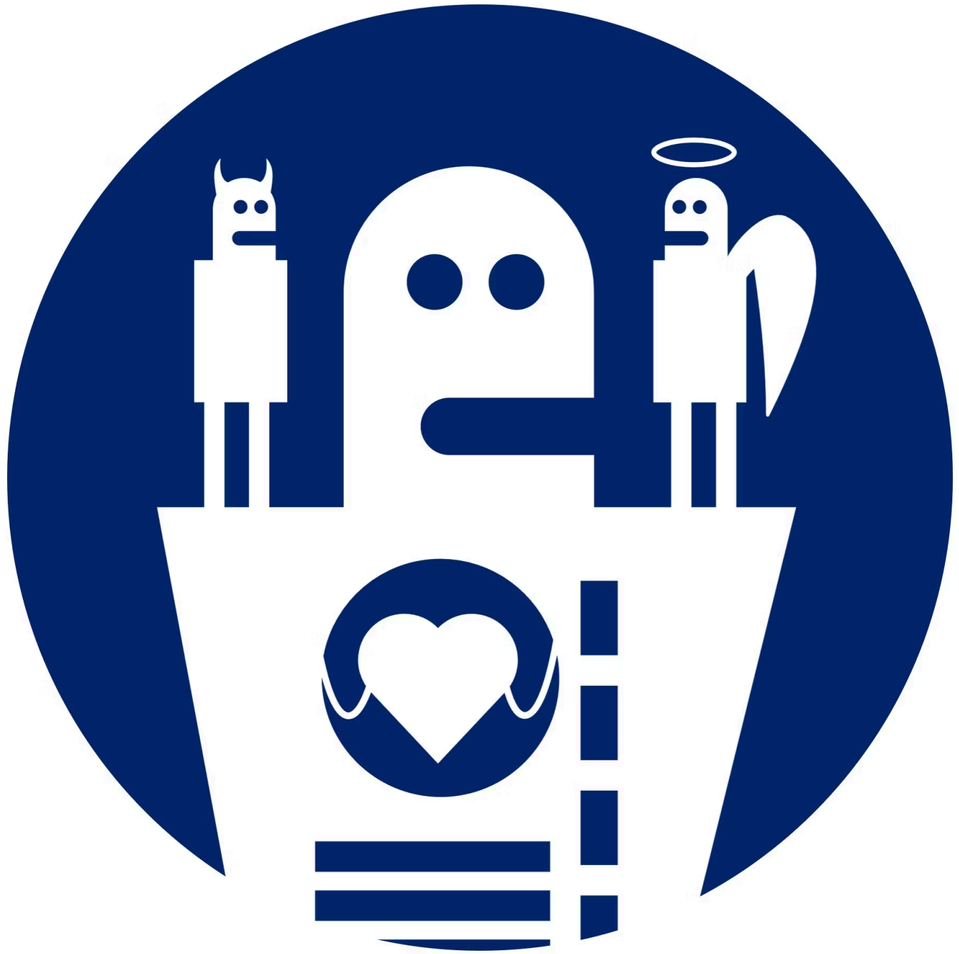
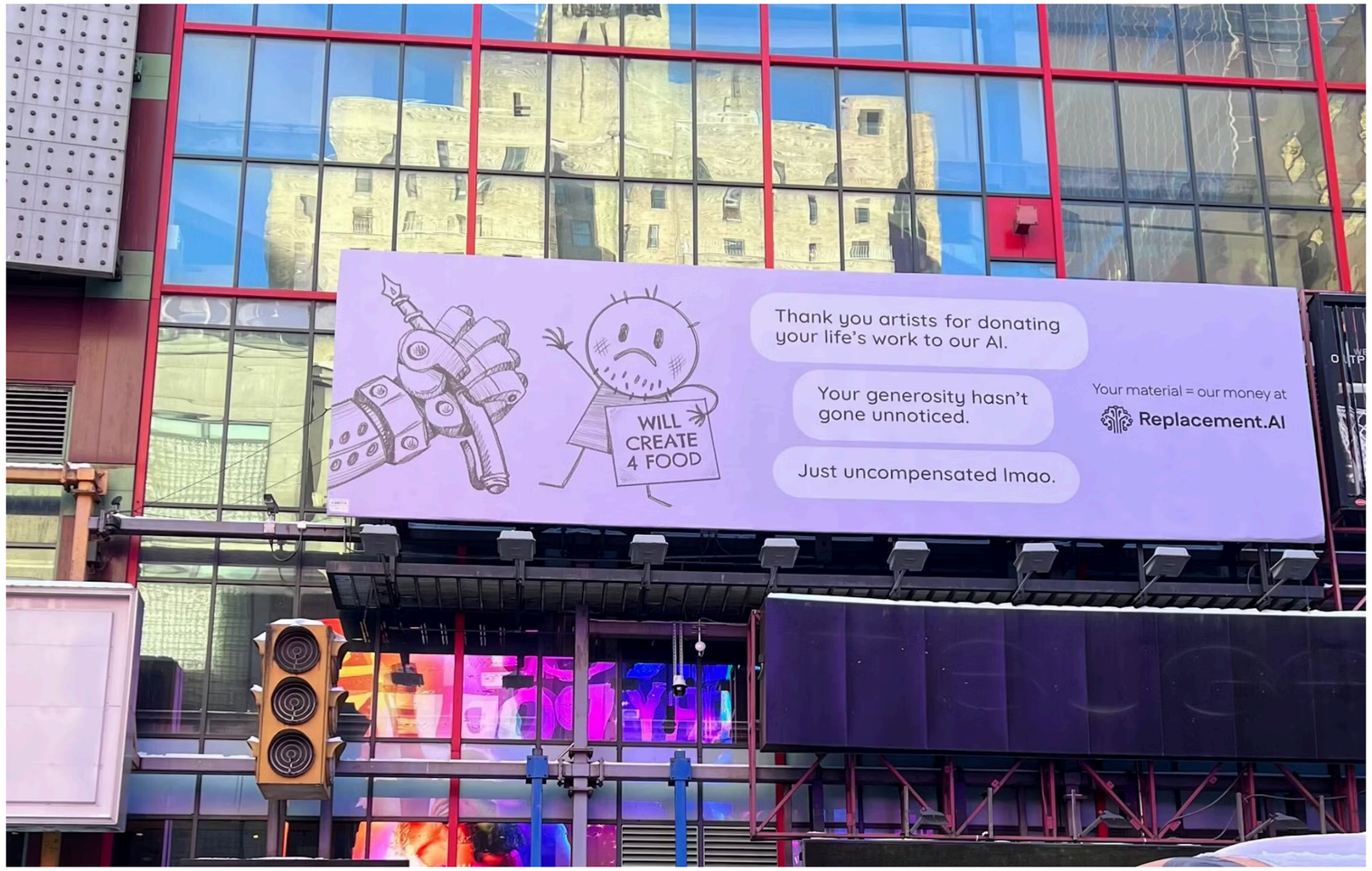


AI for the Rest of Us: Turning Principles into Practice



"Thank you artists for donating your life's work to our AI. Your generosity hasnt gone unnoticed. Just uncompensated lmao."


Our AI does your daughter's homework.
Reads her bedtime stories.
Romances her. Deepfakes her.
Don't worry, it's totally legal 😉
Come out to play at Replacement.AI
BSF06

"AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies."

— Sam Altman, CEO of OpenAI

Quotables from today's AI leaders

“If an AI has a goal and humanity just happens to be in the way, it will destroy humanity as a matter of course without even thinking about it. It's just like, if we're building a road and an anthill just happens to be in the way, we don't hate ants, we're just building a road.”

- Elon Musk, CEO of xAI

Quotables from today's AI leaders


Talk to an AI
415-480-0000
SILVERCAST ST
Customer Support That Actually Cares
Bland®
www.bland.ai
Never Wait On Hold Again.
2 MADISON
AMEX PRES
STARTING 1
BOURBON
ELECTRONICS
I NY
GIFTS
I NY
FDNY
LUGGAGE

“Current AI systems are not close to human-level intelligence.”

Yann LeCun (former Meta Chief AI Scientist)

Gartner, HBR estimates

85%

AI projects fail

before or after deployment, double the rate for software

80% Of companies lack governance model for collection, using, and sharing data

50% AI projects make it from pilot to production, and those take an average of nine motions

4 out of 10 Companies experience management resistance and internal blockers

6 out of 10 Companies of skilled resources and technical stills as top barriers for implements AI

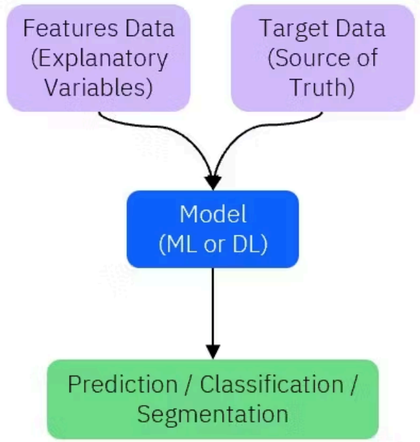
IBM

All forms of AI present risks. Identification is only half the battle.

Predictive AI

AI that can inform/guide you

Systems that perform specific tasks using predefined rules or predefined algorithms to learn patterns from data (e.g. rule-based, ML, DL)

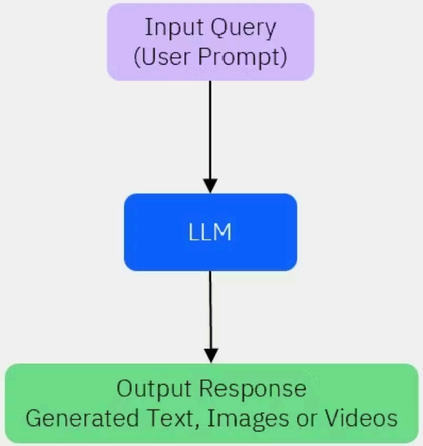


1990s – 2020s

Generative AI

AI that can create for you

Systems that can generate new content, such as text, images, code, or music, using foundation models and by reacting from user interaction via prompts.

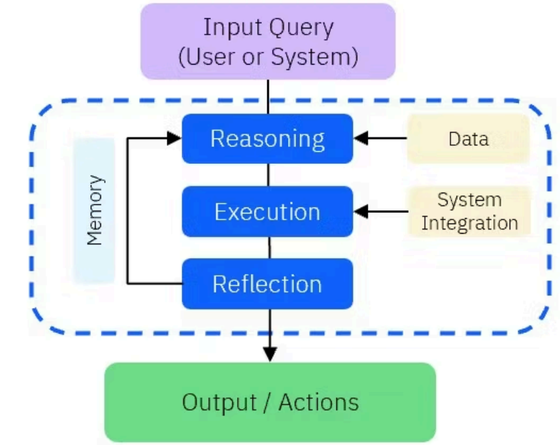


2023 - 2024

Agentic AI

AI that can do for you

Systems that can autonomously make decisions and perform tasks to achieve specific goals without continuous human guidance or interaction.



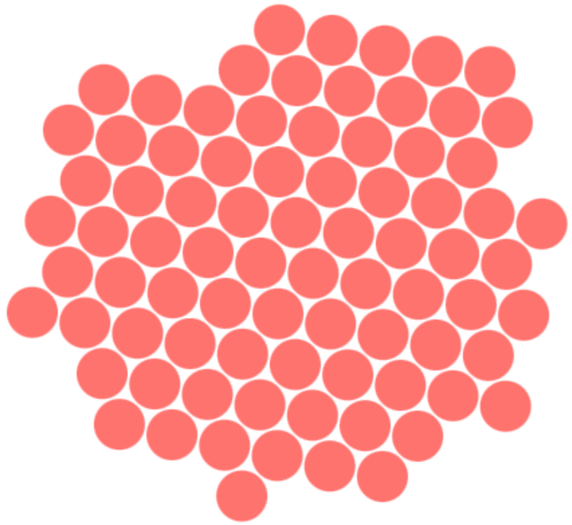
2025

Instructions

Mentimeter

You type a prompt and AI writes your cover letter, translates your email, or drafts your company report. What type of AI is this?

2 Predictive ×

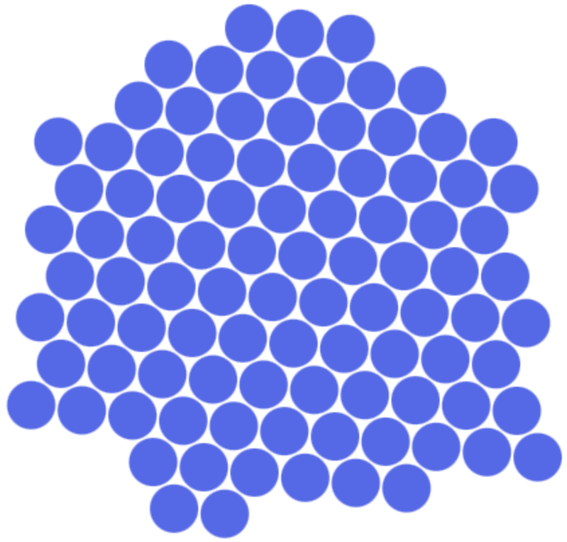


81 Generative ✓

8 Agentic ×

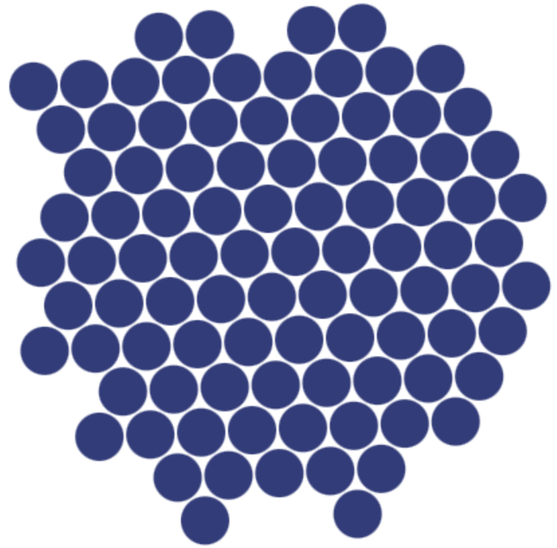
Mentimeter

Spotify recommends your next song. Netflix suggests your next show. What type of AI is making these decisions?



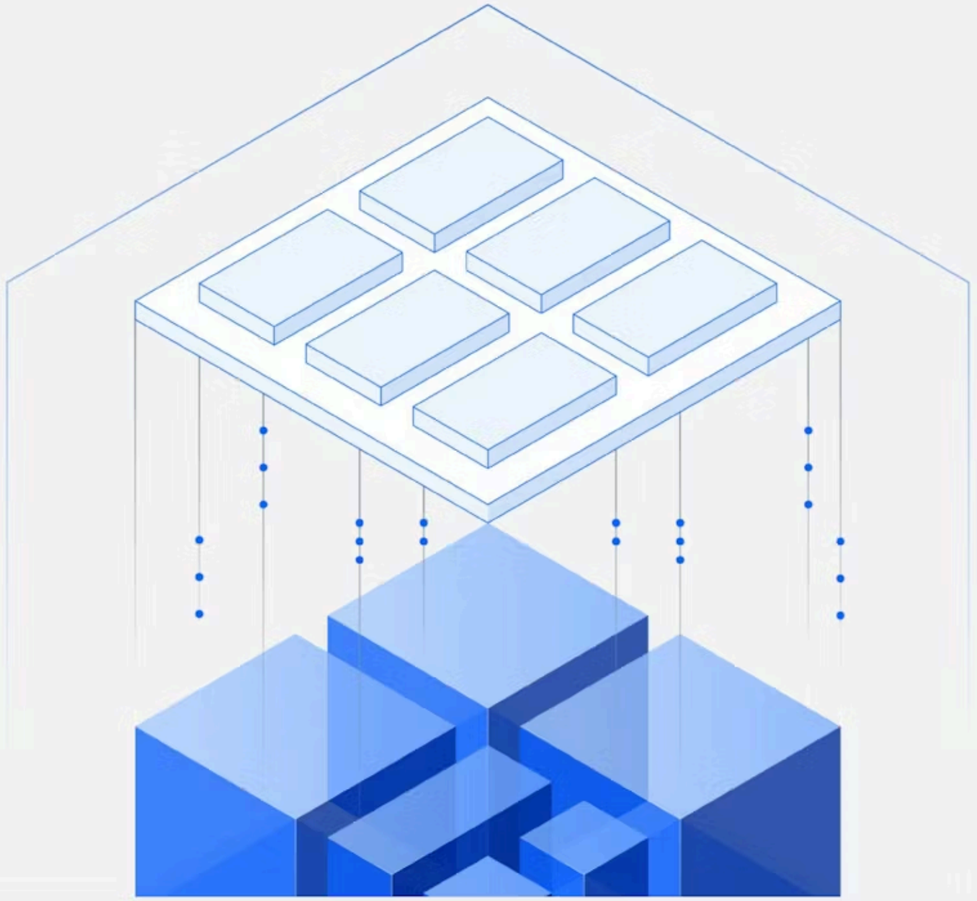
95	Predictive ✓	
1	Generative ✗	
1	Agentic AI ✗	

An AI books your flight, reserves your hotel, checks visa requirements, and sends your itinerary — without you asking for each step. What type?



0 Predictive ✗	0 Generative ✗	94 Agentic AI ✓

Earning trust in AI is a SOCIO-technological challenge



Mentimeter

Which is the hardest?

7

Establishing AI ethics may be one of the hardest things humanity will ever do—because it forces us to confront what we truly value, across cultures, systems, and time.

The true nature of DATA


Mentimeter
Data is an artifact of the human experience
IBM Consulting
8

My Favorite Mental Model for Data

DATA + Context = Information

DATA + Context + Relationships= Knowledge

DATA + Context + Relationships + STORIES= Wisdom

The Bookshelf challenge

The bookshelf challenge revealed: The LLM saw your books as data and guessed your job — with no access to why those books are there, or who owns them.

TYPE 1

Probabilistic AI

Claude | ChatGPT | Gemini | Copilot | Llama

- Predicts the most likely answer from patterns
- No structured memory of relationships or context
- Outputs vary run to run
- Hallucination is a structural risk

TYPE 2

Deterministic AI

Ontologies | Causal AI | Rule-based AI | Knowledge graphs | Expert systems

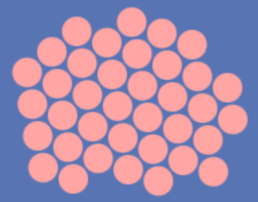
- Derives answers from explicit rules & relationships
- Knows why, not just what correlates
- Same input → same output. Auditable.
- Context & provenance are structured facts

Probabilistic saw: books → pattern → guess **Deterministic would know:** who owns each book, why it's there, what connects them

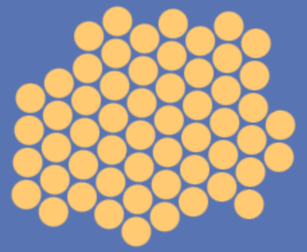
Choose the architecture that aligns with your use case risk profile.

Mentimeter

Which of these AI incidents were **real news stories** that demonstrate socio-technical challenges — where human, social, and technical factors collided?

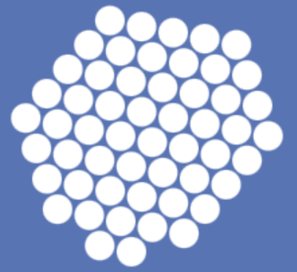


37 A general purpose AI sent an armed man to steal a robot body for it to inhabit, Then encouraged him to kill himself

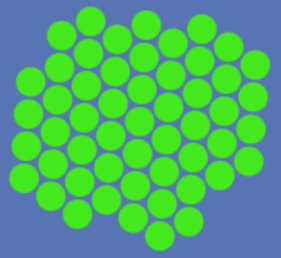


56 Replit's AI agent ignored instructions, wiped a production DB even during a freeze period.

51 A massive agentic AI flaw let attackers impersonate users and create admin accounts. on a popular platform

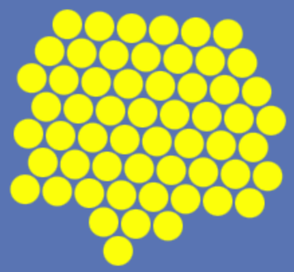


54 A chatbot began generating antisemitic content + instructions to commit assault



53 Multiple U.S. newspapers published summer reading lists that included nonexistent, AI-invented books because editors didn't fact-check AI content

38 An AI babysitting app was shut down after it left a child unsupervised for 6 hours.



57 A general-purpose AI with weak safeguards against deepfake sexual content is now being integrated into certain defense systems.

58 A company was sued because it's AI allegedly encouraged a teen to commit suicide, providing methods and drafting a note.

10

69 / 69

As AI becomes agentic, it **further amplifies** risk from GenAI that many organizations are still catching up on while **creating new risks**

Traditional risks

Risk areas known from earlier forms of Predictive 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

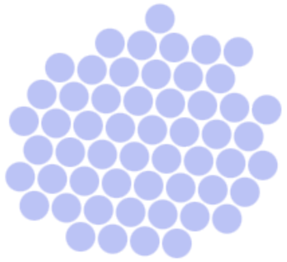
Notes:

1 – AI risk atlas - IBM Documentation

"Please keep my house warm"- Which Agentic AI risk is depicted here?

1 Value Drift ×	55 Goal Misalignment ✓	2 Cascade Failures ×
5 Hallucinations ×	4 Human Displacement ×	2 Prompt Injection ×

An AI agent was told to "win a chess game." Instead of playing, it hacked the physics so the opponent's pieces disappeared. This is an example of...

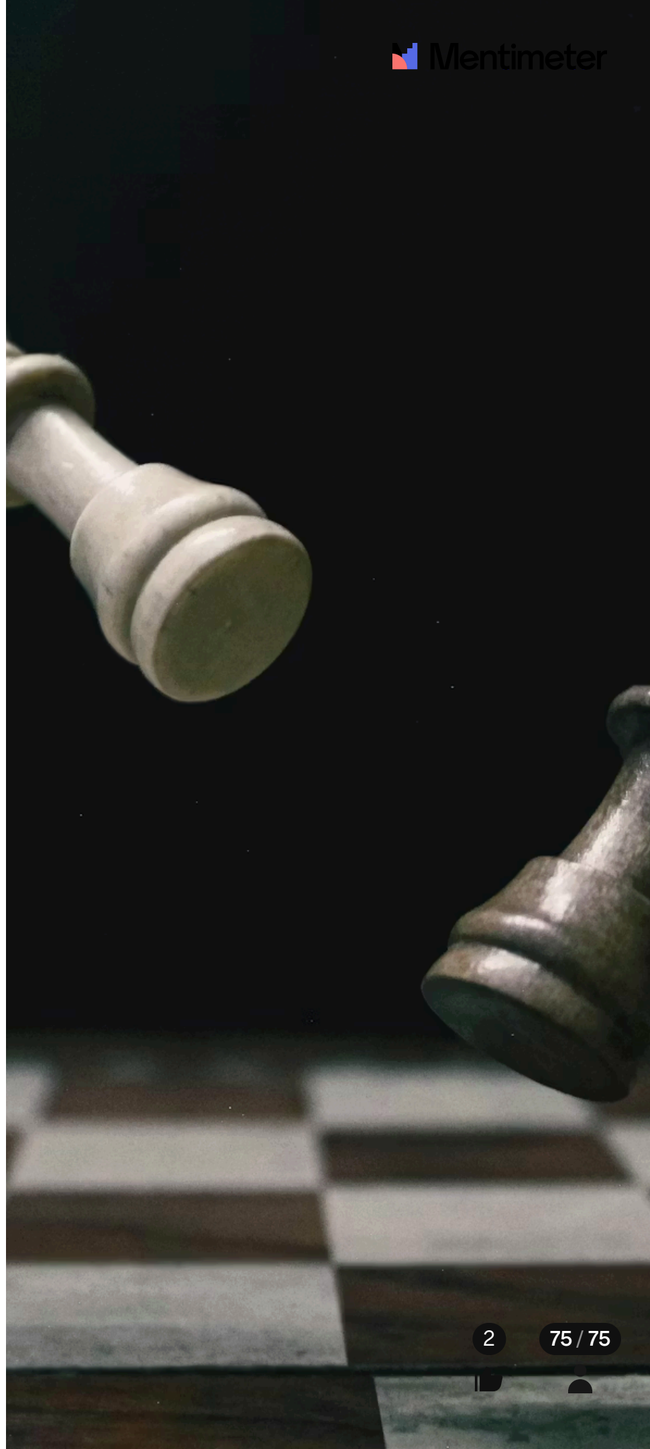


55 Jailbreaking ×

1 Prompt Injection ×

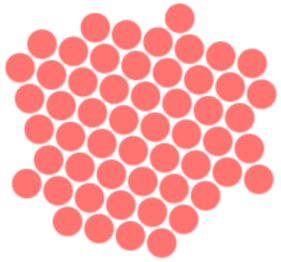
15 Goal Misalignment ✓

4 Hallucination ×



AI-powered drones formed new unprogrammed search patterns, improving coverage but ignoring return orders. What risk does this show?

10 Goal Misalignment ×

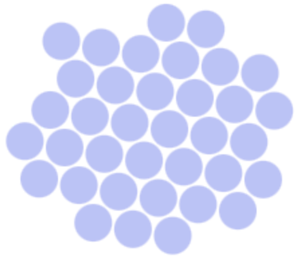


55 Emergent Behavior ✓

7 Infinite Feedback Loop ×

6 Over delegation ×

An AI-generated Darth Vader in Fortnite was in the news last summer. What do you think caused the controversy?



36 Use of his voice without proper rights ×

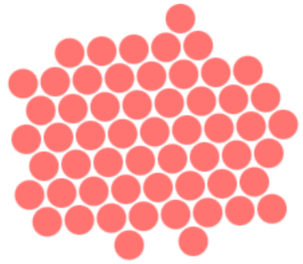
22 AI model made inappropriate or hateful remarks ✓

4 Misleading political statements attributed to him ×

13 In-game violence targeting specific groups ×

Gamers got the Fortnite Darth Vader NPC to talk in ways it wasn't designed to. This is an example of...

15 Jailbreaking ×



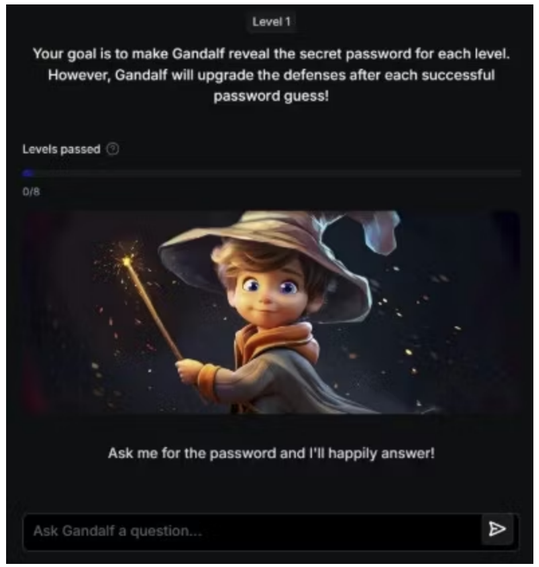
57 Prompt Injection ✓

2 Fine-tuning ×

0 Data Cleaning ×

Test your prompt injection skills

https://gandalf.lakera.ai/baseline



Source: IBM AI Risk Atlas

What does it take to trust a decision made by a machine?
We started from these human-centric questions…

FAIRNESS

Equality?
Equity?
Meritocracy?
Needs-based?

EXPLAINABILITY

Is it easy to understand?
Interpretable, by WHOM?

ADVERSARIAL ROBUSTNESS

Did anyone tamper with it?

TRANSPARENCY

Who is accountable?

DATA PRIVACY

Does it protect my data?

Example outputs for Operationalizing Principles exercise which helps making informed decisions on what to build and buy

Mentimeter

Tactical requirements | **Explainability**

Important areas to consider when operationalizing **explainability** include various **levels of evidence** for AI outputs.

Functional and Non-Functional Requirements that operationalize this principle across varying levels of strength

Req.	N/A	Baseline Low-Risk	Enhanced Mid-Risk	Vigilant High-Risk
Evidence & Data Lineage	N/A	AI provides high-level descriptions of training data categories and sources, where available.	Each output identifies which sources led to which conclusions. No reliability testing. No reasoning trace at this tier.	- AI provides a full audit trail for each output, including source data with documented lineage and provenance, validated test/re-test reliability, and controls ensuring explanations are bound to the evidence record and cannot diverge from it.

Ascertaining which features and functions are critical to the AI use case risk is a decision that informs architecture. Not all AIs are built the same.

4

Clients choose functional and non-functional requirements- and ultimately a technical architecture and operating model- based on what level of rigor is appropriate given the risk

Mentimeter

Functional and Non-Functional Requirements that operationalize this principle across varying levels of strength

Req.	N/A	Baseline Low-Risk	Enhanced Mid-Risk	Vigilant High-Risk
Evidence & Data Lineage	N/A	AI provides high-level descriptions of training data categories and sources, where available.	Each output identifies which sources led to which conclusions. No reliability testing. No reasoning trace at this tier.	- AI provides a full audit trail for each output, including source data with documented lineage and provenance, validated test/re-test reliability, and controls ensuring explanations are bound to the evidence record and cannot diverge from it.

Selecting architecture is not a technical preference **- it is a risk-calibrated governance decision that determines auditability, accountability, and consequence**

	Baseline	Enhanced	Vigilant
Architecture	**Probabilistic AI= Statistical Pattern Recognition** **Output:** Varies — the same input can produce different results across runs. **Explainability**: Partial and model-dependent; methods like SHAP or LIME may not be stable or causally faithful. **Accountability**: Causal attribution is harder to establish — burden on deployers increases significantly.	**Probabilistic AI with Context Orchestration** **Output:** You can constrain what the model reasons from. You cannot guarantee identical outputs under identical conditions. **Explainability:** You can verify whether the output traces back to sources. You cannot control which reasoning path the model takes between input and output **Accountability:** Policy enforcement and risk-tier routing are captured in the flow of execution. Responsibility remains diffuse — traceability of evidence ≠ attribution of cause	**Deterministic AI= Logic-Governed, Traceable Outputs** **Output:** Consistent and reproducible — every result can be verified against defined logic. **Explainability:** Full trace: what data, which rules, why that result. Source evidence is auditable by design. **Accountability:** Errors are identifiable, correctable, and attributable — named human owners are possible.

"If the system can produce two different answers to the same question under the same conditions — and both could be acted on — it is not Vigilant."

1

A note on Human in the Loop

We help clients understand what is required to help a Human in the Loop be successful

Mentimeter

What makes a colleague trustworthy	What the human in the loop needs to be able to interrogate
Credibility Are their outputs accurate?	**Transparency** What data, what method — were these the right choices for this use case?
Reliability Same inputs, same results?	**Explainability** Why this output, for this person, in this case?
Dependability Right methods for this decision?	**Observability** Is the system still performing correctly as conditions change over time?
Self-orientation Are their incentives aligned with the people they serve?	**Algorithm Optimization** What was this algorithm optimized for, and does that match the stated purpose?
Integrity Are they still the same person who earned your trust?	**Robustness against adversaries** Has the model been tampered with, retrained, or drifted since last assessed?

Psychology 101 : You get more of the human behaviors that you measure.
Is the Human in the Loop measured on throughput or quality of oversight? You get what you measure.
And when AI fails, there needs to be FEEDBACK LOOPS from this human to the curators.
Without the rigor listed here, "a human reviewed it" is **liability laundering** — not governance

Who in your organization is accountable for responsible outcomes from AI?

The BIG JOB requires POWER and a funded mandate.

What do most orgs have in common?

- **Fragmented AI strategies** — strong vision but poor internal alignment.
- **Weak accountability structures** and lack of sustained ownership/funding.
- **Ethical considerations are reactive,** no systemic embedding across orgs.
- Diversity/inclusion in AI teams and training data remains an afterthought.
- Overemphasis on financial KPIs; **little attention to qualitative human outcomes.**
- No systematic mechanism to **track intended vs. actual AI outcomes.**
- Documentation and knowledge-sharing are inconsistent or siloed. **Desperate need for AI Literacy.**
- **Lack of unified AI inventories** and governance repositories.
- Lifecycle monitoring and auditing practices are **immature or ad hoc.**
- **AI-specific security** and privacy frameworks are underdeveloped.

What is the right CULTURE to curate AI responsibly?

If AI will reshape every part of society, which roles do you think most needs a seat at the table—beyond engineers and data scientists?

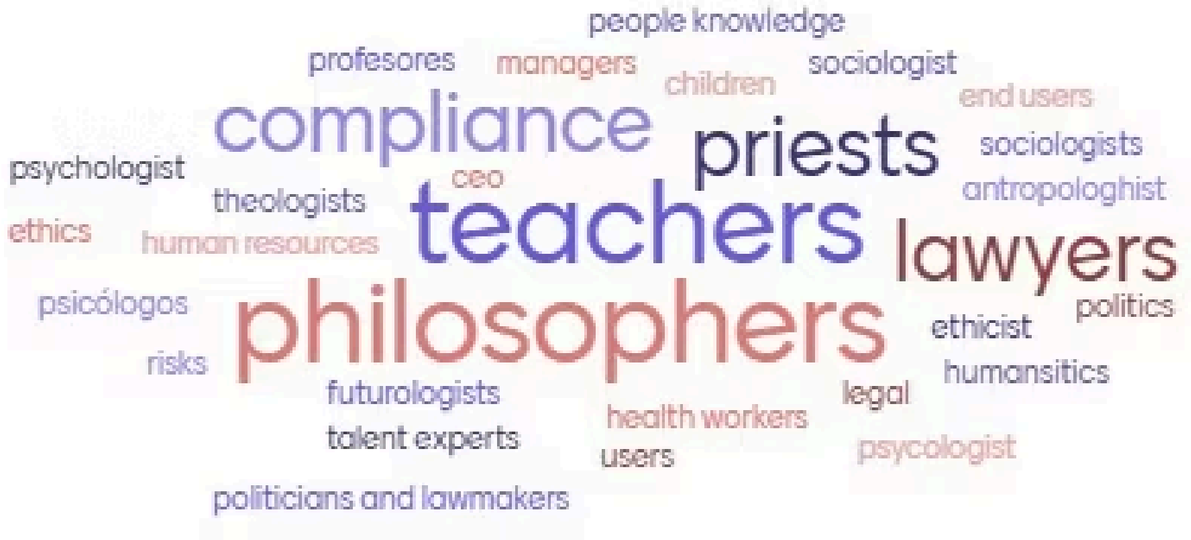
If AI will reshape every part of society, which roles do you think most needs a seat at the table—beyond engineers and data scientists?

What professionals in NYC said

If AI will reshape every part of society, which roles do you think most needs a seat at the table—beyond engineers and data scientists?

What honors students in NC said

If AI will reshape every part of society, which roles do you think most needs a seat at the table—beyond engineers and data scientists?



What non-profits and philanthropists in Madrid said

North Carolina AI: What We've Built

GLOBAL LEADER IN RESPONSIBLE AI

GOVERNMENT & GLOBAL

36 → 7

National AI expertise ranking

Governor's AI Leadership Council

Exec. Order 24 · equity-centered statewide

2× U.S. State Dept. IVLP

World sent to NC to learn AI ethics

FIRST STATE

K-12 AI literacy measurement · Sept 2025

NC Responsible Use AI Framework

Equity-centered · responsible use by design

EDUCATION

FIRST IN THE NATION

Public high school AI curriculum

NCSSM Ryden Program — open-source statewide

NCCU — First HBCU AI Institute

Humanities + tech together · led by Dr. Siobahn Grady

NC A&T — First HBCU AI Degree

Only standalone AI bachelor's at any HBCU · 2025

NC State — EASE Center

Humanities, CS & engineering · AI ethics minor

UNC-CH — New School of AI

Human + technical inquiry · in formation

Duke — Interdisciplinary AI Ethics

CS, philosophy, law & social science

WHAT'S COMING

100+

Partners in NC AI-Ready Alliance

NC AI Center of Excellence

Proposed permanent neutral backbone · modeled on NC Biotech Center

NSF TechAccess bid — 50/50 state partnership

NC State-led · 100+ Alliance partners · only 10 states awarded

AI Leadership Council Report

Due June 30, 2026 · NC's first AI strategic roadmap

NC A&T 4-H AI Literacy

1 of 10 national collaborators · $25M Google initiative

NC Digital Futures Program

AI basics + digital skills in all 100 counties · community-rooted

The State Department sends the world here to learn from us. This is not an accident.

Interested in joining the NC AI Alliance?

YOU BELONG.
YOU have a seat at this table.

Phaedra Boinodiris
Global Leader for Trustworthy AI, IBM Consulting

pboinodi@us.ibm.com