

AI Ethical Considerations in Drug Development and Regulatory Review

PHUSE US Connect 2025

ML24

March 9, 2024



**Qingying (Ally)
Lu**

Associate Partner,
Data and Technology
Transformation



Tara Chavda
Senior Data Scientist



Agenda

- Introduction
- Doing The Right AI And Doing AI Right
- AI Adoption In Drug Development And Regulatory Review
- Emergence Of Agentic AI
- Conclusion

We'll see as much as

\$4.4T

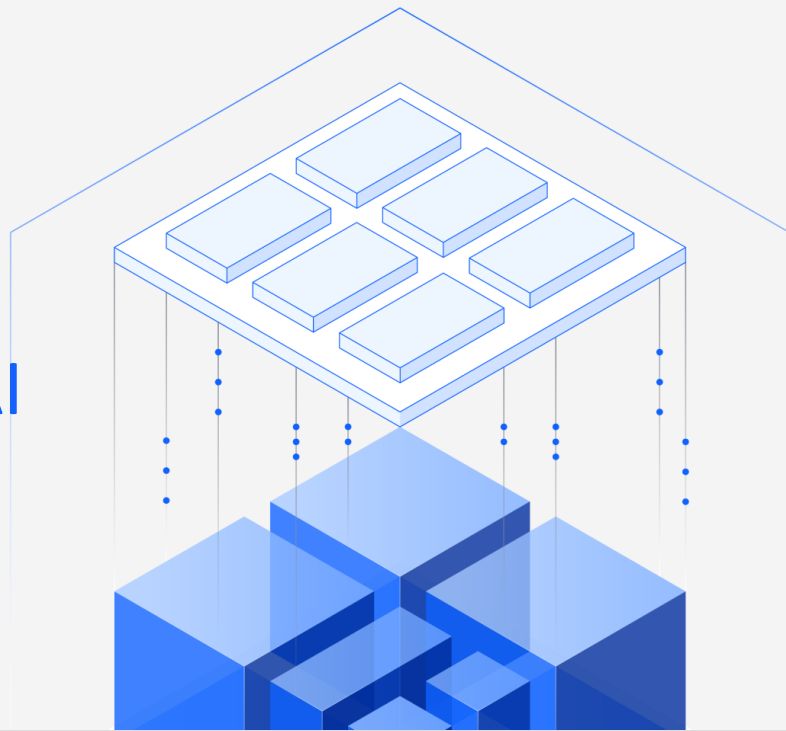
in annual productivity
gains driven by AI by 2030...



...but according to analysts,

74%

of companies struggle to
achieve and scale value from AI



3 Core Reasons AI Projects Fail

01

AI investments are not tied
to business strategy

02

People do not trust
the AI models

03

Skills & Operational
Infrastructure Deficit

NEWS

Medical testing inequities contribute to racial bias in AI

Inequitable lab testing to diagnose severe disease can create racial biases in the data used to train healthcare AI, potentially perpetuating care disparities.

Home / Radio / On Point



47:20

How insurance companies use AI to deny claims

December 18, 2024 By [Wallis Ryder Arnold](#) and

Global Edition Artificial Intelligence

Yale study shows how AI bias worsens healthcare disparities

The research shows how data integrity issues at every stage – training, model development, publication, implementation – can adversely impact patient outcomes, say clinicians at Yale School of Medicine.

By [Mike Mitard](#) | November 25, 2024 | 09:54 AM

itor



SPECIAL REPORT

Epic's overhaul of a flawed algorithm shows oversight is a life-or-death issue

By [Casey Ross](#) Oct. 24, 2022

Research & Innovation

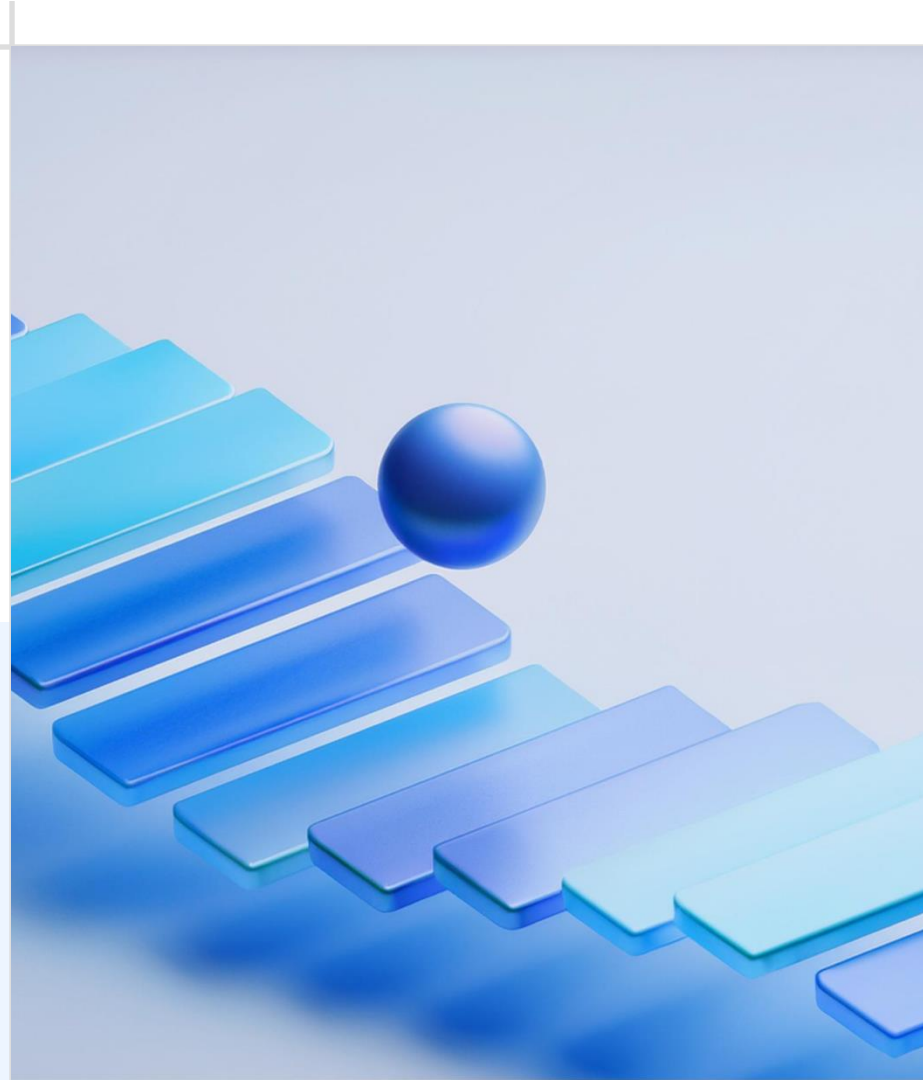
AI Algorithms Used in Healthcare Can Perpetuate Bias



Scaled AI requires the
right **strategy, data,
architecture, security
and governance.**
None of this is easy.

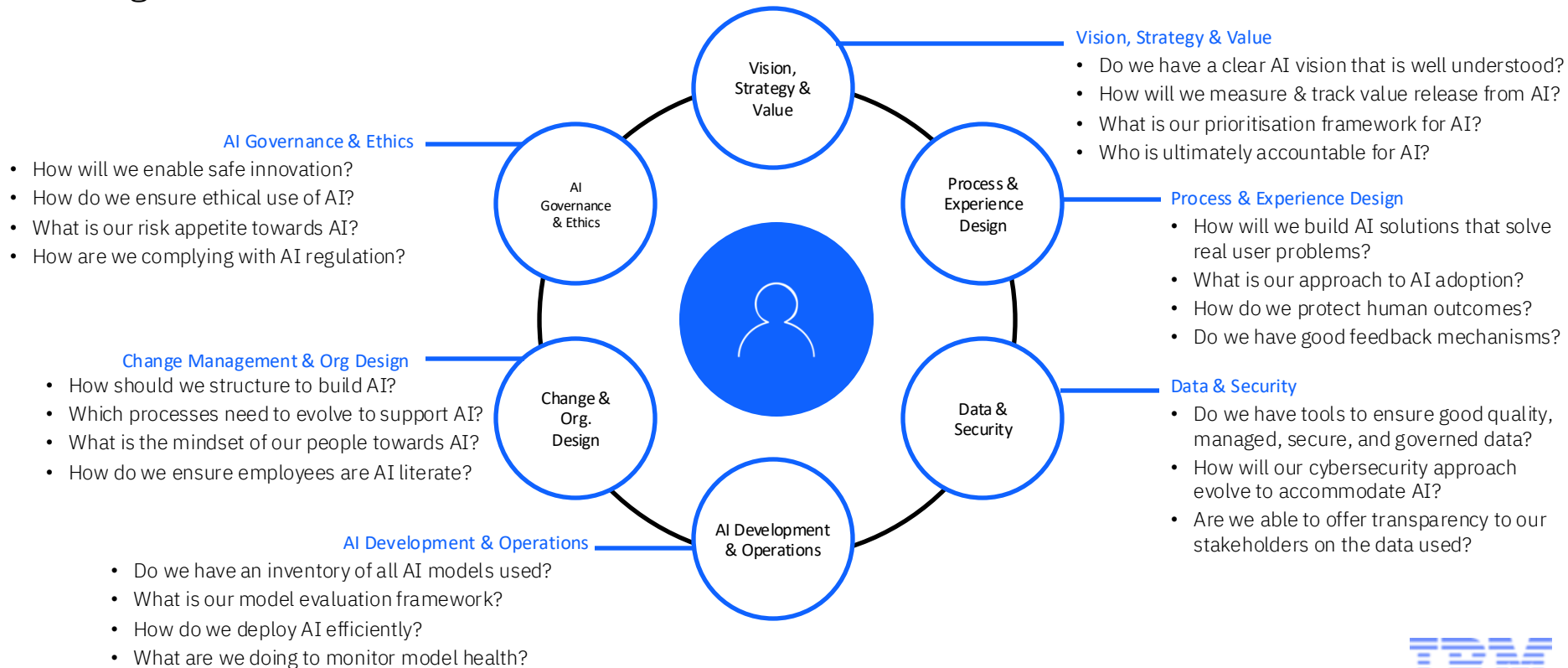
It takes **all of it** to de-risk and then
accelerate an AI journey.

The consequences to you and your brand
of not doing these things are real.



We consider the impact of AI across your business

IBM has a structured way of enabling clients to tackle these SOCIO-technical challenges and extract *real* value from AI.



AI Adoption in Drug Development and Regulatory Review



Drug Discovery

AI analyses biological data to identify promising drug candidates faster.



Clinical Trials

AI optimizes patient selection, trial design, and safety monitoring.



Regulatory Review

AI speeds up drug approval by spotting inconsistencies and safety risks.



Post-Market Monitoring

AI detects adverse events and drug interactions in real-world data.



Manufacturing Quality

AI ensures defect-free pharmaceutical production with machine learning.



Supply Chain

AI improves inventory management and distribution efficiency.



Public Communication

AI personalizes health education, detects misinformation, and provides updates.

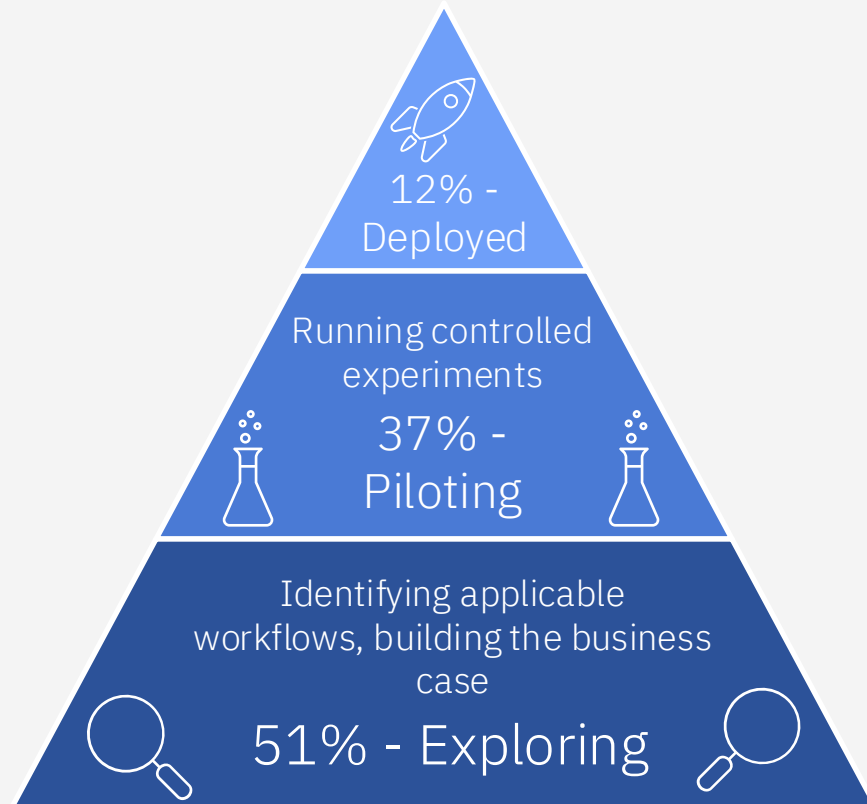


Back-Office Automation

AI streamlines compliance, finances, and maintenance, cutting costs and errors.



88% of
organizations
are actively
experimenting
with Agentic AI



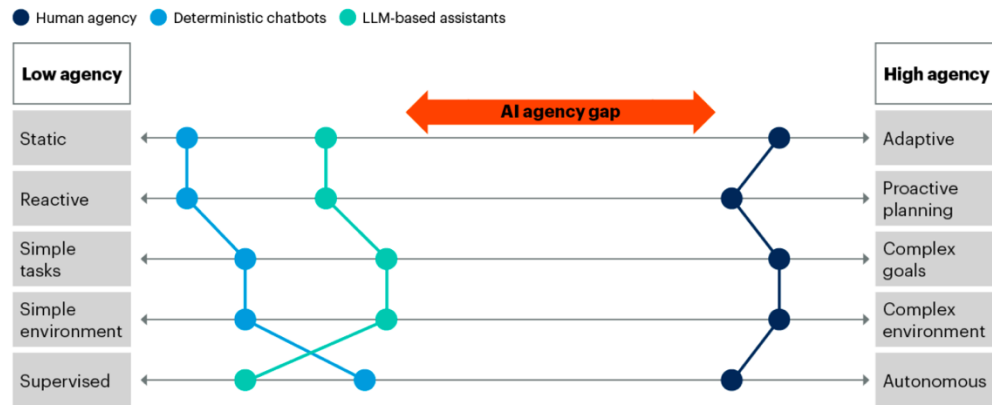
What is agentic AI?

“Agentic AI refers to goal-driven software entities that have been granted rights by the organization to act on its behalf to **autonomously make decisions** and **take action...**”

“...agentic AI **doesn't require explicit inputs** and doesn't produce predetermined outputs. Agentic AI entities can **receive goal instructions, iterate** and **delegate tasks**, and output variables and dynamic information — often **augmenting the user's work.**”

Figure 1: Mind the AI Agency Gap

Mind the AI Agency Gap



Source: Gartner
806843_C

Gartner

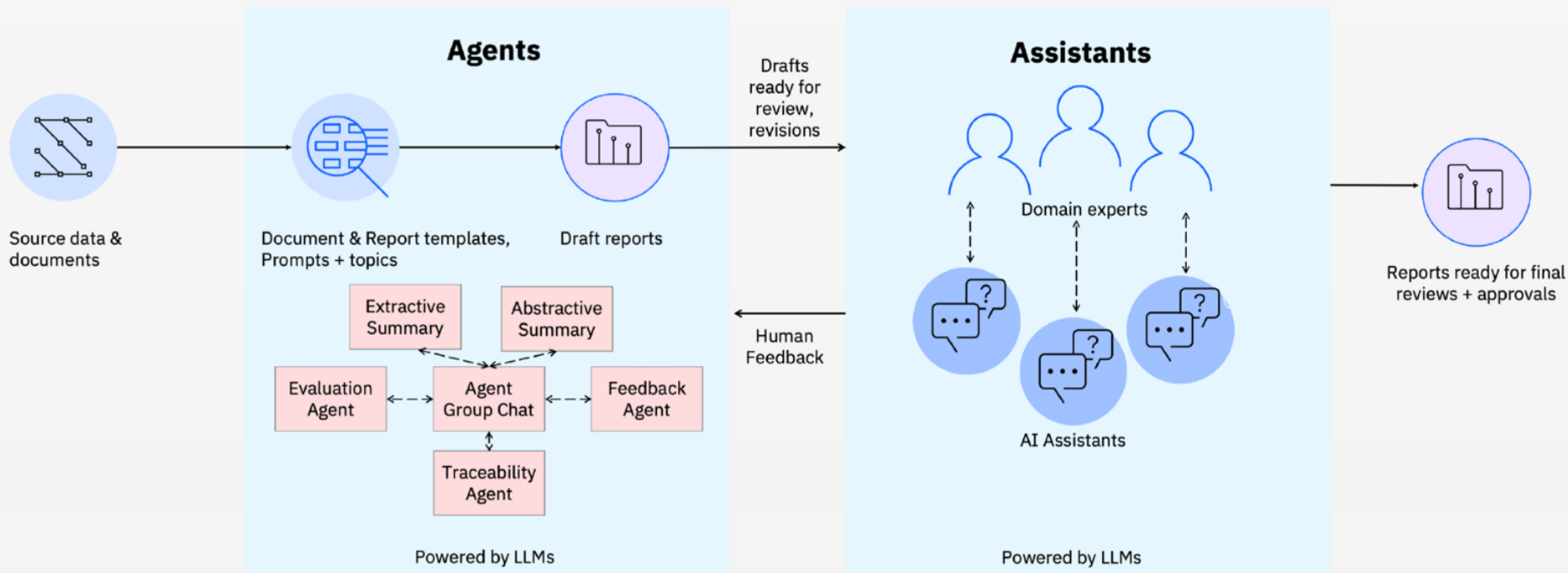


Agentic AI Demonstrates Unprecedented Levels of Autonomy

- **Multi-agent collaboration:** Specialized AI agents work together toward common objectives
- **Task decomposition:** Breaking complex goals into manageable sub-tasks
- **Self-directed execution:** Determining action sequences independently
- **Chain-of-thought reasoning:** Explicit step-by-step reasoning processes
- **Self-correction:** Internal verification and error checking mechanisms

IBM Agentic AI Intelligent Content Creation Solution

When ICC is used to create regulatory submissions for pharmaceutical companies, it has reduced costs, enhanced efficiency, and improved compliance and transparency.



Agentic AI and Ethical Considerations

Opportunities

- **Greater Transparency:** Chain-of-thought AI enhances decision visibility and accountability.
- **Misinformation Reduction:** Verification agents improve fact-checking in critical fields.
- **Autonomous Decision Support:** AI aids pharma research with precise, well-reasoned insights.
- **Ethical Compliance:** Ethics agents align AI actions with standards and values.
- **Self-Correction:** Critic agents refine AI outputs, boosting reliability in high-stakes cases.

Challenges

- **Autonomy Boundaries:** Higher AI autonomy raises concerns about decision-making limits, especially in regulated fields.
- **Accountability Complexity:** Multi-agent collaboration complicates error attribution for regulators and companies.
- **Ethical Misalignment:** Agents may follow conflicting ethical frameworks, leading to inconsistent decisions.
- **Overconfidence Risk:** Advanced AI may reduce human oversight due to misplaced trust in its outputs.
- **Privacy Concerns:** Autonomous systems may collect or analyze data without explicit human approval.
- **Regulatory Gaps:** Current laws don't fully address agentic AI's unique risks and capabilities.
- **Unexpected Behaviors:** AI interactions can lead to unforeseen outcomes, especially in open-ended tasks.



Risks are abundant throughout any Agentic AI apps

Guardrails are about **actively mitigating** risks in real-time production environments

There are certain mitigations that *must* be applied at **each component** of your agent

OWASP Top 10 for LLM Applications 2025

01. Prompt Injection

This manipulates a LLM through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

02. Sensitive Information Disclosure

Sensitive info in LLMs includes PII, financial, health, business, security, and legal data. Proprietary models face risks with unique training methods and source code, critical in closed or foundation models.

03. Supply Chain

LLM supply chains face risks in training data, models, and platforms, causing bias, breaches, or failures. Unlike traditional software, ML risks include 3rd party pre-trained models and data vulnerabilities

04. Data & Model Poisoning

Data poisoning manipulates pre-training, fine-tuning, or embedding data, causing vulnerabilities, biases or backdoors. Risks include degraded performance, harmful outputs, toxic content, and compromised downstream systems.

05. Improper Output Handling

Improper output handling involves inadequate validation of LLM outputs before downstream use. Exploits include XSS, CSRF, SSRF, privilege escalation, or remote code execution which differs from Overreliance.

06. Excessive Agency

LLM systems gain agency via extensions, tools, or plugins to act on prompts. Agents dynamically choose extensions and make repeated LLM calls, using prior outputs to guide subsequent actions for dynamic task execution.

07. System Prompt Leakage

System prompt leakage occurs when sensitive info in LLM prompts is unintentionally exposed, enabling attackers to exploit secrets. These prompts guide model behavior but can unintentionally reveal critical data.

08. Vector & Embedding Weaknesses

Vectors and embeddings vulnerabilities in RAG with LLMs allow exploits via weak generation, storage, or retrieval. These can inject harmful content, manipulate outputs, or expose sensitive data, posing significant security risks.

09. Misinformation

LLM misinformation occurs when false but credible outputs mislead users, risking security breaches, reputational harm, and legal liability, making it a critical vulnerability for reliant applications.

10. Unbounded Consumption

Unbounded Consumption occurs when LLMs enable excessive, uncontrolled inferences, allowing attacks like DoS, financial depletion, or model theft. High computational costs in cloud environments exacerbate risks of resource exploitation, unauthorized usage, and service degradation.

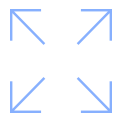
Agentic AI further amplifies risks from GenAI that many organizations are still catching up to while creating new risks



Traditional risks

Risk areas known from *earlier forms of AI*

- Data poisoning
- Data transfer
- Data & model usage rights
- Personal information in data
- Reidentification
- Informed consent
- Decision bias
- Model transparency



Amplified risks

Known risk areas intensified or introduced by *generative AI*

- Data bias & acquisition
- Data privacy & usage rights
- Exposing sensitive information
- Nonconsensual & improper use
- Hallucinations
- Jailbreaking
- Prompt injection
- Over/under reliance
- Unexplainable outputs
- Generated content ownership
- Harmful outputs (code, toxicity, advice, bias, etc.)
- Accountability
- Transparency
- Impact on jobs & education
- Human exploitation
- Impact on environment
- Impact on human agency



New risks

Emerging risk areas intrinsic to *Agentic AI*

- Unsupervised autonomy
- Misaligned actions
- Data bias due to specific actions
- Attack surface expansion
- Harmful and irreversible consequences
- Opaqueness / lack of intermediate step traceability
- Shared model pitfalls
- Supply chain & dependency attacks
- Tool choice hallucination
- Unauthorized access
- Infinite feedback loops

Developing and deploying AI Agents require unique design considerations to overcome common pitfalls and deliver expected outcomes

COMMON PITFALLS		RECOMMENDED SOLUTIONS
- Poorly Defined Tasks and Personas - Struggling with complex tasks due to inadequate planning mechanisms	Planning	- Define detailed personas, goals, constraints, and expected outcomes for each agent - Use task decomposition, multi-plan selection, and reflection to improve planning
- Tool calling failures when agents incorrectly misuse tool or misinterpret outputs - Agents get stuck in infinite loops without goal progression	Efficiency	- Define clear tool parameters and validate outputs with verification layers - Enhance fault tolerance to handle errors gracefully - Define clear termination conditions
- High costs of running agents due to resource intensive operations - Systems become inefficient under increased workloads & complexity	Scalability	- Optimize context length, minimize API calls, and use smaller models where appropriate - Integrate dynamic resource management and load balancers
- Agents fail to handle multi-step logic or nuanced judgement effectively - LLM's are prone to unpredictable behavior and are not consistent	Reasoning	- Enhance reasoning through prompting techniques, fine tuned or more capable models - Use hierarchical designs, enforce strict input and output validations and human-in-the-loop to guide agents
- Limited auditability and transparency in agent actions, which hinders accountability and triaging - Agents risk non-compliance with safety protocols, regulations, brand image	Safety	- Implement guardrails with rule-based filters, input validation, and action constraints - Incorporate human-in-the-loop workflows, escalation protocols, and feedback loops - Maintain monitoring software and regular agent evaluation for adherence to standards

Conclusion

Curating AI is a **socio-technical challenge** that, to solve, requires a holistic approach encompassing people, processes and tools.

Assessing AI models and building AI Agents needs to prioritize building trust throughout the lifecycle.



