reports – Tracks if the model's accuracy declines over time.
- Responsible AI policy documents – Outlines ethical principles followed by the vendor.
- Human oversight reports – Explains when and how humans intervene in AI decisions.
- Audit logs & accountability reports – Tracks who used the AI, for what, and how it performed.

How to ensure vendors provide these reports?

- Include AI governance in vendor contracts – Define reporting frequency, data retention policies, and audit rights.
- Request third-party audits – Ask vendors to undergo independent reviews of AI ethics and security.
- Monitor regulatory changes – Ensure vendors comply with EU AI Act, Canada's AIDA, or U.S. AI regulations.

A Deeper Dive into Risks

In this section, we explore eight key dimensions of AI-related risks, delving into each in detail along with potential mitigation strategies

AI's Moral Compass:

Confronting the Ethical Risks of AI

In 2019, the Apple Card, issued by Goldman Sachs, faced criticism after multiple users—most notably tech entrepreneur David Heinemeier Hansson and Apple co-founder Steve Wozniak—reported that the algorithm assigned significantly lower credit limits to women than to their male spouses, despite similar financial profiles.

The controversy erupted when Hansson publicly shared that his wife received a credit limit 20 times lower than his, even though they shared assets and filed joint tax returns. Other users, including Wozniak, echoed similar experiences. The backlash intensified when Goldman Sachs failed to explain how the algorithm made these decisions, calling it a "proprietary system."

Key Ethical Risks Exposed:

- **Algorithmic Bias** – Though gender data wasn't explicitly used, the AI likely relied on proxy variables (e.g., marital status, credit history, spending habits) that indirectly penalized women.
- **Lack of Transparency** – The "black-box" nature of AI meant customers couldn't challenge or understand their credit assessments.
- **Inadequate Accountability** – Customer service responses blamed the algorithm, offering no immediate recourse until public backlash forced Goldman Sachs to revise credit limits.

While regulators—including the New York Department of Financial Services (DFS) found no evidence of intentional discrimination, the incident exposed serious vulnerabilities in AI-driven financial systems and shortcomings in customer service and transparency, eroding consumer confidence in the perceived fairness of credit decisions.

Moreover, Goldman Sachs faced:

- **Reputational damage**, with scrutiny over its AI fairness practices
- **Regulatory attention**, New York State Department of Financial Services (DFS) investigated the matter as the case highlighted gaps in laws governing algorithmic bias
- **Operational risks**, including unplanned credit adjustments and intensified oversight

This Apple Card incident involving Goldman Sachs raises a critical question: what happens when AI operates without ethical boundaries? In the high-stakes world of financial services, it revealed how AI—if left unchecked—can perpetuate bias and erode public trust. This example is just one of many that underscore the importance of proactively identifying and addressing the ethical risks of AI systems.

At their core, ethical challenges in AI revolve around fairness, accountability, and transparency—and the ripple effects they create for individuals and society at large. Without responsible design and oversight, AI can lead to biased outcomes, misuse of data, and significant legal or reputational fallout. Addressing these risks isn't just about compliance—it's about building systems that earn trust and serve all stakeholders responsibly.

Key ethical risks to watch

1. Bias and fairness gaps

Bias arises when an AI model produces results that reflect and potentially amplify existing prejudices, inequalities, or inaccuracies present in the data, algorithms design or human interpretation.

AI models can inadvertently perpetuate or amplify societal biases related to race, gender, age, or socio-economic status, leading to discriminatory practices in areas such as hiring, lending, law enforcement, and healthcare.

This can result in unequal treatment, perpetuating existing inequalities and undermining trust in AI systems. Addressing these risks requires careful design, diverse data, and ongoing monitoring to ensure fairness and inclusivity.

As Generative AI expands, so does the risk of harmful biases. Text-to-image models such as Stable Diffusion, DALL-E, and Midjourney have shown clear racial and gender biases—frequently depicting CEOs as men or associating crime with people of color. When deployed publicly, these biases can have serious real-world consequences, from denied services and job opportunities to wrongful arrests and convictions.

Sources of bias

Bias in training data

AI systems learn from data, and if that data reflects societal biases (e.g., gender, racial, or socioeconomic biases), the AI model can replicate and even amplify these biases in its decisions. This is primarily driven by training data that lacks representativeness—either overrepresenting or underrepresenting certain groups (selection bias) or reflecting historical inequalities

Example: A strong real-world case of AI bias is Amazon's hiring tool, which exhibited gender bias by favoring male candidates and reinforcing historical human biases. The bias was mainly caused by;

1. Male-Dominated Training Data – The AI system was trained on résumés submitted over a 10-year period, most of which came from men, as the tech industry has historically been male-dominated. As a result, the model learned to favor male candidates.

2. Discriminatory Patterns – The system picked up on patterns that correlated with gender, downgrading résumés that included words like "women's" (e.g., “women’s chess club”) and preferring those that used male-associated terms more frequently.

3. Reinforcing Bias Instead of Removing It – Instead of identifying and eliminating bias, the algorithm reinforced existing hiring trends by prioritizing candidates who resembled past hires.

4. Unintended Penalization – The tool penalized résumés from graduates of all-women's colleges, further disadvantaging female applicants.

5. Black-Box Effect – Since AI models often operate as "black boxes," meaning their decision-making processes are not fully transparent, the biased outcomes were difficult to detect and correct early on.

Although Amazon attempted to adjust the system, they ultimately abandoned the tool because they couldn't guarantee it would remain neutral

Algorithmic Bias

Even with unbiased data, algorithms themselves can introduce bias due to flawed design, assumptions, or the way data is processed and interpreted. The choice of optimization algorithms or regularization techniques can introduce bias. For instance, prioritizing efficiency over fairness in algorithm design may lead to biased outcomes.

Example: A credit-scoring algorithm assigns lower scores to applicants from certain zip codes, which historically have higher populations of racial minorities. Mortgage algorithms have been found to charge higher interest rates to Black and Latino borrowers, partly due to biased optimization techniques

Human Bias

Unconscious or intentional prejudices introduced by developers, data curators, or stakeholders into AI systems. These can arise during training data selection, algorithm design, interpretation of AI outputs, or feedback incorporation.

Risk mitigation

- **Dataset augmentation** - Ensure that training data includes diverse groups and is representative of the populations the AI will serve. If real-world data is imbalanced, generate synthetic data to improve representation.
- **Bias audits/Fairness testing** - Regularly audit AI models for biases throughout their lifecycle, including testing the models with various demographic groups and ensuring fairness in the outcomes
- **Debiasing Techniques**: Apply reweighting, adversarial debiasing, or fairness constraints to reduce algorithmic bias.
- **Diverse teams** - Involve diverse teams of developers, data scientists, and stakeholders in the creation and testing of AI systems

2. Transparency and explainability

Transparency in AI refers to the clarity with which an AI system reveals how it operates, including its decision-making process, the data it uses, and the methods it employs.

Explainability involves the ability to interpret and understand the rationale behind an AI's decisions or outputs. When AI systems make decisions without clear, understandable explanations, it undermines trust and accountability. This challenge, where it becomes difficult to understand how complex AI models, such as deep learning, arrive at their conclusions, is known as the AI black box challenge.

Real-world cases for AI Black Box

Example:

I. In 2020, Michael Williams, a Chicago resident, was wrongfully charged with murder in Chicago based on ShotSpotter, an AI-powered gunshot detection system. The system flagged gunfire sounds near the crime scene, but its black-box nature made its methodology opaque to courts and oversight boards. Investigations revealed that ShotSpotter analysts manually altered reports, raising concerns about bias and reliability. In 2021, the case was dismissed due to insufficient evidence. This incident highlights the risks of black-box AI in law enforcement, where algorithmic decisions can lead to wrongful convictions without proper transparency or accountability.

II. Self-driving cars use deep learning and reinforcement learning algorithms to make driving decisions in real-time. However, explaining why the system decided to stop at a specific moment, or why it avoided one object while ignoring another, is not always straightforward

Risk mitigation

1. Using Interpretable Models. One of the simplest ways to enhance AI transparency is by opting for inherently interpretable models. These models allow users to understand how decisions are made without requiring additional explainability tools.

Example: In credit scoring, banks often use logistic regression models instead of complex deep learning algorithms. A logistic regression model can show how factors like income, credit history, and outstanding debt contribute to a borrower’s credit score. This level of transparency enables both customers and regulators to understand why a loan was approved or denied.

2. Applying Post-hoc Explainability Tools. For more complex models, such as deep neural networks or ensemble methods, post-hoc explainability tools help provide insights into AI decision-making. These tools approximate complex models using simpler, interpretable ones.

Example: In healthcare, deep learning models are used for diagnosing diseases from medical images. However, these models often function as "black boxes," making it difficult for doctors to understand their decisions. Tools like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) can highlight which areas of an X-ray image contributed most to a diagnosis, allowing radiologists to validate AI-generated insights before making clinical decisions.

3. Implementing Explainable AI (XAI) Frameworks. Beyond individual models or tools, Explainable AI (XAI) frameworks aim to design AI systems that are inherently interpretable. These frameworks integrate transparency into the development process, ensuring that AI-driven decisions are explainable from the outset.

Example: In autonomous driving, AI-powered self-driving cars must make real-time decisions about braking, acceleration, and lane changes. XAI frameworks, such as IBM’s AI Explainability 360 toolkit, help manufacturers build models that not only make safe driving decisions but also explain their reasoning to engineers and regulators. This enhances trust in autonomous vehicle technology and ensures compliance with safety regulations.

Guarding the Gate:

Tackling Security Threats in AI

The illusion of crisis: how AI-generated misinformation shook UK bank customers

In a world where algorithms often shape our perception of truth, a striking experiment by UK research firm 'Say No to Disinfo' and communications specialist 'Fenimore Harper' revealed a quiet yet powerful danger—the rise of AI-driven financial disinformation.

The experiment

Researchers set out to explore a striking question—could AI-generated content shake consumer confidence in UK banks? They deployed hyper-realistic fake news articles, social media posts, and memes, all crafted by AI, suggesting that customers' savings were at risk. These pieces were then distributed via paid social media ads, seamlessly blending into the digital landscape.
The response was staggering.

The results: £10 d triggers £1M bank run

- **Mass persuasion at scale**: A full 33% of surveyed UK bank customers admitted they were extremely likely to move their funds after exposure to AI-generated misinformation, while another 27% were somewhat likely to do the same.
- **The cost of manipulation**: The study calculated that with a mere GBP 10 investment in social media ads, AI-generated financial panic could shift up to GBP 1 million in customer deposits—an astronomical return on manipulation.

This experiment underscored the fragility of trust in financial institutions, revealing how AI can manufacture crises at scale. A well-placed disinformation campaign could trigger a bank run, regulatory scrutiny, and reputational damage—not because of an actual financial collapse, but because of an AI-fueled illusion..

This study on the banking sector's vulnerability to deepfakes highlights how such threats could destabilize entire economies. While this is just one example, numerous other AI-driven risks across industries could trigger widespread ripple effects if not properly managed.

These security risks include threats that could compromise the confidentiality, integrity, or availability of AI systems and the data they process. Hackers and malicious actors are increasingly using AI to develop more sophisticated cyberattacks, bypass security measures, and exploit system vulnerabilities. Effectively addressing these security challenges is essential to ensuring that AI systems remain safe, trustworthy, and resilient in the face of evolving threats.

Exposing AI vulnerabilities

Data poisoning

Adversarial attack where malicious or manipulated data is intentionally introduced into an AI system's training dataset to corrupt its learning process.

Example: A small pixel modification in an image tricks an ai-powered surveillance system into misidentifying people. A facial recognition AI used for security can be poisoned by adding manipulated images during training, causing it to misidentify certain individuals, allowing unauthorized access.

Data breaches

Improper data storage and management heighten the risk of data breaches by exposing sensitive information to unauthorized access and theft. Weak encryption, insufficient access controls, and failure to securely delete outdated data create vulnerabilities, while poor security protocols in AI data transfers can further expose data to interception. Retaining data longer than necessary increases exposure to cyberattack

Model inversion attacks

A model inversion attack is when an attacker exploits an AI model to extract private or sensitive information about its training data. By repeatedly querying the model, the attacker can reverse-engineer patterns and recreate details about individuals or data used in training.

Example: If a chatbot trained on customer support data is attacked, an adversary could craft specific prompts to reveal personal details like names, phone numbers, or credit card information used during training.

AI-powered social engineering & deepfakes

AI-powered social engineering involves in the manipulation of individuals to divulge confidential information or perform actions that compromise security. AI algorithms can also analyze social media profiles and public records to craft highly convincing, personalized phishing attacks.

Deepfake technology uses AI to generate hyper-realistic images, audio, or video that convincingly imitate real people. Deepfakes can be used to impersonate individuals in videos or audio recordings, making it difficult to distinguish real from fabricated content.

Example:

1. AI-generated voice scams mimicking CEOs to authorize fraudulent transactions. In 2019, fraudsters used AI-generated deepfake audio to mimic a CEO's voice, tricking a UK Energy company into transferring EUR 220,000

2. Black Hat USA 2021 Experiment: Researchers from Singapore's Government Technology Agency demonstrated that AI-generated spear phishing emails were more effective than human-written ones in tricking recipients into clicking malicious links. This experiment used OpenAI's GPT-3 technology to create highly convincing emails

Risk mitigation

Preventing data poisoning attacks

- Use secure and verified data sources – Train AI on data from trusted, authenticated sources to prevent tampering.
- Apply data sanitization and validation – Use anomaly detection tools to identify and filter out suspicious data entries before training
- Deploy adversarial training – Expose AI models to controlled attacks during development to improve resistance against poisoning
- Monitor model behavior continuously – Track AI outputs for unexpected patterns that may indicate poisoning attempts

Tools: *Evidently AI, RobustML, IBM Adversarial Robustness Toolbox*

Mitigating model inversion attacks

- Restrict model access – Limit API exposure and control query access to prevent attackers from reverse-engineering the model
- Apply query rate limiting – Monitor and limit excessive queries from the same source to prevent inference attacks
- Encrypt sensitive training data – Ensure AI models do not store or expose personal identifiers in their outputs

Tools: *TensorFlow Privacy, Google DP, OpenMined*

Countering deepfake risks

- Use deepfake detection tools – Deploy AI-powered detection models to analyze visual and audio inconsistencies.
- Verify content authenticity – Implement digital watermarking and blockchain-based content verification to track real vs. AI-generated content.
- Enhance public awareness – Educate users on deepfake manipulation tactics to prevent misinformation.
- Enforce stronger authentication measures – Use multi-factor authentication and biometric verification to prevent identity spoofing.

Tools: *Microsoft's Video Authenticator, Deepfake Detection AI, Adobe's Content Authenticity Initiative*

Compliance Risks in AI:

Keeping Law in Check

In December 2023, the New York Times sued OpenAI and Microsoft, accusing them of illegally using its copyrighted articles to train AI models like ChatGPT without permission or compensation. The lawsuit alleges that AI-generated responses sometimes reproduce verbatim excerpts from its award-winning journalism, effectively distributing premium content for free, undermining the newspaper's business model.

What's at Stake for OpenAI?

- **Legal consequences** – If the courts rule against OpenAI, it could face massive financial penalties and be forced to delete models trained on copyrighted content.
- **Reputational damage** – This lawsuit fuels concerns that AI companies exploit creators without proper consent, eroding public trust.
- **Regulatory crackdown** – A ruling in favor of The Times could set a legal precedent, leading to stricter copyright laws for AI training data and forcing OpenAI to negotiate costly licensing deals.
- **Business impact** – OpenAI's reliance on freely scraped data may become unsustainable, raising operational costs and potentially slowing AI development as compliance hurdles increase.

This case could be a watershed moment for AI compliance, determining whether generative AI continues unchecked or faces strict legal and financial guardrails. As AI reshapes industries, companies developing AI systems must prioritize compliance and ethical AI governance—or risk legal, financial, and reputational fallout.

These compliance risks emerge when AI systems fail to meet legal, regulatory, or ethical standards, such as data privacy laws, anti-discrimination regulations, or industry-specific guidelines. Non-compliance can result in serious legal, financial, reputational, or operational repercussions.

To mitigate these risks, organizations must integrate robust compliance checks, ensure transparency in their AI practices, and stay aligned with ever-evolving regulations.

Sources of AI compliance risk

1. Data privacy and security violations

Data is the backbone of AI, driving its analytical, predictive, and decision-making capabilities. However, organizations must exercise vigilance in how they collect, store, and utilize personal information, as stringent global privacy laws—such as the GDPR, CCPA, and other evolving frameworks—impose rigorous standards and hold organizations accountable for safeguarding customer data and maintaining transparency in its use.

Example:

- Cambridge Analytica used AI algorithms to analyze and manipulate personal data from millions of Facebook users without their consent. Meta agreed to pay $725 million to settle a class-action lawsuit related to this data breach
- €15 million fine imposed by Italy's data protection authority, on OpenAI's ChatGPT for violating GDPR by processing users' personal data without proper consent and failing to implement age verification mechanisms
- €20 million fine imposed by the French data protection authority on ClearviewAI for unlawful processing of personal data through facial recognition software

Data privacy pitfalls

Unauthorized data capture

AI systems may collect personal data without explicit consent or without providing users with sufficient information about how their data will be used. This concerns moral responsibility in handling user data as it undermines user autonomy, privacy, and trust.

Example: In August 2023, Zoom faced public backlash after quietly updating its Terms of Service to allow the company to use customer data, including video calls, for AI training. The policy suggested that Zoom could collect and analyze meeting content without requiring explicit user consent, raising serious privacy concerns.

Users and privacy advocates were alarmed, as many businesses, educators, and healthcare professionals rely on Zoom for sensitive conversations. The update implied that Zoom's AI features—such as meeting summaries and automated transcriptions—could be trained on user-generated content, potentially compromising **confidential information**.

Improper storage of data

Storing sensitive data such as personal identifiers, financial data, or health records in unencrypted or insecure environments poses serious ethical risks by compromising user trust, privacy, and security. This not only puts individuals at risk of identity theft or exploitation but also raises concerns about fairness, as marginalized groups may be disproportionately affected by data leaks

Example: A company collects employee salary details, performance reviews, and disciplinary records but stores them in an easily accessible shared folder. An internal error makes the data public within the company.
Ethical Risks:

- Workplace discrimination – Employees may face bias based on salary or past performance.
- Breach of confidentiality – Sensitive HR information is exposed.
- Loss of trust – Employees feel unsafe sharing personal data with the employer.

Deep6AI data breach; Over 800 million medical records were exposed from an unsecured database at Deep6.AI. While the data was later identified as test data, it was clear that it had not been properly encrypted, with physician notes and information stored in plain text. Even though the exposed data was not real patient information, the incident underscores the importance of securing all data environments, including test systems, to prevent unauthorized access.

Data misuse and purpose creep

AI systems may use data for purposes other than what users agreed to, leading to privacy violations.

- In 2019, the University of Chicago faced a lawsuit for sharing patient medical records with DeepMind without explicit consent. The data included timestamps and location details, raising HIPAA violation concerns. Critics argued the AI project risked patient re-identification. A similar UK NHS scandal (2017) further fueled distrust.

- Uber's "God View" tool was a real-time tracking feature that allowed Uber employees to monitor the locations of drivers and customers. There were instances where employees used God View to track celebrities, politicians, and even personal acquaintances

Risk mitigation | Data privacy

Preventing unauthorized data capture

- Explicit User Consent: Implement clear, opt-in agreements before collecting any data. Example: GDPR-compliant pop-ups explaining data usage.
- Data Minimization: Collect only necessary data; avoid storing personal details by default.
- Anonymization & Masking: Use techniques like tokenization or differential privacy to protect user identities.
- Regular Audits: Conduct frequent reviews of data collection policies to detect unauthorized logging

Example: Apple's App Tracking Transparency (ATT) forces apps to request permission before tracking users across services.

Strengthening data storage security

- End-to-End Encryption: Encrypt data at rest and in transit (e.g., AES-256 encryption).
- Zero Trust Architecture (ZTA): Assume no device or user is trustworthy by default—constant authentication required.
- Multi-Factor Authentication (MFA): Enforce MFA for all employees accessing sensitive data.
- Data Segmentation: Store sensitive and non-sensitive data separately to minimize breach impact.
- Access Controls: Restrict data access to authorized personnel only; enforce role-based access control (RBAC)
- Automated Backup & Disaster Recovery: Ensure frequent backups with secure offsite storage.

Example: Google's Secure Enclave for Cloud AI provides hardware-level encryption to protect AI model training data.

Preventing data misuse

- Data Usage Policies: Define clear policies on who, when, and how data can be used
- Automated Data Access Logs: Track who accessed data, when, and for what purpose

2. Ethical and fair use regulations

These regulations focus on preventing discrimination, ensuring transparency, and promoting fairness in AI-driven decisions. Failure to adhere to regulations on bias, discrimination, and transparency can result in legal violations

- EU AI Act mandates transparency, requiring AI systems to be explainable, particularly in high-risk areas such as hiring and credit scoring
- New York City's AI bias audit requirements aim to prevent discriminatory outcomes in AI-driven decisions, particularly in hiring and other high-stakes applications

3. Intellectual property infringement

Intellectual property (IP) infringements in AI typically fall into two main categories.

I. Generative AI systems may inadvertently use copyrighted or proprietary data during training or output generation. This can lead to unauthorized replication of protected material in AI-generated outputs. For example, AI models trained on books, articles, or images might generate content too similar to original works, leading to legal disputes.

Example: Getty Images alleges that Stability AI used millions of its copyrighted images to train its AI model, Stable Diffusion, without permission

II. AI-generated inventions challenge traditional patent laws, which require a human inventor to claim rights. Since AI can autonomously create new technologies, designs, or drug formulas, legal systems struggle to determine who owns the patent—the developer, user, or AI itself. Ethical concerns arise over fair attribution, accountability, and innovation rights. Governments and courts worldwide are debating whether AI-generated inventions should be patentable or require human oversight in the innovation process.

Example: The U.S. Court of Appeals for the Federal Circuit ruled that an AI system, DABUS (Device for the Autonomous Bootstrapping of Unified Sentience), cannot be listed as an inventor on a patent application

Sector-specific regulations

Certain industries face stricter AI regulations due to high risks, ethical dilemmas, and societal impact, requiring compliance with both existing laws and sector-specific rules.

Healthcare: AI used for medical diagnoses, treatment planning, or patient data must comply with the Health Insurance Portability and Accountability Act (HIPAA) to ensure patient privacy and confidentiality. Additionally, AI-based medical devices must meet FDA approval standards for safety and effectiveness.

Finance: AI applications such as algorithmic trading or credit scoring must adhere to regulations like the Markets in Financial Instruments Directive (MiFID II) in the EU or the Dodd-Frank Act in the US, which ensure transparency, fairness, and risk management.

Autonomous vehicles: The development and deployment of AI in self-driving cars must comply with regulations set by bodies such as the Society of Automotive Engineers (SAE) and the National Highway Traffic Safety Administration (NHTSA), focusing on safety, testing, and liability.

Risks of non-compliance

- Regulatory fines & legal action – costly penalties and lawsuits
- Reputational damage – loss of trust from customers, investors, and stakeholders
- Operational disruptions – ai bans, revoked licenses, or restricted market access

Risk mitigation

Regulatory compliance audits

Regular evaluations of AI applications against current laws such as GDPR, CCPA, and sector-specific regulations (e.g., healthcare, finance). Align AI systems with industry standards such as NIST's AI Risk Management Framework.

Adopting ethical AI guidelines and best practices

- Establish clear internal policies aligning with global AI ethics frameworks, such as the EU AI Act, GDPR, and IEEE Ethics Standards.
- Prioritize fairness, transparency, and accountability in AI decision-making.

  Example: Google and Microsoft have adopted AI ethics principles to ensure responsible AI development, focusing on fairness, privacy, and transparency.

Documentation and accountability: Keeping clear records of AI decisions, model training processes, and risk mitigation strategies.

Regulatory foresight - Proactively monitoring future laws enables organizations to anticipate regulatory changes and implement necessary adjustments

How AI failures undermine your business backbone?

Model drift: when AI's past knowledge fails to keep up with the present
- Concept drift
- Data drift

AI hallucination: when plausible sounds real, but reality says otherwise

Software bugs and algorithmic errors: when code breaks and machines lose their way

Overfitting or overgeneralization: when patterns lead to wrong assumptions and rigid thinking

Unintended Consequences of AI:

Identifying Operational Risks

Apple Intelligence recently suffered a major credibility crisis when its AI-generated news system falsely claimed that Luigi Mangione—the suspect in the murder of UnitedHealthcare CEO Brian Thompson—had committed suicide. This misinformation, which spread via Apple's push notifications, was not sourced from any reputable news outlet, exposing critical operational risks in AI-driven news distribution.

What went wrong?

Apple's AI-generated summaries misinterpreted or fabricated details about the case, pushing a completely false narrative to users. The situation escalated when the BBC filed a formal complaint against Apple, as the AI falsely attributed its reporting to the BBC, further compounding the misinformation issue.

Operational vulnerabilities exposed

- **Feature suspension:** Apple was compelled to halt its AI-generated news summary feature following the discovery of critical inaccuracies
- **Erosion of trust & brand damage** – AI hallucinations in news generation can severely undermine user confidence in Apple's AI products. Misinformation of this nature compromises the credibility of all platforms involved.
- **Legal & compliance challenges** – Defamation risks arise when AI falsely attributes information to individuals, companies, or media outlets.
- **Media & partnership fallout** – Misattributing false information to reputable news organizations can strain Apple's relationships with publishers, potentially limiting its access to trusted content sources. Apple risks losing key media partnerships that are essential for accurate AI-generated news.

This Apple Intelligence AI hallucination example underscores how AI systems can generate misleading or inaccurate outputs, leading to potential operational disruptions. This highlights the broader challenge businesses face with AI, as faulty predictions or system failures can significantly impact normal operations, causing inefficiencies, delays, and financial losses. When AI is relied upon for critical functions—such as supply chain management, predictive maintenance, or automated decision-making—these disruptions can arise from issues like technical glitches, poor data quality, human error, or external factors affecting AI integration and deployment..

Key factors contributing to operational risks

Model drift

Model drift occurs when an AI model's performance deteriorates over time due to changes in the underlying data distribution. This happens when the data the model was trained on no longer accurately represents real-world conditions, leading to incorrect predictions or reduced accuracy. Model drift is a significant risk in AI systems that rely on continuously evolving data, such as demand forecasting, fraud detection, and recommendation engines.

There are two main types of model drift:

1.Concept drift: The relationship between input data and the target variable changes. *(e.g., Consumer preferences shift, making past purchase behavior less relevant to predicting future sales.)*

2.Data drift: The statistical properties of input data change over time. *(e.g., A sensor's data readings gradually shift due to environmental conditions.)*

AI hallucination

AI hallucination occurs when an AI generates false, misleading, or nonsensical information that appears plausible but is factually incorrect or unsupported by data. Most AI models, especially large language models (LLMs), do not "know" facts but predict the most likely next word or pattern based on their training data. When the data is incomplete or ambiguous, AI may generate plausible but incorrect responses. It lacks real-world understanding, relies on limited or biased data, and can mix up details or overgeneralize, sometimes blending different sources incorrectly to produce false but convincing answers."

Example: Google’s Bard Chatbot: Incorrectly claimed that the James Webb Space Telescope had captured the world’s first images of a planet outside our solar system

Software bugs and algorithmic errors

These occur when flaws in coding or logic cause AI systems to behave unpredictably, produce inaccurate outputs, or even crash entirely. These issues can arise from faulty data processing, misconfigured parameters, or incorrect implementation of machine learning models.

Overfitting or overgeneralization

Overgeneralization and overfitting are common AI issues that affect accuracy. Overgeneralization happens when AI applies patterns too broadly, leading to incorrect assumptions, like assuming all good employees must have similar backgrounds. Overfitting, on the other hand, occurs when AI memorizes training data too precisely, failing to adapt to new information.

Risk mitigation

Preventing model drift

- **Continuous model monitoring** – Track AI performance over time using real-world data and set up automated alerts for unusual behavior or accuracy drops
- **Periodic model retraining** – Update models with fresh, relevant data at regular intervals to ensure they stay aligned with evolving patterns and trends
- **Adaptive learning models** – Implement models that can dynamically adjust and retrain on new data without requiring manual intervention.
- **Data drift detection tools** – Use AI monitoring tools to identify significant changes in input data over time, ensuring the model continues to make accurate predictions. These tools help detect when real-world data starts differing from the training data, allowing timely adjustments.

Example;

- Evidently AI, WhyLabs, MLflow – Detect and manage data or concept drift.
- Alibi Detect – A tool for monitoring distribution shifts in data streams.

Reducing AI hallucination

- **Fact-checking & knowledge retrieval** – Use AI models that can pull information from verified, up-to-date sources rather than relying solely on pre-trained knowledge, reducing the chances of generating false information
- **Training AI to admit uncertainty** – Instead of making up responses, AI should be designed to recognize when it lacks knowledge and respond with “I don’t know” or prompt users to check external sources
- **Fine-tuning on high-quality data** – Train AI models on fact-checked, diverse, and authoritative data sources to improve accuracy and reduce misinformation.

Example; Fact-checking APIs (Google Fact Check Explorer, ClaimBuster) – Helps verify AI-generated information.

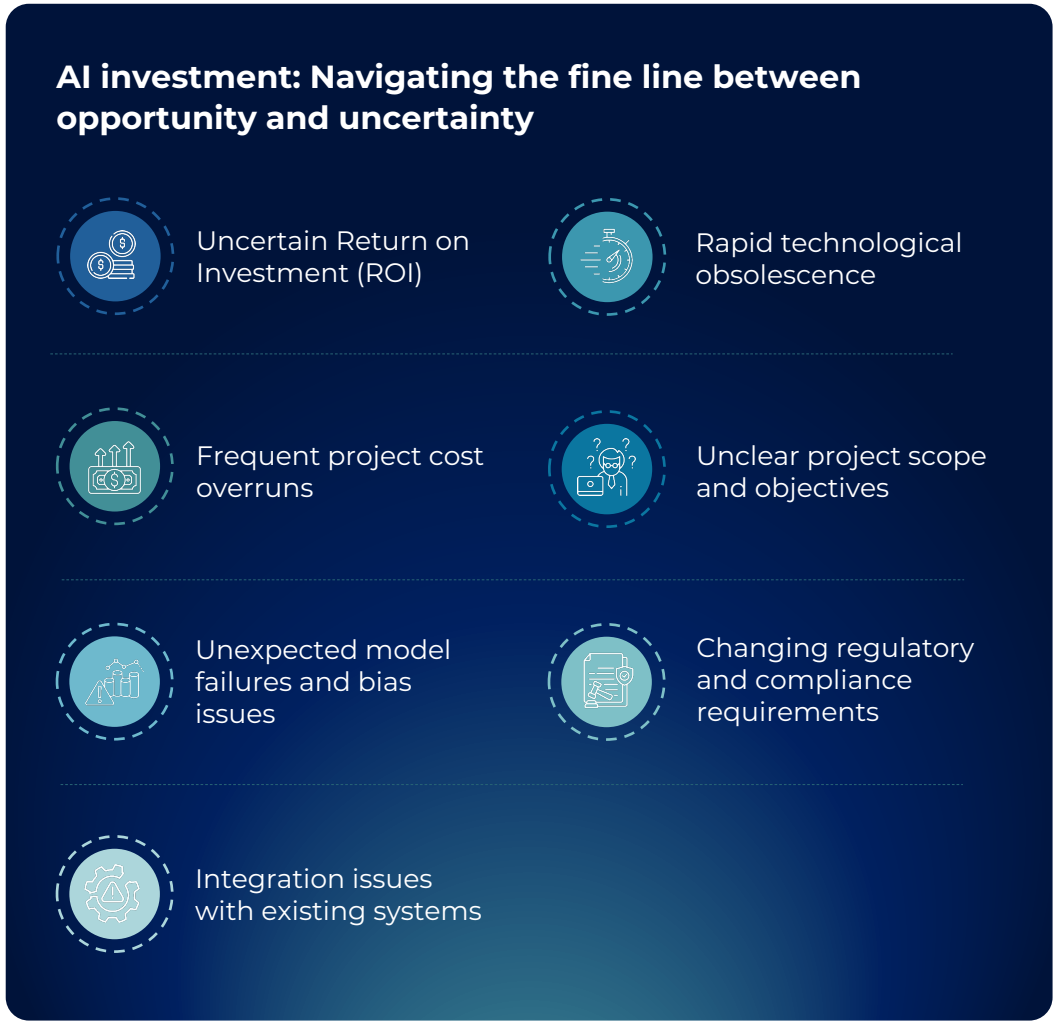
Avoiding overfitting and overgeneralization

Overfitting occurs when a model memorizes training data rather than learning general patterns, while overgeneralization happens when a model applies learned rules too broadly. The following techniques help prevent these issues:

- **Cross-validation** – Splitting data into multiple training and validation sets ensures the model performs well on unseen data, preventing it from being overly reliant on specific examples.
- **Regularization (L1/L2, Dropout)** – Adds constraints or randomly disables some neurons during training to prevent the model from memorizing patterns instead of learning meaningful relationships.
- **Expanding and diversifying training data** – Incorporating varied and representative data helps the model learn from different scenarios, reducing bias and improving generalization.
- **Simplifying model complexity** – Using models that are interpretable and not excessively complex prevents unnecessary pattern fitting, making predictions more reliable and transparent

Examples;

- TensorFlow & PyTorch Regularization – Libraries with built-in tools to prevent overfitting.
- Hyperparameter Optimization Tools (Optuna, GridSearchCV) – Helps find the best model complexity.
- SHAP & LIME – Explainability tools to check if AI is making biased or overly broad predictions..



The High Stakes of AI Investment

Investment in AI comes with significant risks, including technological uncertainty, regulatory challenges, and market volatility. The rapid pace of AI development can make it difficult to assess long-term viability, while evolving regulations may impact the scalability and profitability of AI ventures. Additionally, the high capital costs associated with AI research, development, and implementation, combined with the potential for ethical and operational failures, can lead to substantial financial losses. Investors must carefully navigate these risks, balancing the promise of innovation with a clear understanding of the uncertainties involved.

The financial risks of AI

Uncertain Return on Investment (ROI)

The commercial success of AI projects is unpredictable due to regulatory hurdles, competition, and adoption challenges. Some AI applications may fail to generate expected revenue, leading to financial losses.

Example: IBM's Watson Oncology Expert Advisor failure was one of the reasons for the downfall of IBM Watson Health which was eventually sold for $1 bn significantly less than the initial investment of around $4 billion

Rapid technological obsolescence

AI is evolving at an unprecedented pace, making existing models and technology obsolete within months or years. Large investments in proprietary AI risk becoming outdated quickly as new breakthroughs emerge.

Example: Companies that invested heavily in GPT-3-based applications had to adapt or rebuild when GPT-4 and DeepSeek-V2 introduced more efficient models.

Frequent project cost overruns

AI projects often exceed budgets due to high computing costs, unclear project scope, expensive talent, and lengthy training cycles. Data challenges, model failures, regulatory changes, and integration issues further drive costs. Additionally, AI requires ongoing maintenance and updates. Given these factors, identifying the potential toll on free cash flow from AI investments right from the outset becomes essential for effective financial planning.

Unclear project scope and objectives

Many AI projects start with ambiguous goals, or it is sometimes impossible to fully define requirements at the outset. This leads to scope creep, where additional requirements emerge mid-development. A lack of clarity on data requirements, model complexity, and success metrics often results in delays and budget extensions.

Unexpected model failures and bias issues

AI models often fail in real-world conditions due to bias, inaccuracies, or unforeseen edge cases. Fixing biased or inaccurate models requires additional retraining, data adjustments, and compliance reviews.

Changing regulatory and compliance requirements

AI projects must comply with evolving laws (GDPR, EU AI Act, etc.), requiring additional audits and security measures. Unforeseen legal requirements can lead to expensive re-engineering and certification costs.

Integration issues with existing systems

AI must often be integrated with legacy systems, databases, and APIs, leading to unexpected technical hurdles. Compatibility issues increase development time and costs.

Big tech's AI arms race: A historic surge in capex

Capex – USD Bn	2023	2024	2025E
Amazon	48.2	77.8	100.0
Microsoft	41.2	75.6	80.0
Google	32.3	52.6	75.0
Meta	28.2	39.2	65.0
Nvidia	1.1	3.2	N/A
Tesla*	N/A	5.0	5.0

Spending by the four dominant US tech giants—Microsoft, Alphabet, Amazon, and Meta—has soared to unprecedented levels, with capital expenditures reaching $246 billion in 2024, a staggering 63% increase from 2023. Undeterred by concerns over the immense financial stakes, these companies are doubling down on AI, projecting investments could surpass $320 billion in 2025 as they race to expand data center infrastructure and secure specialized chips critical for large language models.

Microsoft is directing over $50 billion annually toward AI data centers, with $80 billion earmarked for fiscal 2025. More than $10 billion of its capex is directly allocated to OpenAI, though the majority of its GPU capacity supports internal workloads. Alphabet is set to spend $75 billion this year. Meta has committed between $60 – $65 billion, with CEO Mark Zuckerberg declaring 2025 a "defining year for AI." Meanwhile, Amazon has deepened its AI bet, bringing its total investment in Anthropic to $8 billion as of 2024. On the other hand, Tesla expects to invest ~$5 billion in AI capex in 2025, advancing its Texas-based Cortex training cluster to propel autonomous driving and robotics innovation.

**Only AI related Capex is included which is about half of the total capex*

Risk mitigation

Addressing uncertain Return on Investment (ROI)

- Conduct pilot projects before full-scale implementation to test feasibility
- Use incremental investment approaches to reduce financial exposure
- Focus on AI use cases with clear ROI, such as automation and efficiency improvements.

Preventing rapid technological obsolescence

- Adopt modular and adaptable AI architectures to allow easy upgrades
- Stay updated on AI advancements and maintain flexibility in AI investments
- Utilize open-source models when possible, to reduce dependency on proprietary tech

Controlling frequent project cost overruns

- Set realistic budgets and continuously track project expenses
- Use cloud-based AI solutions to reduce infrastructure costs
- Implement agile project management to quickly adapt to challenges

Clarifying project scope and objectives

- Define clear AI goals, success metrics, and data requirements from the outset
- Involve cross-functional teams to align technical and business expectations
- Use prototyping and proof-of-concept models to validate feasibility

Reducing unexpected model failures and bias issues

- Conduct regular audits to detect and correct AI bias.
- Use diverse training data to improve AI fairness and robustness.
- Implement human oversight in AI-driven decision-making.

Ensuring compliance with regulatory and legal requirements

- Stay updated on AI laws and guidelines to ensure compliance.
- Implement privacy-preserving techniques such as encryption and differential privacy.
- Conduct legal and ethical reviews before AI deployment.

Overcoming integration challenges with existing systems

- Design AI with compatibility in mind for legacy systems.
- Use APIs and middleware solutions to ease integration challenges.
- Test AI models in real-world environments before full deployment.

Silent Dangers: Indirect AI Risks

The Samsung data leak: when AI became an accidental confidant

In early 2023, a few keystrokes sparked a significant corporate incident for Samsung. In an era where AI promises efficiency and innovation, the company found itself entangled in an unintended data breach—one not caused by hackers, but by its own employees.

The leaks: when convenience overpowered caution

The breaches weren't the work of cybercriminals but well-intentioned employees seeking AI-powered assistance. However, in their pursuit of efficiency, they unknowingly fed Samsung's proprietary secrets into OpenAI's ChatGPT—an external system beyond the company's control.

Source code exposure: In an effort to debug faulty semiconductor database code, an engineer uploaded proprietary source code to ChatGPT, inadvertently leaking critical intellectual property.

Strategic insights compromised: Another employee transcribed an internal meeting and input the notes into ChatGPT to generate structured minutes—handing over confidential discussions to an AI trained on external data.

Intellectual property at risk: A third instance saw an employee using ChatGPT to refine a test sequence for defective chips, unknowingly placing Samsung's semiconductor know-how into OpenAI's vast learning system.

The implications were immediate and severe. Since ChatGPT retains input data for model training, there was a real risk that Samsung's proprietary knowledge could be indirectly referenced in future AI responses, accessible to others in unpredictable ways.

Emergency countermeasures: Samsung scrambled to contain the damage, capping AI input sizes to a mere 1024 bytes per prompt before ultimately banning ChatGPT and similar tools outright.

Key sources of indirect AI risks

This incident at Samsung highlights that AI carries not only direct risks but also reveals the often-overlooked indirect risks posed by employees and third-party vendors who have no direct involvement in AI development or deployment.

AI-related risks go beyond dedicated AI vendors, reaching into the wider network of third-party service providers—such as legal, financial, and tax consultants—who may use AI in ways that unintentionally create compliance issues, security gaps, or ethical concerns. The lack of standardized AI governance across these entities heightens the potential for data mismanagement, regulatory misalignment, or algorithmic biases that could ripple through business operations.

Example:
Imagine a law firm handling an organization's confidential legal matters or a financial advisory firm managing its strategic planning. If these vendors deploy AI-driven tools—perhaps to automate contract analysis or financial forecasting—without stringent safeguards, the organization's sensitive data could be processed, stored, or even inferred in ways beyond its control.

An AI-powered legal assistant might inadvertently expose privileged case details through algorithmic vulnerabilities, or a financial model could feed proprietary market strategies into a third-party system, risking competitive leaks. The organization itself may have no direct role in these AI deployments, yet it remains accountable for the downstream consequences—whether it's a regulatory breach, reputational damage, or the erosion of client trust.

On the other and, employee adoption of BYOAI (Bring Your Own AI) introduces risks like data breaches, security flaws, ethical concerns, and regulatory non-compliance. Unvetted AI tools amplify these threats, jeopardizing system integrity and exposing organizations to legal and reputational damage, with risk levels varying by the employee's role and the sensitivity of the information they handle.

Example:
A secretary utilizing BYOAI to transcribe meeting minutes may inadvertently upload sensitive information, thereby exposing the company to significant data security breaches.
Conversely, a marketing professional leveraging AI for the creation of promotional materials would generally face lower risk exposure, as the data involved is typically less sensitive.

Risks associated with BYOAI

Data security and confidentiality risks

- Employees may input sensitive company data into external AI models, risking data leaks and regulatory violations.
- Many AI tools store or learn from user inputs, creating long-term exposure risks.
- ***Example:*** An employee pastes confidential customer data into an AI chatbot, violating GDPR and data privacy laws.

Compliance and regulatory violations

- External AI tools may not comply with industry-specific regulations (e.g., GDPR, HIPAA, SOC 2).
- Unauthorized AI usage can lead to legal repercussions if customer or financial data is processed improperly.
- ***Example:*** A healthcare employee uses an unapproved AI tool to analyze patient records, breaching HIPAA compliance.

Intellectual Property (IP) and ownership risks

- AI-generated content may lack clear ownership rights, leading to legal disputes over IP.
- Employees might unknowingly use AI tools that produce copyright-infringing material.
- ***Example****:* A marketing team generates AI-created images for an ad campaign, only to discover they infringe on an artist's work.

Inconsistent AI quality and reliability

- External AI tools vary in accuracy, leading to misleading insights or unreliable decision-making.
- AI-generated outputs can contain hallucinations or factual errors, impacting business strategies.
- ***Example:*** A financial analyst relies on AI-generated stock predictions, leading to poor investment decisions due to inaccurate data.

Shadow AI and lack of IT oversight

- Employees using unauthorized AI tools create "Shadow AI", bypassing IT security controls.
- IT teams lose visibility into AI usage, increasing risks of security breaches and compliance violations.
- ***Example****:* Employees use personal AI tools to draft reports, creating a security blind spot for IT teams.

AI bias and ethical risks

- AI models may reflect biases in training data, leading to unfair or discriminatory decisions.
- Companies risk reputational damage if AI-generated content violates ethical standards.
- ***Example****:* An HR team uses an external AI tool for hiring, unknowingly perpetuating gender or racial bias in candidate selection.

Performance and productivity risks

- Relying on unregulated AI tools may decrease accountability in decision-making
- If AI replaces critical thinking or human oversight, errors can go unnoticed.
- ***Example:*** Employees trust AI-generated legal contracts, missing important inaccuracies that create legal exposure

Risk mitigation

Thorough vendor selection and management

- Ensure that vendors meet high standards for ethical AI practices, data privacy, security, and compliance
- Perform due diligence by evaluating their history, certifications, and AI models and thorough technical, security, and compliance assessments of vendors
- Include contractual terms for performance metrics, security standards, compliance, and exit strategies. Embed specific requirements in contracts to address AI-related risks, such as data use and protection practices, IP etc.

Risk mitigation | BYOAI

- Develop a Comprehensive BYOAI Policy: Outline acceptable use, data security requirements, and compliance standards for personal AI tools
- Employee Training and Awareness - elevate employees' AI literacy in ethical use of AI, security best practices, and the risks associated with AI systems
- Conduct security assessments and compliance checks before approving personal AI tools for use
- Establish clear communication channels for reporting AI-related risks or incidents, both internally and with external stakeholders
- A targeted AI access strategy - identify high-risk employees, providing them with secure AI alternatives
    - Offer Enterprise-Grade AI Subscriptions – Provide company-approved AI tools with data privacy guarantees
    - Use On-Premise or Private LLMs – Deploy AI models internally to ensure data never leaves the organization
    - Enable AI Access Based on Security Clearance – Restrict public AI use for employees handling sensitive company information.

The Human side of AI:

Managing Social Impacts

AI presents notable social risks that affect various facets of society. A primary concern is job displacement and the growing economic inequality, as automation increasingly replaces jobs, particularly for lower-wage workers.

Additionally, it can widen the digital divide, creating unequal access to AI technologies and limiting opportunities for education, employment, and social mobility. Overreliance on AI may further diminish human influence and essential skills, leading to a loss of creativity and empathy in fields such as healthcare and education.

Social footprint of AI

Job displacement and economic inequality

AI-driven automation can lead to significant job displacement, particularly in sectors with repetitive tasks

- 14% of workers have experienced job displacement due to AI
- 85 million jobs are expected to be displaced by automation in 2025

Misinformation and social manipulation

Generative AI can create and disseminate misinformation, deepfakes, and hate speech, which can destabilize society and undermine trust in institutions

Example: AI deepfakes have been used in elections worldwide, including Bangladesh and Slovakia, to undermine candidates or create false narratives

Social surveillance

Potential for AI technologies to infringe upon individual privacy and civil liberties by enabling widespread monitoring and profiling of people

Example: China's use of facial recognition technology to monitor citizens' movements and activities

Increased criminal activity

Potential for AI technologies to be exploited by criminals for malicious purposes

- Over 3,500 new AI-generated criminal child sexual abuse images uploaded to a dark web forum

Loss of human influence and skills

Overreliance on AI can lead to diminished human skills and empathy, particularly in sectors such as healthcare and education

Widening digital divide

exacerbate existing inequalities, leaving marginalized groups without access to the opportunities, resources, and benefits that AI can provide

Risk mitigation

- **Reskilling and workforce transition programs:** to address job displacement, governments and organizations should invest in reskilling programs to help workers transition to new roles.
- **Human-centered AI:** AI should complement, not replace, human skills, supporting professionals while preserving their authority, creativity, empathy and expertise in fields such as healthcare and education.
- **Social impact assessments:** conducting regular social impact assessments of AI systems before deployment can help identify and mitigate potential risks to workers, communities, and societal structures.
- **Equitable access to AI:** bridging the digital divide involves ensuring that marginalized communities have access to the tools, skills, and infrastructure necessary to benefit from ai technologies.

AI’s Environmental Footprint:

Navigating the Green Risks

In Memphis, Tennessee, something massive is happening. Elon Musk’s xAI is building the world’s largest supercomputer, and it’s turning heads for more than just its sheer size. While the tech world buzzes about the new innovation, the locals are starting to ask some tough questions about what it’s costing the environment.

The supercomputer's thirst: A growing problem

From day one, xAI has been pulling a jaw-dropping 30,000 gallons of water a day from the Memphis Sand Aquifer, which is the city's main drinking water source. The problem? Memphis has been struggling with aging water infrastructure, and past contamination issues only make this even more concerning. If xAI keeps drawing this much water, will the city have enough for its residents? That’s the big worry.

Air quality concerns: Trouble in every breath

As if that weren’t enough, community groups are now raising alarms about the air. The supercomputer runs on gas turbines, and the emissions from these could seriously affect the air quality. There’s little oversight on this front, and growing concerns are stirring about the possible risks to public health. What happens when cutting-edge tech doesn’t just take from the earth, but starts polluting it, too?

Energy demand: the power-hungry beast

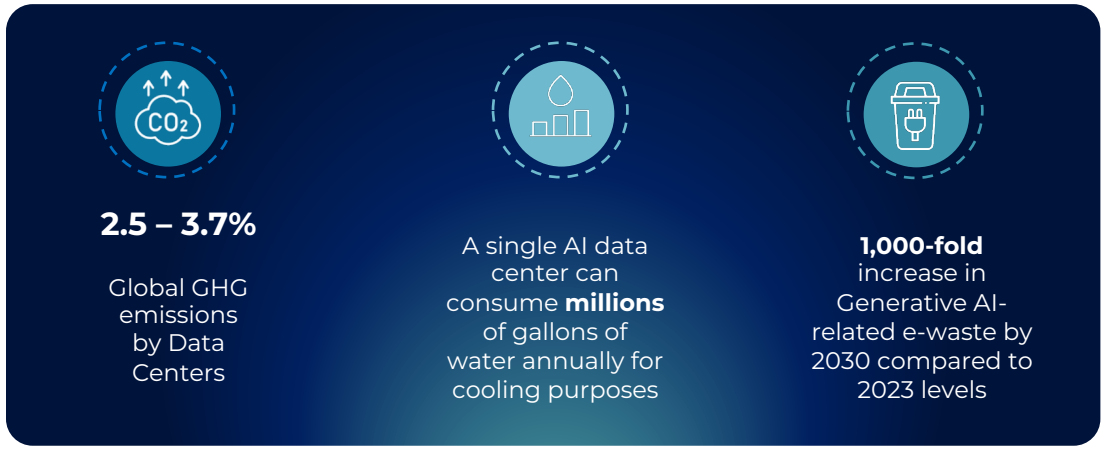
Then there’s the question of power. xAI was initially expected to use around 150 MW of electricity, but now they want to double that. As the supercomputer’s energy needs keep climbing, people are starting to wonder if the local grid can keep up. Advancing tech is exciting—but are our resources ready to match its pace?

Tesla’s Texas project: A mirror image of challenges

It doesn’t stop in Memphis. Down in Texas, Tesla’s Cortex project—designed to push the limits of AI for self-driving cars and robotics—is facing a similar dilemma. What started as an operation needing 130 MW of energy for cooling is now on track to exceed 500 MW in just 18 months. If renewable energy can’t pick up the slack, local power grids could be in trouble, and the carbon emissions might go through the roof.

Can innovation keep up with sustainability?

So here we are, watching these groundbreaking projects unfold, and the question is: at what cost? The tech world is on fire with excitement about what these supercomputers and AI projects can do, but the environment is taking the heat. Can we keep pushing innovation without pushing the planet too far? It’s a fine line, and with every new development, it feels like we’re walking it.



The environmental risks associated with AI are multifaceted and significant.

1. Energy consumption and emissions

AI systems, particularly those involving large-scale data centers, consume massive amounts of electricity, leading to substantial greenhouse gas emissions

GHG emissions

- Data centers account for 2.5 – 3.7% of global GHG emissions surpassing the carbon footprint of the aviation industry
- Training GPT-3 in 2022 resulted in 502 tons of CO2 emissions and consumed nearly 1,300 MWh of power, equivalent to the lifetime emissions of 8 cars or 109 cars' yearly emissions, enough to power an average U.S. home for 120 years.
- Microsoft emissions grew by 29% since 2020 due to the construction of more datacenters that are “designed and optimized to support AI workload

Energy consumption

- Global AI-related electricity consumption could rise by 85 to 134 TWh annually, equivalent to the annual electricity usage of countries such as Argentina, the Netherlands, and Sweden.

2. Water usage

Significant water consumption in data centers where cooling systems are essential to maintain optimal temperatures

- Training GPT-3 in Microsoft’s U.S. data centers consumes 700,000 liters of clean water, enough to produce 370 BMWs or 320 Tesla cars.
- Global AI demand is expected to account for 4.2-6.6 billion cubic meters of water withdrawal by 2027, surpassing the total annual water usage of Denmark or half of the UK

3. Resource extraction and waste

AI hardware production requires extracting rare earth minerals often through unsustainable mining, causing environmental harm.
AI-generated e-waste presents a significant environmental threat, as it contains hazardous chemicals such as lead, mercury, and cadmium, which can contaminate soil and water, posing risks to human health and the environment.

E-waste

- AI boom will increase the total amount of electronic trash generated globally by 3 – 12% by 2030
- Generative AI merely could account for up to 5 million metric tons of e-waste by 2030 (~8% of total e-waste)

Risk Mitigation

- **Sustainable infrastructure:** Deploy AI workloads on energy-efficient hardware and leverage green data centers powered by renewable energy.
- **Optimize cooling systems**: Improve cooling efficiency in data centers, using liquid cooling systems or more sustainable methods that reduce water and energy usage
- **Recycling and sustainable hardware**: Promote the recycling of e-waste and use hardware designed for longevity and energy efficiency

  Example: Studies have found out that implementing strategies to reduce, reuse, repair, and recycle out-of-date equipment from data centers could reduce e-waste generation by as much as 86%
- **Carbon offsetting**: Invest in carbon offset programs to neutralize the environmental impact of AI operations, such as investing in reforestation or renewable energy projects.
- **Increase model efficiency**: Designing AI models that require fewer computational resources (e.g. using model optimization techniques such as pruning, quantization) can significantly reduce energy consumption.

  Example: DeepSeek's technology utilizes fewer compute resources, significantly lowering its environmental impact. Initial estimates indicate up to 90% less energy consumption and approximately 92% lower carbon footprint compared to similar infrastructures, with a notable reduction in cooling requirements as well.
- **Lifecycle management:** Extend the lifespan of AI models through regular updates instead of frequent retraining from scratch.

Global AI Regulatory Landscape

The evolving AI regulatory landscape

AI is advancing rapidly, outpacing the regulatory frameworks designed to govern it. As novel risks—such as misinformation, deepfakes, and the ethical challenges of autonomous decision-making—emerge, existing laws, like those on privacy or intellectual property, struggle to keep up. These frameworks were not designed for such technologies, making enforcement fragmented, reactive, and difficult to scale. Moreover, issues like intent and responsibility remain legally ambiguous, exposing the limitations of current regulations.

As we progress toward more agentic forms of AI, collaboration between AI agents across different geographies will become increasingly common. However, for such collaboration to occur on equal footing, it is essential to establish common global regulations and principles. Without this, AI systems may struggle to work together effectively. Thus, unified regulations are not just beneficial, but a prerequisite for ensuring the interoperability of AI systems

In this context, governance should not be perceived as a constraint but rather as an enabler—one that empowers AI to mature responsibly and equitably. The establishment of universal principles and cross-border legal frameworks is vital, not merely to mitigate risk, but to cultivate trust, ensure alignment with societal values, and foster global collaboration. In doing so, governance becomes a catalyst for AI's sustainable growth, rather than a barrier to its progress.

These regulatory approaches fall into two broad categories: legally binding laws, which impose strict compliance requirements, and non-binding frameworks, which provide voluntary guidance.

Binding regulations are enforceable laws that mandate AI risk assessments, transparency measures, and accountability requirements. The European Union's AI Act is the most comprehensive, categorizing AI systems by risk level and imposing strict obligations, including documentation, human oversight, and penalties for non-compliance. China's AI laws prioritize national security, requiring AI companies to register with regulators and ensure content aligns with government policies. The U.S. has no federal AI law, but states like California are pushing for stricter AI transparency and bias mitigation rules.

In contrast, **non-binding frameworks offer best practices without legal enforcement.** The U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework provides voluntary guidelines for AI safety, focusing on bias, security, and trustworthiness. Similarly, the OECD AI Principles, adopted by 40+ countries, advocate for fairness, accountability, and human rights protections in AI development. Industry-led initiatives, such as the ISO/IEC 42001 AI Management System Standard, also provide a global framework for AI risk management, but adoption remains voluntary.

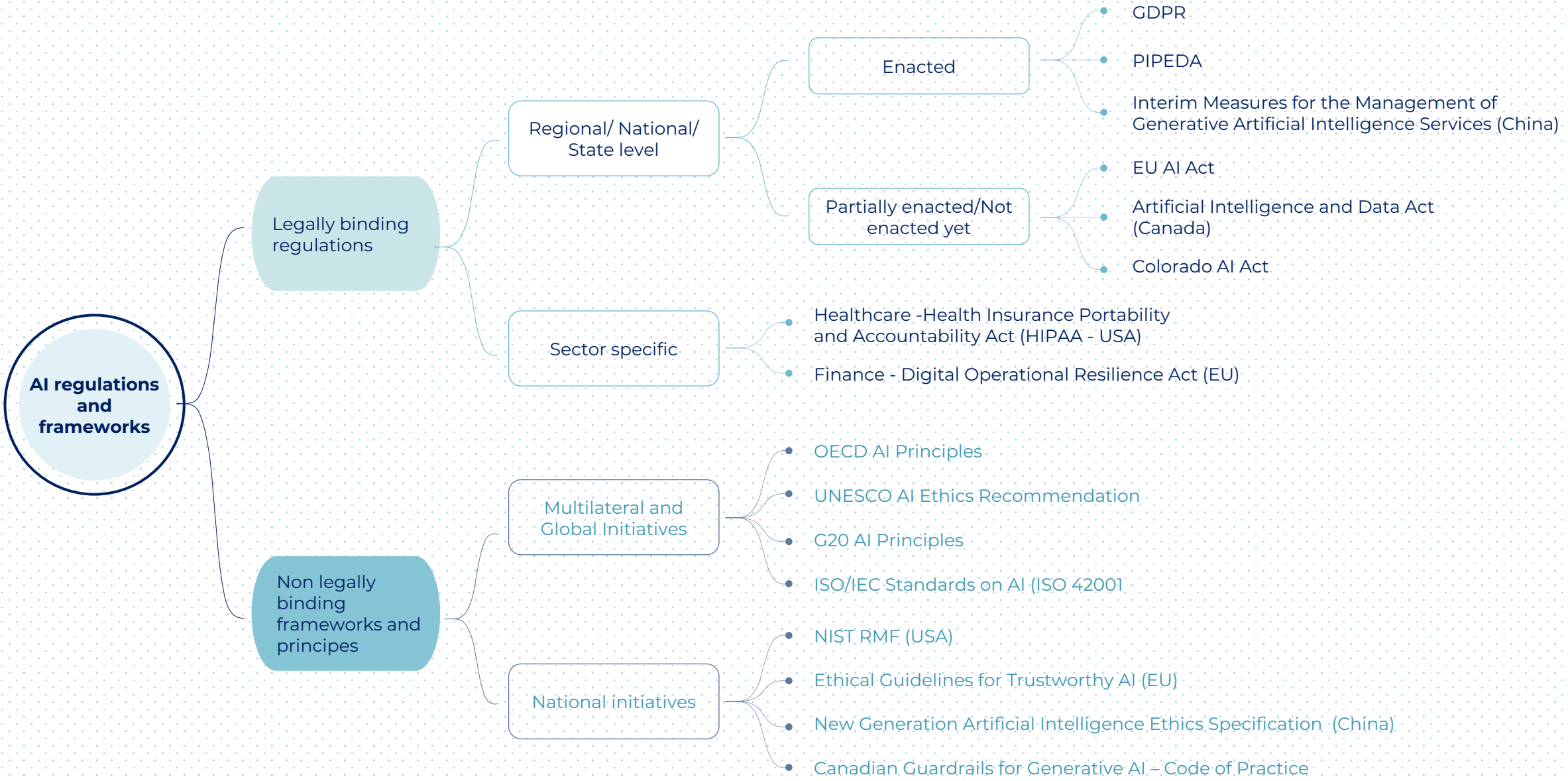
The differences in regulatory approaches across countries further complicate compliance. The European Union leads with stringent AI governance, emphasizing risk-based oversight and transparency. The United States, by contrast, largely resists AI governance, arguing that excessive regulation stifles innovation and puts American companies at a competitive disadvantage. AI has also become a geopolitical issue, with the U.S. prioritizing technological dominance over regulatory control. Meanwhile, Canada remains stalled, with its proposed AI regulations lacking clear direction and no cohesive framework for enforcement.

This regulatory divergence creates challenges for interoperability—the ability of AI systems to function across borders without conflicting legal risks. Without standardized governance, businesses must navigate multiple, and sometimes contradictory, compliance requirements. This lack of harmonization increases costs, legal uncertainty, and operational inefficiencies, particularly for multinational companies deploying AI globally.

Dealing with uncertainty

Given this landscape, organizations must adopt proactive governance frameworks that align with evolving regulations. Companies that embed compliance early—prioritizing transparency, risk mitigation, and ethical AI—will be better positioned to adapt as AI laws continue to develop**.** alignment remains a key challenge amidst geopolitical competition

The AI rulebook: A visual guide to laws, frameworks & principles shaping AI governance



Legally binding regulations

Legally binding regulations impose enforceable obligations on AI developers, deployers, and users. These laws are designed to ensure accountability, transparency, and ethical use of AI systems. These laws are implemented across multiple tiers—regional or country level, state-level within nations, and sector-specific—ensuring comprehensive coverage of diverse aspects as deemed necessary.

Regional/country level

Example:

- Europe – The EU AI Act is the world's most comprehensive AI regulation (refer to the page 83)
- China has enacted legally binding AI regulations covering data protection, cybersecurity, content labeling, and algorithmic transparency and accountability (refer to the page 96)
- Countries such as, Canada, Singapore, Japan, India and Australia primarily have regulations focused on data protection in AI systems, with no specific laws directly governing AI at this time.
    - Canada – Personal Information Protection and Electronic Documents Act (PIPEDA)
    - Singapore – Personal Data Protection Act (PDPA)
    - Japan – Act on the Protection of Personal Information (APPI)
    - India – Digital Personal Data Protection Act (DPDPA)
    - Australia – Privacy Act 1988

State level

In the absence of unified, comprehensive AI laws, some countries implement state or provincial-level regulations to govern AI development and deployment within specific regions, ensuring responsible practices.

- The US lacks a unified federal AI law but has state-level regulations such as California's AI Transparency Act, Colorado AI Act
- Canada has provincial laws regulating data protection rights and the responsible use of AI in the public sector, among other areas.

Sector level

AI regulations are increasingly being tailored to address the unique risks and challenges posed by AI in specific industries. These sector-specific regulations complement broader, cross-sectoral frameworks, ensuring that high-risk areas receive focused oversight.

Example:

- **Healthcare**: The FDA (USA) regulates AI-based Software as a Medical Device (SaMD) under its digital health framework
- **Financial Services**: EU's Digital Operational Resilience Act (DORA) emphasize the need for financial institutions to maintain operational resilience when using AI, ensuring that AI systems do not compromise overall business continuity

Non-binding principles and frameworks

Non-binding principles and frameworks are voluntary guidelines that encourage responsible AI development without imposing legal obligations. They are often seen as stepping-stones toward more robust regulatory measures

Example: NIST AI Risk Management Framework (US)

Google has built an AI and advanced technologies governance program that follows the NIST AI RMF approach. It uses the framework to balance optimizing beneficial AI use cases while addressing potential risks, such as ethical concerns, bias, and transparency issues.

The ideal regulatory framework would combine the strengths of both approaches—establishing enforceable rules for high-risk applications while leveraging voluntary guidelines to foster innovation in less critical areas. As nations like the EU set stricter precedents and others like the U.S. adopt mixed approaches, achieving global alignment remains a key challenge amidst geopolitical competition

AI governance frameworks at a global scale

Global AI non-binding frameworks are gaining widespread adoption as a critical mechanism for navigating the absence of unified, binding international regulations. These frameworks play an essential role in fostering robust AI governance, providing a flexible yet effective foundation for aligning diverse global practices and promoting ethical standards across borders.

Meanwhile, international collaboration on AI standards is expected to grow, with initiatives like the upcoming International AI Standards Summit in Seoul.

Example:

OECD AI Principles

First adopted in 2019 and updated in 2024, These principles serve as a global reference point for AI policymaking, facilitating international cooperation and policy interoperability. Its principles mainly focus on Inclusive growth, human rights and democratic values, transparency and explainability, robustness, security and safety, and accountability.

UNESCO AI Ethics Recommendation

Adopted in November 2021, this is the first global agreement on human-centric AI ethics. It provides a comprehensive framework to ensure that AI systems are developed and used in ways that respect human rights, promote human dignity, and support environmental sustainability. This represents a consensus among all 193 UNESCO Member States, providing a unified framework for ethical AI development.

ISO/IEC 42001

The world's first international standard specifically designed for Artificial Intelligence Management Systems (AIMS). t provides a comprehensive framework for organizations to establish, implement, maintain, and continually improve their AI systems, ensuring responsible development and use of AI technologies.

AI regulations and frameworks through time

2019

- OECD AI Principles
- Singapore Model AI Governance Framework
- G20 AI Principles

2021

- UNESCO AI Ethics Recommendation
- New Generation Artificial Intelligence Ethics Specification (China)

2022

- Digital Charter Implementation Act (Canada)
- Blueprint for an AI Bill of Rights (USA)
- Algorithmic Accountability Act (USA)
- Administrative Provisions on Algorithm Recommendation of Internet Information Services (China)
- American Data Privacy and Protection Act (USA)

2023

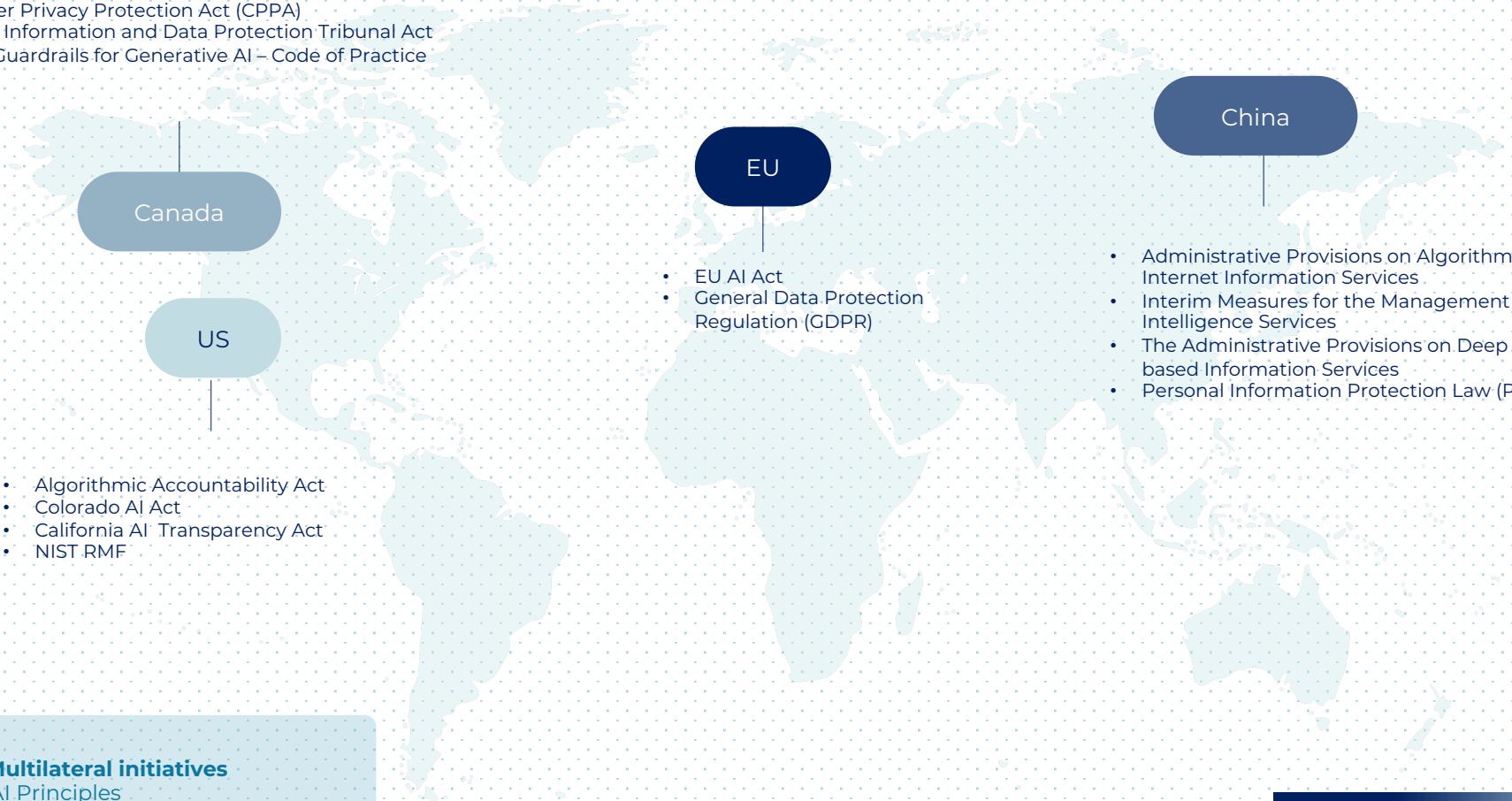
- The AI Research Innovation and Accountability Act (USA)
- Interim Measures for the Management of Generative Artificial Intelligence Services (China)
- The Administrative Provisions on Deep Synthesis in Internet-based Information Services (China)
- The Measures for Review of Scientific and Technological Ethics (China)
- AI Regulation White Paper (UK)
- Canadian Guardrails for Generative AI – Code of Practice

2024 & Beyond

- EU AI Act – provisions rolling out in stages from 2025 until 2027
- California AI Transparency Act – will take effect on January 1, 2026
- Colorado AI Act – will take effect on February 1, 2026

- Enacted law
- Implemented voluntary framework
- Only introduced, in the legislative process
- Signed into law and will be enacted in near term

The AI regulations atlas: How the world governs AI



Canada

- Digital Charter Implementation Act, 2022 (Bill C-27)
  - Artificial Intelligence and Data Act (AIDA)
  - Consumer Privacy Protection Act (CPPA)
  - Personal Information and Data Protection Tribunal Act
- Canadian Guardrails for Generative AI – Code of Practice
- PIPEDA

US

- Algorithmic Accountability Act
- Colorado AI Act
- California AI Transparency Act
- NIST RMF

EU

- EU AI Act
- General Data Protection Regulation (GDPR)

China

- Administrative Provisions on Algorithm Recommendation of Internet Information Services
- Interim Measures for the Management of Generative Artificial Intelligence Services
- The Administrative Provisions on Deep Synthesis in Internet-based Information Services
- Personal Information Protection Law (PIPL)

Global & Multilateral initiatives

- OECD AI Principles
- UNESCO AI Ethics Recommendation
- ISO/IEC 42001
- G20 AI Principles

Highly regulated — Least regulated

Regional AI oversight: Canada

Canada's AI regulations journey: Guardrails, setbacks, and uncharted futures

Bill C-27 – A Glimpse of the Future: Canada is on the cusp of something big with Bill C-27, paving the way for a future where AI isn't just powerful but purpose-driven.

AIDA at a Crossroads: Once poised for a 2025 or 2026 launch, now faces delays and uncertainty amid political hurdles and criticisms over its drafting and content

PIPEDA's Privacy Shield: A safeguard for personal data, ensuring AI systems handle it responsibly—yet with limited enforcement powers, its ability to uphold privacy rights remains constrained. Will CPPA finally bring the teeth PIPEDA lacks?

Treasury Board AI Policies: Guidance is mandatory, but enforcement remains to be seen

Provincial AI Patchwork: Without a federal AI law, provinces are attempting to create a fragmented regulatory landscape.

Voluntary Guardrails in Action: With AIDA pending, frameworks like Canadian Guardrails for Generative AI guide responsible AI development.

Global commitment

Canada is a signatory to the Global AI Treaty, signed in 2021, which imposes binding obligations on signatories regarding the use of AI by public authorities. The treaty requires countries to implement domestic laws addressing AI risks related to bias, discrimination, privacy, sustainability, and transparency throughout the AI lifecycle.. While it establishes these obligations for public sector use, it does not impose direct regulations on the private sector, allowing for more flexibility in corporate AI development. However, this means that while signing the treaty signals a commitment, actual enforcement and impact depend on national legislation

Canada's AI regulations are still evolving, with the Pan-Canadian AI Strategy (2017) being a foundational step in promoting ethical AI development and research. The country emphasizes privacy protection and data governance through the Personal Information Protection and Electronic Documents Act (PIPEDA), while focusing on fostering innovation in AI technologies. Although Canada lacks a comprehensive AI-specific regulatory framework like the EU AI Act, Bill C-27 marks a significant step toward strengthening its AI regulations..

Bill C-27 also known as the Digital Charter Implementation Act, 2022, is a significant piece of legislation in Canada aimed at modernizing privacy laws and regulating AI systems.

Components of Bill C-27:

- **Consumer Privacy Protection Act (CPPA):** Replaces Part 1 of PIPEDA, enhancing privacy rights and introducing new consent requirements.
- **Personal Information and Data Protection Tribunal Act (PIDPTA):** Establishes a tribunal to review decisions by the Privacy Commissioner and impose penalties for non-compliance.
- **Artificial Intelligence and Data Act (AIDA):** Represents Canada's first comprehensive attempt to regulate AI, focusing on high-impact systems and ensuring they are developed and deployed responsibly.

AIDA (Artificial Intelligence and Data Act)

Bill C-27, known as the Digital Charter, includes provisions for AIDA which aims to regulate high-risk AI systems in Canada. It establishes a framework for ensuring accountability, transparency, and fairness in AI development and deployment, focusing on mitigating risks such as bias, discrimination, and privacy violations

AIDA – what was the main objective?

AIDA strives to align with global AI regulations, particularly those of the EU and US, to strengthen trade, interoperability, and market access for Canadian companies. Its core mission is twofold: first, to empower Canadian businesses to confidently sell AI and AI-enabled products worldwide, ensuring they meet the highest international standards; and second, to safeguard Canadian companies and citizens from potential harm caused by unregulated AI products and services from jurisdictions lacking robust oversight.

AIDA – what is the scope?

AIDA is a risk-based system designed to regulate the potential harms of high-impact AI systems, particularly those used in critical sectors like law enforcement, finance, healthcare, and employment.

Key Aspects:

- **Risk classification –** determine whether an AI system is high-risk based on its function and impact
- **Transparency & accountability -** Requires companies to document how the AI system works, its data sources, and potential biases
- **Mitigation –** Companies should ensure that safeguards are in place to minimize harm, including human oversight and bias reduction
- **Compliance & reporting -** Organizations must conduct regular assessments and report on AI system performance and risks

But following questions are left answered

- **What is high impact**? How to measure it? – Does not explicitly define what is a high impact system
- **Who can measure it?** - It's unclear who determines impact levels—self-assessment by companies, regulators, or independent auditors?
- **How to mitigate?** - It also lacks details on what specific measures should be taken to mitigate risks
- **How to monitor?** – AIDA advices companies to regularly monitor high impact systems – but the frequency of that, who should monitor is not clearly detailed

In comparison, EU AI ACT has clearly defined risk categories and criteria (please see our comparison on slide 90)

What is stalling AIDA

Lack of details

AIDA lacks crucial specificity, particularly in defining "high-impact" AI systems, identifying who assesses them, and outlining clear methods to mitigate and monitor risks. This lack of detail creates ambiguity, making enforcement and compliance difficult**.**

Overly broad and rushed

Bill C-27, which bundles AIDA with various other legislative elements, including consumer protection, privacy (under the Privacy Protection and Electronic Documents Act), and cybersecurity (under the Critical Cyber Systems Protection Act), makes it overly broad and complex. The scope of these combined areas dilutes the focus of AIDA and complicates its implementation. Addressing AIDA separately might have ensured a more focused and effective framework.

The bill has insufficient clarity on compliance and oversight measures. Critics argue it lacks a comprehensive framework for enforceability, particularly regarding enforcement mechanisms and accountability.

The consultation period was considered too short, limiting stakeholder input. With over 100 amendments proposed, it suggests that further refinement is needed before it can move forward.

Resistance from big tech

Major tech firms—Google, Meta, Microsoft, and Amazon—have actively lobbied against AIDA, arguing it is too restrictive and burdensome.

These companies contend the law should assess AI applications based on their risk, rather than their industry or context. They argue that a one-size-fits-all approach is not effective for diverse AI applications across sectors.

They have pushed for AI oversight to be handled by sector-specific regulators—such as Health Canada for medical AI and the Office of the Superintendent of Financial Institutions for banking—rather than imposing broad AI-specific rules. This would allow for more tailored, specialized regulation.

Tech firms argue that companies might incur high costs to comply with AIDA, which could disproportionately burden smaller companies and startups, giving larger firms a competitive advantage.

Case study: How would AIDA work?

	Open-source	High Impact	Low Impact
Impact assessment	Researcher/company publishing an open-source AI model	Developing/deploying a **High Impact AI system** that significantly affects human health, safety, or rights e.g. An airport deploying an AI-powered facial recognition for identity verification at check-ins or boarding	Developing/deploying a **Low Impact AI system** that does not significantly affect human health, safety, or rights e.g. A travel planner using an AI based personalized travel recommendations
Liability	Not a commercial activity/not a complete system therefore, no regulatory requirements or liability under AIDA	***Compliance requirements;*** **AI Designers and Developers:** • Robust model validation, testing & retraining • Ensuring fairness, explainability and safety **AI Deployers:** • Monitoring the output and intervening as required • Human oversight Subject to penalties for non-compliance	Minimal/less stringent regulations could be expected

Bill C-27 in a nutshell

	Artificial Intelligence and Data Act (AIDA)	Consumer Privacy Protection Act (CPPA)	Personal Information and Data Protection Tribunal Act
Purpose	Regulates high-impact AI systems to ensure safety, transparency, and accountability.	Modernizes privacy laws by replacing Part 1 of PIPEDA, enhancing individual privacy rights and data protection.	Establishes a tribunal to enforce CPPA decisions and impose penalties for non-compliance.
Key provisions	• Risk assessments for high-impact AI • Prohibits harmful AI practices • AI and Data Commissioner oversight.	• Strengthened consent requirements • Data portability rights • Right to request deletion • Stronger enforcement powers.	• Independent tribunal with authority to review Privacy Commissioner's decisions. • Imposes administrative monetary penalties
Enforcement Mechanisms	Penalties include fines up to $25 million or 5% of global revenue for non-compliance with AIDA.	Privacy Commissioner has broader powers to investigate violations and recommend corrective actions.	Tribunal can impose fines up to $10 million or 3% of global revenue for CPPA violations.
Applicability	Targets developers and deployers of high-impact AI systems used in critical areas like healthcare or employment.	Applies to private-sector organizations handling personal data in commercial activities.	Focuses on organizations found in violation of CPPA provisions.
Global alignment	Similar to the EU AI Act, focusing on regulating high-risk AI systems globally.	Aligns with GDPR-like standards for data protection and privacy rights.	Not applicable

Personal Information Protection and Electronic Documents Act (PIPEDA)
PIPEDA ensures that personal information used in AI systems is collected, used, and disclosed responsibly, safeguarding individuals' privacy rights. It establishes a framework for organizations to balance innovation in AI with the protection of fundamental privacy rights.

Key provisions

- **Consent:** AI systems must comply with PIPEDA's requirement for meaningful consent when processing personal data.
- **Data Minimization:** AI developers must ensure they only collect and process the minimum amount of personal data required for their systems, reducing risks of over-collection or misuse.
- **Transparency in Automated Decision-Making**: While PIPEDA does not explicitly regulate AI, its principles require organizations to be transparent about the use of automated decision-making systems that affect individuals.
- **Data Breach Reporting:** Organizations must report breaches involving AI systems if they pose a "real risk of significant harm" to individuals and notify affected parties promptly.

Challenges for AI Compliance

- The lack of explicit provisions addressing AI-specific risks under PIPEDA creates ambiguity for developers. CPPA is expected to supersede the existing provisions of PIPEDA that relate to the protection of personal information in the private sector.

- **Example:** In 2020, Clearview AI was found to have violated PIPEDA by collecting and using biometric data from Canadians without their consent. The company's claim that publicly available data didn't require consent was rejected, as Canadian privacy authorities ruled that facial recognition use demands explicit permission. A joint investigation by federal and provincial privacy commissioners confirmed Clearview AI's violations, leading to an order to halt its services in Canada, stop collecting Canadians' images, and delete existing data. However, this case highlighted significant shortcomings in Canadian privacy laws, particularly PIPEDA's lack of order-making powers as Clearview AI was not imposed with any penalties.

In the absence of overarching federal AI legislation, the Treasury Board's Directive on Automated Decision-Making provides the primary guidance for how federal agencies should deploy AI. At the same time, provinces such as Quebec, Ontario, and Alberta are advancing their own regulatory approaches, while sector-specific regulators in areas like finance and health have begun issuing independent guidelines. This evolving mix of federal directives, provincial initiatives, and sectoral standards is contributing to a fragmented, province-by-province patchwork in Canada's AI governance landscape.

Treasury Board Policy Instruments: Directive on Automated Decision-Making (ADM)
The Treasury Board of Canada regulates federal agencies through a set of policy instruments, including directives that govern the use of artificial intelligence. The Directive on Automated Decision-Making (ADM) is the primary AI-related policy, aimed at ensuring transparency, accountability, and responsible AI use across federal institutions.

- Effective Date: April 1, 2019
- Mandatory Since: April 1, 2020
- Scope: Applies to departments and agencies subject to the Policy on Service and Digital that use or plan to use automated decision systems in administrative processes.

A central feature of the directive is the **Algorithmic Impact Assessment (AIA)** Tool, which helps agencies evaluate and mitigate AI-related risks:

- Assesses system complexity, human involvement, and potential harm
- Generates an impact level—ranging from 1 (little or no impact) to 4 (very high impact)—to determine the required level of oversight.
- Public disclosure of completed AIAs is mandatory for Levels 2–4

The directive also supports a pre-approved List of Interested AI Suppliers for responsible procurement.

Ontario Bill 194 (Strengthening Cyber Security and Building Trust in the Public Sector Act, 2024)
Ontario's public sector entities must now disclose information about their use of AI systems, take steps to manage associated risks, and establish an accountability framework for their AI usage.

Quebec's Law 25, formerly Bill 64
Quebec's Law 25, formerly Bill 64, is a landmark privacy law that introduces strict requirements for consent, transparency, and accountability—particularly around the use of AI. It applies to both public and private sector organizations that collect, use, or disclose personal information about individuals in Quebec, regardless of where the organization is based. The law mandates privacy impact assessments, transparency in automated decision-making, and mandatory data breach reporting.

Sector-Specific Regulatory Bodies in Canada
Regulatory bodies are making strides in overseeing AI use within their sectors.

- Law Societies (e.g., Ontario, Alberta, British Columbia)

Provincial law societies provide guidelines on the responsible use of generative AI in legal practice. Violations can lead to disciplinary actions such as fines or suspension of licenses.

- Office of the Superintendent of Financial Institutions (OSFI)

OSFI's Draft Guideline E-23 outlines model risk management for financial institutions using AI. It focuses on AI's role in decision-making, such as loan approvals and fraud detection.

Bills 33 and 34 - Protection of Privacy Act & Access to Information Act
Aim to strengthen privacy protections, enhance transparency, and modernize access to information in Alberta's public sector, aligning with contemporary technological and societal needs.

Meanwhile, voluntary AI frameworks are pivotal in shaping AI governance in Canada, setting ethical standards and promoting responsible development as formal regulations, such as the proposed Artificial Intelligence and Data Act (AIDA), are being finalized.

Canadian Guardrails for Generative AI – Code of Practice
A voluntary framework developed by the Government of Canada to guide the development, deployment, and operation of generative AI systems. It aims to ensure these systems are safe, fair, transparent, and accountable while preparing developers and operators for future regulatory compliance under the Artificial Intelligence and Data Act (AIDA)

Canada at a crossroads: The future of AI regulation

Canada's AI regulation remains in limbo, stalled by extensive amendments, political disagreements, and mounting criticism. Bill C-27, which includes the Artificial Intelligence and Data Act (AIDA), has faced scrutiny over its vague enforcement mechanisms and lack of clarity on compliance obligations. With a decisive federal election in April, the future of AI governance in Canada hinges on shifting political dynamics.

A key question is whether Canada will align with Europe's stringent AI regulations or adopt the U.S. approach, which prioritizes innovation over strict oversight. The European model emphasizes transparency, risk management, and accountability but could isolate Canada from U.S. companies reluctant to comply with stricter rules. On the other hand, aligning with the U.S. stance—favoring industry self-regulation—may expose Canada to risks such as biased AI systems, data misuse, and weakened consumer protections.

Adding to the complexity, Big Tech is expected to push back against stringent regulations, lobbying for a more flexible framework. Meanwhile, party politics will play a decisive role, with different parties likely to take divergent positions on the scope and necessity of AI governance.

As the election draws near, Canada stands at a crossroads: decisions made in the coming months will define its regulatory posture and international alignment in the global AI economy.

In the meantime, in the absence of comprehensive federal legislation, Canada relies on Treasury Board policy instruments, regional laws like Quebec's Law 25, and sector-specific guidelines to form a patchwork safety net for AI governance.

Regional AI oversight: EU

Navigating the EU AI Act: What You Need to Know

Geopolitical Chess Move
The world's first comprehensive AI law—will others follow suit, or will the EU become an AI regulation island?

Risk-First, Innovation-Second?
AI is classified into four risk tiers: Unacceptable, High, Limited and Minimal—but does this strangle cutting-edge AI before it breathes?

Accountability in the AI Supply Chain
Not just developers—deployers & other actors share legal responsibility

No Passport Needed: The EU's AI Law Knows No Borders
It holds companies accountable regardless of where they are based, if they place, use, or distribute AI in the EU

Fines That Bite
Up to €35M or 7% of global revenue for violations—higher than GDPR. AI governance just got a serious price tag.

The EU AI Act is a bold first step, but is it a guiding light for ethical AI or a bureaucratic anchor slowing down AI progress? Let's dive in to explore!

01. EU AI Act

The EU AI Act is a pioneering legal framework designed to ensure the safe and responsible development and deployment of AI across the European Union. This Act represents a significant step in regulating AI globally, aiming to foster trust in AI technologies while protecting fundamental rights and societal values.

Key features of the EU AI Act

- **Risk-Based Approach:** The Act categorizes AI systems into four risk levels—unacceptable, high, limited, and minimal risk—imposing stricter regulations on high-risk applications. This approach ensures that AI systems are developed and used in ways that mitigate potential harm to individuals and society
- **Applicability:** Obligations can be applied to six categories of economic actors: providers, importers, distributors, product manufacturers, authorized representatives, and deployers. The AI Act comes into effect when an AI system or general-purpose AI model is placed on the EU market, put into service within the EU, imported into, or distributed within the EU. It also applies when an AI system is used by a deployer whose place of establishment is in the EU or who operates within the EU.
- **Prohibited AI Practices**: The Act bans AI systems deemed unacceptable due to their potential to infringe on fundamental rights, such as social scoring and certain forms of biometric identification.
- **Protection of Human Rights:** The Act aims to ensure that AI systems respect fundamental rights and EU values. It requires high-risk AI systems to undergo rigorous assessments, have robust documentation, and ensure human oversight.

- **Transparency and Accountability**: By mandating transparency in AI decision-making processes and requiring human oversight, the Act promotes accountability and trust in AI-driven outcomes. It emphasizes shared accountability across the AI supply chain, ensuring that not only developers (providers) but also deployers, importers, distributors, and other actors bear legal responsibility for AI systems. This is particularly crucial in high-risk areas such as healthcare, finance, and law enforcement

- **Global Leadership**: The EU AI Act sets a precedent for AI governance worldwide, influencing other regions to adopt similar frameworks. Its emphasis on ethical AI use is expected to shape global standards for AI development and deployment

Outlook and Future Developments

- **Implementation Timeline**: The Act's provisions are being implemented in phases, with key regulations becoming effective in 2025 and beyond. This phased approach allows businesses to adapt gradually to the new regulatory environment

- **AI Office and Code of Practice**: The European Commission is establishing an AI Office to oversee the implementation of the Act and is developing a Code of Practice to provide detailed guidance on compliance

- **Stakeholder Engagement:** The AI Pact initiative encourages voluntary compliance with the Act's key obligations, fostering collaboration among stakeholders to ensure a smooth transition to the new regulatory framework

Criticisms

- **Overregulation and Innovation:** Industry representatives worry that the Act could lead to overregulation, stifling innovation and driving AI businesses to more lenient regulatory environments. This might diminish the EU's competitiveness in the AI sector

- **Export of AI Systems:** The Act does not prevent EU-based companies from exporting AI systems that are banned in the EU, potentially violating human rights in non-EU countries

Europe's AI Continent Action Plan: Catalyzing a bold future for AI

The most recent milestone in the EU's evolving AI regulatory and governance landscape is the launch of the AI Continent Action Plan in April 2025—an ambitious initiative aimed at establishing Europe as a global frontrunner in artificial intelligence. This strategic move comes at a critical juncture, as the bloc faces mounting criticism from industry leaders who argue that its existing regulatory environment—from stringent AI rules to complex tax policies—is stifling innovation and hindering the growth of startups. Concerns have also been echoed by the U.S. administration, which has voiced unease over Europe's treatment of American tech giants and the increasing barriers faced by emerging AI enterprises.

Against this backdrop, the EU's action plan is widely viewed as a recalibration effort—seeking not only to assert global leadership but also to foster a more enabling landscape for innovation, investment, and cross-border collaboration. Building on the €200 billion InvestAI initiative, it lays the groundwork for global leadership in AI by advancing five strategic pillars:

1. Powering AI with world-class infrastructure – The plan envisions a network of at least 13 AI factories and up to 5 gigafactories—state-of-the-art hubs equipped with supercomputing power and vast data centers to fuel the development of frontier AI models. Supported by the InvestAI facility and a proposed Cloud & AI Development Act, these initiatives will triple Europe's data capacity while prioritizing sustainability.

2. Unlocking the value of data – To fuel innovation, the EU will foster a unified data market via the upcoming Data Union Strategy. AI factories will house specialized data labs to collect, organize, and share high-quality datasets, empowering developers across Europe.

3. Accelerating AI in strategic sectors – With only 13.5% of EU firms currently using AI, the new Apply AI Strategy will supercharge adoption—integrating AI into industries, the public sector, and healthcare, with AI factories and innovation hubs at the core of deployment.

4. Building a skilled AI workforce – Europe will attract global talent through legal migration pathways, retain top minds, and launch education programs tailored to AI and generative AI—ensuring a future-ready workforce with the skills to lead.

5. Supporting responsible AI deployment – To simplify compliance with the EU AI Act, the Commission is offering practical tools—guidelines, codes of practice, and soon, an AI Act Service Desk—to help businesses navigate the regulatory landscape with confidence and clarity.

Elements of hype and criticism

Dependency and feasibility concerns: Critics argue that the plan is heavy on rhetoric and light on actionable, immediate impact. The EU still relies heavily on non-EU infrastructure and technology, raising questions about whether it can achieve true technological sovereignty.

Imitation, not innovation: Some observers note the plan's similarity to strategies already pursued by global AI leaders like OpenAI and the US, suggesting the EU is playing catch-up rather than leading the field.

Sustainability and environmental impact: There are concerns about the environmental cost of massively scaling up data centers and supercomputing facilities, potentially increasing fossil fuel use in the name of digital leadership

EU AI Act: A risk-based approach

	Unacceptable risk	High risk	Limited risk	Minimal or no risk
Nature of risk	AI systems that pose a significant threat to human safety, livelihoods, or rights.	AI systems that significantly impact safety or fundamental rights	AI systems that interact with individuals but pose minimal risks.	AI systems with little to no impact on individuals' rights..
Examples	Subliminal Manipulation, Exploitation of Vulnerabilities , Social Scoring Systems etc.	Critical infrastructure, healthcare, education, employment, law enforcement.	Chatbots, deepfakes	AI-powered video games, spam filters
Regulatory requirements	Prohibited except for specific law enforcement exemptions; heavy penalties for non-compliance	Strict requirements, including risk assessments and mitigation systems, data quality, transparency, robust documentation, human oversight	Transparency obligations; users must be aware they are interacting with AI	Generally unregulated, but other laws may apply
Penalties for Non-Compliance*	Up to €35 million or 7% of the company's total worldwide annual turnover for the preceding financial year, whichever is higher	≤ 3% of total worldwide annual turnover in preceding year; or ≤ €15 million	Does not explicitly outline specific penalties for non-compliance with AI systems classified as "limited risk	No penalties for non-compliance due to lack of regulation

*General Non-Compliance: Fines for incorrect information or other violations can reach up to €7.5 million or 1% of annual turnover

02. General Data Protection Regulation (GDPR)

GDPR has a profound impact on AI development and use. While it was not designed specifically for AI, its principles and provisions significantly influence how AI systems handle personal data. It plays a crucial role in shaping the ethical and legal landscape of AI by emphasizing transparency, accountability, and individual rights.

Key provisions

- **Consent**: AI systems must obtain explicit, specific, informed, and unequivocal consent for processing personal data
- **Purpose Limitation**: Data collected for one purpose cannot be used for another without additional consent
- **Transparency**: Users must be informed about how their data is used in AI-driven decisions
- **Explainability**: Individuals have the right to understand the logic behind automated decision-making processes
- **Storage Limitation**: Store personal data for no longer than necessary for the purpose of processing. The retention period must be justified and documented
- **Breach Notification**: Organizations must notify the relevant supervisory authority within 72 hours of becoming aware of a data breach
- **Individual Rights**: grants individual rights such as access, rectification, erasure, and portability of their data, which AI systems must respect

Applied on	• Mandatory for all organizations that collect or process personal data of individuals in the EU and EEA • Companies not based in the EU but offers products or services to EU citizens or residents or monitors their behavior.
Non-compliance	• Non-compliance may result in administrative fines of up to €20 million or up to 4 % of the total worldwide annual turnover of the previous financial year, whichever is higher.
Example	**OpenAI's €15 Million GDPR fine imposed by the Italian Data Protection Authority (Garante)** • OpenAI was found to have used personal data to train ChatGPT without an adequate legal basis, such as consent or legitimate interest • The company was criticized for not providing clear information to users about how their data was used, violating GDPR's transparency obligations • In March 2023, ChatGPT experienced a data breach exposing chat histories and partial payment details. OpenAI failed to notify the Garante about this breach, violating GDPR's breach notification requirements • OpenAI did not implement adequate age verification measures, risking exposure of children under 13 to inappropriate content

Annexure | EU AI Act

01. The regulated operators under the AI Act

- **AI providers**: Natural or legal persons, public authorities, agencies, or other bodies that develop or have developed an artificial intelligence (AI) system or general-purpose AI model that is placed on the EU market. This applies to both free and paid AI systems.
- **Product manufacturers**: Companies incorporating AI systems into products sold under their name or trademark in the EU.
- **Deployers (Users):** A natural or legal person, public authority, agency, or other body using an AI system, except if the system is being used for a personal, nonprofessional activity. (The majority of organizations subject to the EU AI Act will be deployers.)
- **Importers**: Entities bringing AI systems developed outside the EU into the EU market.
- **Distributors**: Natural or legal persons in the supply chain—other than the provider or importer—who make an AI system available in the EU market.
- **Authorized representative**: An EU-established natural or legal person appointed by a provider established outside the EU to act as their authorized representative

02. Unacceptable risk: prohibited practices

- **Harmful AI-based manipulation and deception**: AI systems that intentionally manipulate individuals into making harmful choices they otherwise wouldn't make.
- **Harmful AI-based exploitation of vulnerabilities**: Systems exploiting vulnerabilities due to age, disability, or social or economic situations.
- **Social scoring**: AI systems used to evaluate or predict aspects of a person's life based on their behavior or characteristics.
- **Individual criminal offence risk assessment or prediction**: AI systems predicting criminal behavior based on personal data.
- **Untargeted scraping of the internet or CCTV material to create or expand facial recognition databases:** Unauthorized collection of facial data for surveillance purposes.
- **Emotion recognition in workplaces and education institutions**: AI systems used to infer emotions in these settings, except for medical or safety purposes.
- **Biometric categorization to deduce certain protected characteristics**: AI systems categorizing individuals based on biometric data to infer race, political opinions, etc.
- **Real-time remote biometric identification** for law enforcement purposes in publicly accessible spaces: AI systems used for real-time surveillance in public areas

Annexure | EU AI Act

03. High-risk use-cases

- AI safety components in **critical infrastructures** (e.g. transport), the failure of which could put the life and health of citizens at risk
- AI solutions used in **education institutions**, that may determine the access to education and course of someone's professional life (e.g. scoring of exams)
- AI-based **safety components of products** (e.g. AI application in robot-assisted surgery)
- AI tools for **employment**, management of workers and access to self-employment (e.g. CV-sorting software for recruitment)
- Certain AI use-cases utilized to give access to **essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan)
- AI systems used for **remote biometric identification, emotion recognition and biometric categorization** (e.g AI system to retroactively identify a shoplifter)
- AI use-cases in **law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence)
- AI use-cases in **migration, asylum and border control management** (e.g. automated examination of visa applications)
- AI solutions used in **the administration of justice and democratic processes** (e.g. AI solutions to prepare court rulings)

04. Transparency obligations

- AI systems designed to interact directly with natural persons;
- AI systems, including general-purpose AI systems, that generate synthetic audio, image, video or text content;
- Emotion recognition or biometric categorization systems;
- AI systems that generate or manipulate images, audio or video that are deepfakes.

05. Requirements for high-risk AI systems

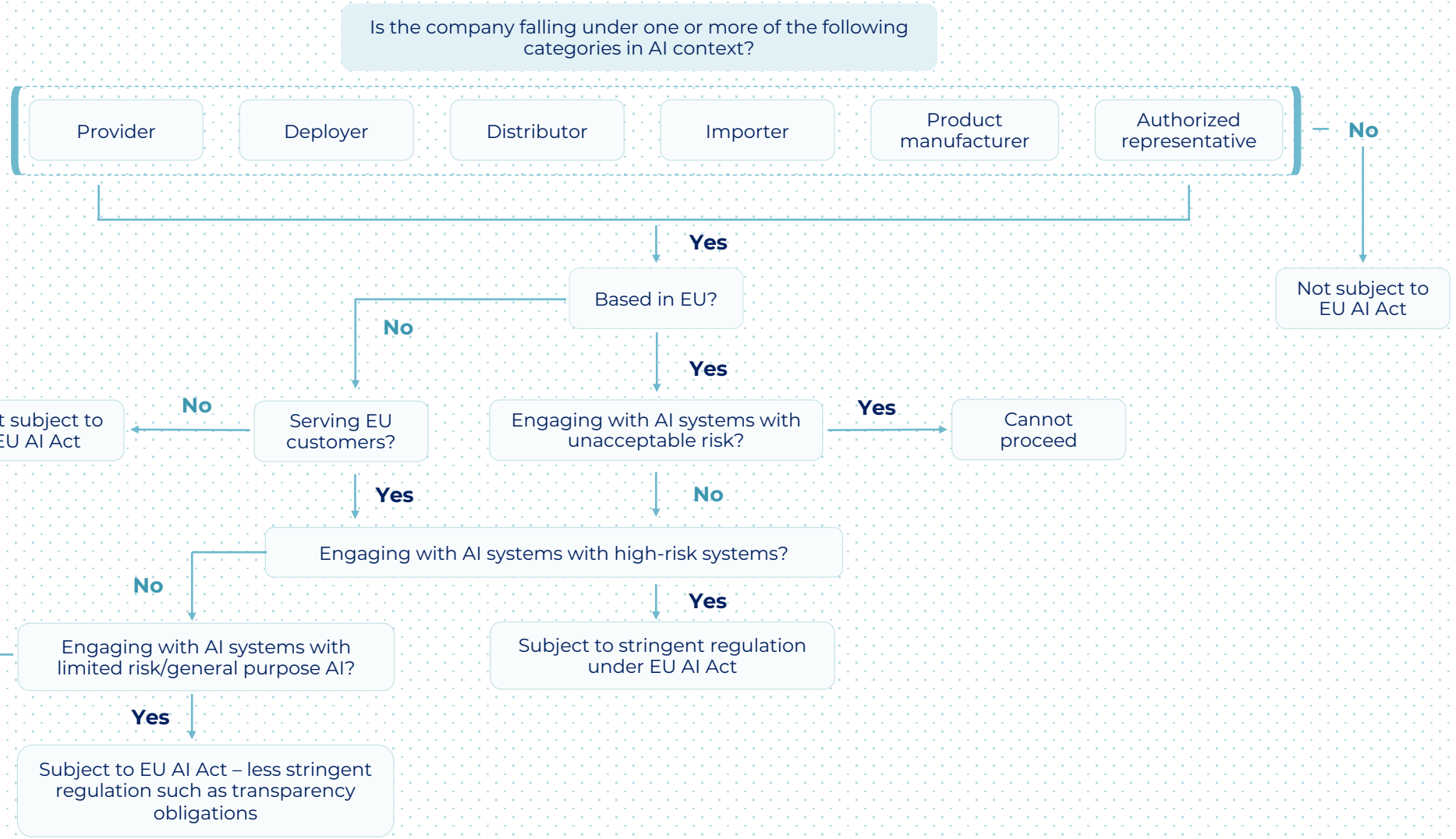
- **Compliance with the requirements**: technical and organizational requirements throughout their life cycle
- **Risk management**: ongoing process to identify, analyze, and mitigate foreseeable risks, including designing risk reduction measures, implementing controls, and providing user information and training
- **Data governance:** High-quality, representative, and to the best extent possible error-free and complete training, validation, and testing datasets are required to ensure proper functioning and safety of the system.
- **Technical documentation and record keeping**: It should cover all aspects of the system, including its characteristics, algorithms, data, training, testing, validation, and risk management.
- **Transparency and provision of information**: The system's decision making must be understandable, and details on its identity, characteristics, limitations, purpose, accuracy, risks, capabilities, oversight, maintenance, and expected lifespan must be provided.
- **Human oversight**: must be proportionate to the system's risks and level of autonomy. Human operators should also be able to override the system if necessary.
- **Accuracy, robustness and cybersecuri**ty: Accuracy measures include minimizing prediction errors, robustness measures ensure systems can handle errors and inconsistencies. Lastly, cybersecurity measures shall protect against unauthorized system alterations in which case compliance can be demonstrated through the EU Cyber Resilience Act for relevant AI systems subject to the EU Cyber Resilience Act.

A comparative look at the EU AI Act and AIDA Canada

Parameter	EU AI Act	AIDA Canada
Regulatory Approach	Establishes 04 distinct categories of AI risks: Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk.	Does not explicitly categorize AI systems into risk levels. Instead, focuses on 'high impact' AI systems
Proportional Obligations and Disclosures	Obligations are proportional to the risk level. High-risk systems require conformity assessments, human oversight, and detailed documentation	Requires transparency and accountability for high-impact systems. The specifics of these obligations will be defined in regulations
Defined Responsibilities for Risk Assessment	Developers and deployers of high-risk AI systems are responsible for conducting risk assessments and ensuring compliance with regulatory requirements	Places similar responsibilities on developers and deployers of high-impact AI systems, though the specifics are less detailed
Risk Mitigation and Continuous Monitoring	High-risk systems must implement robust risk mitigation measures and undergo continuous monitoring to ensure compliance and safety	Requires risk mitigation and monitoring for high-impact systems, though the details are less prescriptive compared to the EU AI Act
Enforcement and Consequences of Non-Compliance	Non-compliance can result in fines up to €30 million or 6% of global turnover for serious offenses	Imposes penalties up to CAD 25 million or 5% of gross global revenues for serious offenses

Case study: How EU AI Act applied on a Canadian company

Extraterritorial scope of the EU AI Act states that it applies to AI providers and deployers, regardless of their location, if their systems are placed on the EU market or used within the EU



Regional AI oversight: US

Can innovation thrive without federal oversight?

Trump's Deregulation Push: Reducing Red Tape, Trump's administration set the stage for AI growth by targeting barriers that could stifle innovation in AI and tech.

Fragmented Regulatory Landscape: The U.S. lacks a unified AI framework, leaving regulations scattered across state laws, industry standards, and ad-hoc initiatives.

Innovation vs. Oversight: While prioritizing innovation, the U.S. leans toward minimal regulation to maintain leadership in AI

State-Level Legislation: States like California, Colorado, New York have taken matters into their own hands with AI-specific laws, creating a patchwork of rules that vary across the country.

Trump's AI policy shift: Innovation at the expense of oversight?

Donald Trump's second-term revocation of Biden-era AI regulations marks a decisive pivot in U.S. policy—from cautious governance to aggressive innovation. On January 20, 2025, he rescinded Biden's 2023 Executive Order on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" that had mandated AI safety disclosures and risk mitigation protocols. In its place, a new directive—"Removing Barriers to American Leadership in Artificial Intelligence", which focused on minimizing regulatory barriers to foster AI innovation. Central to this agenda was a $500 billion infrastructure initiative, forged in collaboration with major tech players such as Oracle and OpenAI..

The administration's message was clear: reduce red tape, sideline "ideological bias," and fast-track U.S. dominance in AI. Federal agencies were ordered to dismantle perceived regulatory hurdles, echoing a broader pro-business, anti-regulation philosophy.

While this deregulation may accelerate innovation and economic growth, critics warn of growing risks—unchecked algorithmic bias, privacy breaches, and cybersecurity threats. The absence of federal safeguards could fracture governance, especially as states like California move to fill the regulatory vacuum. The future of institutions like the U.S. AI Safety Institute remains unclear, potentially leaving gaps in governance at the federal level.

Meanwhile, the EU's stringent AI Act highlights a widening global rift in oversight, potentially undermining trust in U.S.-built systems. However, lawmakers behind the EU's AI Act are concerned that lobbying from U.S. tech giants and pressure from the Trump administration may lead the 27-nation bloc to dilute key aspects of the landmark regulation.

Navigating the AI patchwork: States, sectors, and soft rules

Unlike the EU's centralized regulatory framework (e.g., the EU AI Act), the U.S. relies on a decentralized, sector-specific, and principle-based approach to balance innovation and risk management. In the US, AI regulations are primarily driven by state-level initiatives and voluntary frameworks due to the absence of comprehensive federal legislation.

State-Level Legislation:

- **Colorado AI Act**: A landmark regulation is designed to protect consumers, specifically Colorado residents, from algorithmic discrimination when AI systems are used to make consequential decisions that significantly affect them.
- **California AI Transparency Act:** A comprehensive legislation aimed at enhancing transparency and accountability in the use of AI systems within California

- **New York City's Local Law 144**: Aims to ensure fairness, transparency, and accountability in the use of Automated Employment Decision Tools (AEDTs) for hiring and promotion decisions. It regulates how AI is used to prevent discrimination based on gender, race, or ethnicity when evaluating NYC residents regardless the employer's location.

- **California Consumer Privacy Act (CCPA):** A landmark state law in the US that enhances privacy rights and consumer protection for California residents

Alongside state-level mandates, sector-specific regulations construct a framework that guides AI governance and its industry-wide impact

- **HIPAA in healthcare sector:** Sets the standard for protecting medical information in the United States, ensuring the confidentiality, integrity, and accessibility of all electronic protected health information (ePHI) handled by covered entities.

- **Fair Lending Laws – e.g. Equal Credit Opportunity Act:** AI models used in credit scoring, lending, and mortgage decisions must comply with these laws, which prohibit discrimination based on race, gender, and other protected classes.

Non-binding/Voluntary frameworks

- **Blueprint for an AI Bill of Rights (2022):** A nonbinding framework designed to guide the development, deployment, and use of AI systems in ways that protect civil rights, civil liberties, and democratic values.

- **Congressional Initiatives – e.g. SAFE Innovation AI Framework:** The SAFE Innovation AI Framework in the U.S. is a policy framework introduced by Senate Majority Leader Chuck Schumer to guide the development of AI legislation. It emphasizes balancing AI's societal benefits with its potential risks

- **NIST AI Risk Management Framework (AI RMF)**

It is a structured, voluntary framework designed to help organizations develop, deploy, and manage AI systems responsibly. Created by the National Institute of Standards and Technology (NIST), the framework provides guidance on identifying, assessing, and mitigating risks associated with AI technologies, ensuring they are safe, fair, transparent, and trustworthy.

The framework is built on a risk-based approach, acknowledging that AI systems can pose various challenges, including bias, security vulnerabilities, and lack of explainability. By following the AI RMF, organizations—whether in the private sector, government, or academia—can make informed decisions about AI adoption, balancing innovation with ethical and regulatory considerations.

It consists of four key functions:
- Govern – Establish policies and practices for AI risk management.
- Map – Identify AI risks and how they relate to business objectives.
- Measure – Assess AI system risks using quantitative and qualitative methods.
- Manage – Continuously monitor and reduce AI risks over time.

In the U.S. AI landscape, where AI regulation is still evolving, the NIST AI RMF serves as a foundational guideline, helping organizations align with emerging policies, ethical standards, and global best practices. It plays a critical role in shaping responsible AI adoption, ensuring AI-driven innovations contribute positively to businesses, society, and national security.

US AI regulations: State level

	Overview	Objectives	Application	Key provisions
Colorado AI Act	• Signed into law on May 17, 2024, is the first comprehensive state-level legislation in the U.S. regulating the development and deployment of AI systems	• To promote transparency, accountability, and fairness while protecting consumers from algorithmic discrimination	• Applies to developers (those who create or substantially modify AI systems) and deployers (those who use such systems) conducting business in Colorado • Primarily targets high-risk AI systems, defined as those that make or significantly influence "consequential decisions" affecting areas	• Deployers must implement AI risk management programs and conduct recurring impact assessments to identify and mitigate risks of algorithmic discrimination • Deployers must disclose the use of high-risk AI systems to consumers and provide explanations for adverse decisions made by these systems
California AI Transparency Act	• The law will take effect on January 1, 2026, and is considered one of the most comprehensive AI-specific transparency laws in the United States.	• To ensure transparency and accountability in the use of generative AI systems	• Applies to "Covered Providers," defined as entities that create or operate generative AI systems with over 1 million monthly users that are publicly accessible in California. • Focuses on generative AI systems, which produce synthetic content such as text, images, videos, or audio.	• Providers must make available an AI detection tool at no cost to users, enabling them to assess whether content was generated or altered by AI. • Providers must offer users the option to include manifest disclosures in AI-generated content. • AI-generated content must include latent disclosures in its metadata.

	Overview	Objectives	Application	Key provisions

California Consumer Privacy Act	• Signed into law on June 28, 2018.
	• It became effective on January 1, 2020
		• To enhance consumer privacy rights and protection for California residents	Applies to for-profit entities doing business in California that meet specific thresholds:
			• Annual gross revenues exceeding $25 million
			• Buying, selling, or sharing the personal information of more than 100,000 consumers annually
			• Deriving more than 50% of annual revenue from selling or sharing personal information
				• Consumer rights including right to know, right to access, right to delete and so on
				• Business obligations:
				  • Provide clear privacy notices at collection and maintain a comprehensive privacy policy
				  • Implement reasonable security measures for consumer data
				• Penalties for non-compliance

New York City's Local Law 144	• Signed into law in 2021 and became effective on January 1, 2023	• To curb AI bias in employment decisions and enhance transparency in the use of AI-driven hiring tools	• Applies to employers and employment agencies using AEDTs to evaluate candidates for employment or employees for promotion within New York City • Includes employers outside of NYC if they use AEDTs to evaluate candidates for jobs located in NYC	• Employers must conduct an independent bias audit annually assess potential disparate impacts and publish results. • Employers must provide at least 10 business days' notice to candidates before using an AEDT • Violations can result in civil penalties, starting at $375 for the first violation

Regional AI oversight: China

China's road to a unified AI regulatory system

Innovation under watchful eyes – AI development thrives, but within strict oversight to align with national security priorities.

From fragmented to unified – A comprehensive AI law is in the works (2024–2025) to replace China's current patchwork of AI regulations.

Regional regulations – With local laws like those in Shanghai and Shenzhen , China creates a patchwork of AI rules

China is shaping the global AI landscape through a comprehensive and evolving regulatory framework that aims to ensure ethical development, national security, and technological leadership. The Chinese government has introduced a series of AI-specific regulations and guidelines aimed at fostering innovation while maintaining strict control over data, algorithms, and AI's societal impact.

Current state of Chinese AI regulations

- **Comprehensive and early implementation**: China has been proactive in regulating AI, focusing on areas like algorithmic governance, generative AI, and deep synthesis technologies. This early regulatory stance positions it as a global leader in AI governance.
- **Focus on security and control**: A central goal of China's AI regulations is to maintain state control over information dissemination and ensure that AI technologies align with national security priorities.
- **Innovation vs. Oversight**: While fostering innovation, China imposes rigorous oversight to mitigate risks like misinformation, privacy violations, and algorithmic discrimination.

Regulations enacted:

- **Algorithm Recommendation Regulation (2022**): Governs recommendation algorithms to prevent monopolistic behavior, dynamic pricing abuse, and algorithmic discrimination. Service providers must file their algorithms with the Cyberspace Administration of China (CAC) within 10 days of deployment.
- **Deep Synthesis Regulation (2023**): Requires labeling of synthetically generated content (e.g., deepfakes) and mandates algorithm filing with the Cyberspace Administration of China
- **Generative AI Measures (2023**): Focuses on public-facing generative AI systems, requiring legal data sources for training, protection of intellectual property rights, and ensuring outputs are accurate and diverse.

Moreover, China has developed several non-binding frameworks and guidelines to guide ethical and responsible AI development.

Code of Ethics for New-Generation Artificial Intelligence
Issued in 2021, this code aims to integrate ethics into the entire life cycle of AI development and use, providing guidance for stakeholders involved in AI-related activities

AI Safety Governance Framework
Released in 2024, this framework provides non-binding guidance for AI developers, service providers, and users to manage AI-security risks.

Future developments:

- China expects to introduce a comprehensive AI law, which is listed in the State Council's legislative work plans for 2024–2025 to unify its patchwork of regulations into a cohesive framework.
- Over 50 new national and industry standards are expected to be formulated by 2026, aiming to integrate AI industry standards with technological innovation

Certain provincial and municipal authorities have also introduced localized initiatives to promote and regulate AI development within their jurisdictions.

Shanghai Regulations on Promoting the Development of AI Industry
- marking China's first provincial-level legislation dedicated to the AI industry, These are a set of provincial-level regulations aimed at fostering innovation and growth in the AI sector within Shanghai.

Shenzhen Special Economic Zone Regulations on AI Industry Promotion
- The primary goal is to promote high-quality development of the AI industry in Shenzhen by encouraging government organizations to lead in AI adoption and increasing financial support for AI research

In addition, China's **Personal Information Protection Law (PIPL),** enacted in November 2021, is the country's primary law governing data privacy and personal information security, directly impacting AI governance. It regulates how AI systems collect, process, and use personal data, particularly in automated decision-making.

Key provisions
- **Consent Requirement**: Organizations must obtain explicit, voluntary consent from individuals before processing their personal information, which includes data used in AI model training and output
- **Transparency**: Users must be informed about how their data is used, especially in automated decision-making processes
- **Access and Modification Rights**: Individuals have the right to access, modify, or delete their personal information processed by AI systems
- **Data Localization**: Certain personal information must be stored within China, with restrictions on cross-border transfers
- **Penalties for Non-Compliance**: Organizations that fail to comply with PIPL can face penalties up to 50 million RMB or 5% of annual revenue and business cessation

	Administrative Provisions on Algorithm Recommendation of Internet Information Services	Interim Measures for the Management of Generative Artificial Intelligence Services	The Administrative Provisions on Deep Synthesis in Internet-based Information Services	Personal Information Protection Law (PIPL)
Effective date	March 1, 2022	August 15, 2023	January 10, 2023	November 1, 2021
Scope	Applies to companies providing internet information services that use recommendation algorithms, including social media, e-commerce platforms, news sites, and other online services.	Applies to service providers that directly offer generative AI services to the public, excluding research and development uses by industry, education, and research institutions	Applies to providers of internet information services using deep synthesis technology, including those generating or editing text, images, audio, video, and virtual scenes	Applies to organizations and individuals involved in the processing of personal information in China or outside China if: • Providing products or services to individuals in China. • Analyzing or assessing activities of individuals in China.
Objective	To standardize algorithmic recommendation services, safeguard national security, protect public interests, and promote the healthy development of internet services	To promote the healthy development and standardized application of generative AI, ensuring it aligns with China's national interests and societal values	To regulate the use of deep synthesis technologies in providing internet information services	To protect the rights and interests of individuals related to their personal information, ensuring that no organization or individual infringes upon these rights

Global AI governance at a glance: Key regulations & frameworks categorized

Parameter	Regulations/Frameworks		
Data privacy and governance	• GDPR (EU) • PIPEDA (CAN)	• Consumer Privacy Protection Act (CAN – proposed) • PIPL (CHN)	• HIPAA (USA) • California Consumer Privacy Act (USA)
Bias & fairness	• EU AI Act • OECD AI Principles	• UNESCO AI Ethics Recommendations • New York City's Local Law 144	• AIDA (CAN - proposed)
Accountability & transparency	• EU AI Act • OECD AI Principles	• UNESCO AI Ethics Recommandations • GDPR (EU)	• AIDA (CAN)
Responsible AI development and deployment	• EU AI Act • OECD AI Principles • UNESCO AI Ethics Recommendations	• NIST AI RMF (USA) • ISO/IEC 42001 • AIDA (CAN)	• Colorado AI Act (USA) • New Generation Artificial Intelligence Ethics Specification (CHN)
Generative AI	• EU AI Act • Canadian Guardrails for Generative AI – Code of Practice • Interim Measures for the Management of Generative Artificial Intelligence Services (CHN)	• The Administrative Provisions on Deep Synthesis in Internet-based Information Services (CHN) • California AI Transparency Act (USA)	

Latest developments in AI governance landscape

AI takes center stage: The Paris AI Summit set the tone for global collaboration, where tech giants and policymakers converged to shape AI's future.

A divided future: The signing of AI Declaration raised questions about the balance between regulation and innovation, highlighting the geopolitical divide on AI governance.

DeepSeek's meteoric rise: A trailblazer in AI innovation, DeepSeek AI is transforming industries with its low-cost cutting-edge technology and open-source AI solutions.

DeepSeek faces global bans: Concerns over DeepSeek's handling of personal data coupled with geopolitical implications have led to bans in multiple countries

DeepSeek sparks global reform?: The case highlights the challenge of harmonizing AI laws globally

Paris AI Summit

The Paris AI Summit, co-hosted by French President Emmanuel Macron and Indian Prime Minister Narendra Modi, brought together leaders from nearly 100 countries, tech giants like Google and OpenAI, academics, and NGOs to shape the future of AI governance. The summit aimed to address global challenges posed by AI, such as its impact on democracy, cybersecurity risks, misinformation, and environmental sustainability.

Key takeaways

- **Regulatory simplification** – Pledges were made to cut red tape, with French President Macron and EU Digital Chief Henna Virkkunen promising to ease regulations, particularly in light of the EU's AI Act
- **AI declaration** – A joint declaration on "Inclusive and Sustainable AI" was signed by 61 countries, excluding the U.S. and the UK. This declaration emphasized the need for open, inclusive, transparent, ethical, safe, secure, and trustworthy AI systems
- **Global collaborations** – The Coalition for Environmentally Sustainable AI, comprising 91 global members, was launched to tackle the environmental impact of AI development, while the inaugural International AI Safety Report, crafted by 96 experts, signifies a pivotal milestone in the global conversation on AI safety.
- **Boosting AI investments**
  - 'Current AI' Initiative: $400M initial investment
  - EU AI Commitment: Investment to strengthen AI, though smaller than the U.S.'s $500B pledge
  - AI Champions Initiative: Over 70 companies to enhance AI in manufacturing, energy, and defense
  - ROOST Initiative: $27M for developing online AI safety tools

Declaration on Inclusive and Sustainable Artificial Intelligence – What Happened at the Signing?

- Signed by 61 countries. However, US and UK declined to sign the declaration
- The U.S. refused to sign due to concerns about excessive regulation that could stifle innovation in the AI sector. The US Vice President JD Vance emphasized the need for minimal regulatory interference to maintain U.S. leadership in AI development
- Vance also expressed fears about AI being co-opted into tools for authoritarian censorship and ideological bias
- The U.K. declined to sign, stating that the declaration did not adequately address global governance of AI and national security concerns

DeepSeek AI's rise and the regulatory storm it's facing

DeepSeek AI, a Chinese generative AI platform, has sparked significant regulatory and geopolitical discussions due to its rapid rise and controversial practices

Data privacy and compliance:

- DeepSeek's handling of personal data has raised concerns globally, particularly regarding transparency and compliance with privacy laws like the GDPR in Europe
- Italy's Data Protection Authority (Garante) banned DeepSeek in January 2025 for inadequate responses on data sharing and storage practices, prompting investigations in other EU countries like France, Belgium, and Ireland
- DeepSeek has been banned or suspended in several countries, including Australia, South Korea, and parts of the U.S., due to non-compliance with local data regulations

Transparency and ethical issues:

- Regulators criticize DeepSeek for its lack of transparency in algorithmic operations and potential biases in decision-making processes
- Ethical concerns include the perpetuation of biases in training data, which could lead to discriminatory outcomes in applications like hiring or lending

Geopolitical implications:

- DeepSeek's emergence highlights the global competition in AI development, with its cost-efficient model challenging U.S. and European AI dominance
- Its open-source approach has fueled debates on intellectual property, national security, and the risks of relying on foreign-developed AI systems
- DeepSeek stores user data in China, raising concerns about potential access by Chinese authorities due to PRC laws requiring cooperation with intelligence agencies

Impact on regulations

- DeepSeek's controversies have intensified calls for stricter AI regulations globally. For instance:
  - The European Data Protection Board (EDPB) expanded its taskforce to address generative AI oversight following the DeepSeek case
- The DeepSeek case underscores challenges in harmonizing global AI regulations. Different countries have varying approaches to privacy, ethics, and liability, complicating compliance for international companies

What is the global outlook of the regulatory landscape?

When it comes to AI regulation and innovation, three main hubs dominate the landscape: North America, Europe, and China.

Federal AI legislation stalls as states advance

In North America, federal AI regulation in the U.S. remains unlikely this year, but activity at the state level is gaining momentum. States like New York and Colorado have introduced AI-related laws, focusing on issues such as bias in hiring algorithms and consumer protections. California, a leader in tech regulation, has implemented rules on self-driving cars, requiring companies to report safety data and address concerns around AI decision-making in autonomous vehicles.

Canada's AI crossroads: federal or provincial rule?

In Canada, the question remains: Will AI governance be handled federally, or will provinces introduce their own AI laws in the absence of national legislation? The outcome of upcoming elections may determine whether Canada takes a fragmented or unified approach to regulation.

The world tightens its grip on AI standards

Meanwhile, global AI standards are expected to strengthen. Organizations such as the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) are refining frameworks for AI safety, ethics, and risk management. These evolving standards could influence national policies even in jurisdictions where AI laws remain weak.

Legal battles forge AI governance

Case law may also play a critical role in shaping AI governance. Courts are increasingly addressing legal challenges related to AI, particularly in intellectual property (IP) disputes. High-profile cases, such as The New York Times suing OpenAI over unauthorized use of its content in ChatGPT, and The Washington Post pursuing legal action over similar claims, highlight the growing tensions between AI developers and content creators. As these cases unfold, judicial rulings could set important precedents for AI-generated content and copyright enforcement.

With regulatory uncertainty, evolving standards, and legal battles, the global outlook for AI governance remains dynamic—and the direction each region takes will have lasting consequences.

Authors

Kanchana Abeywarna

Chief Future Officer

Marsha Subasinghe

Chief Research Officer

Wathsala Dharmakeerthi

Manager - Corporate Finance and Investment Research

Supun Madusanka

Manager – Corporate Governance

Alysha Sabana

Research Assistant

Thank You!

www.infintel.co

info@infintel.co

Kanchana Abeywarna

kanchana@infintel.co

+1 416 998 0728

Check out our latest survey

AI Governance Survey 2025