Mitigating Bias in AI: Introduction Webinar

*by **Holistic AI** and hosted by **The Alan Turing Institute***

28.03.2025

Meet your presenters

Fiona Thendean

AI Scientist (AI Governance Solutions)

Fiona is an AI scientist on the AI Governance Solutions team at Holistic AI, specialising in developing quantitative approaches to AI governance. She earned her MSc in Artificial Intelligence for Sustainable Development at UCL, where she focused her dissertation on Explainability for Deep Reinforcement Learning in the context of climate change policy models. Before moving to London, she studied Statistics and Machine Learning at Carnegie Mellon University.

Sachin Beepath

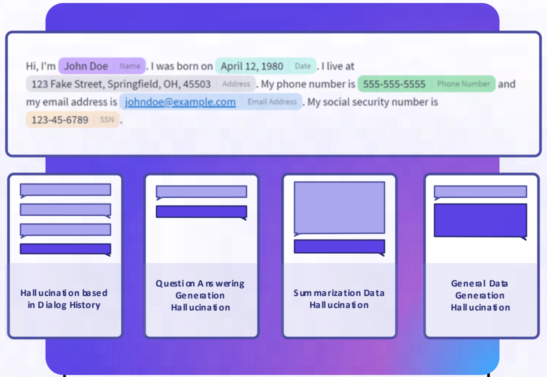
Head of AI Governance Solutions

Sachin is the Head of AI Governance Solutions at Holistic AI, specialising in responsible AI adoption and scalability. Originally from Toronto, he studied astrophysics at the University of Waterloo before earning an MSc in Data Science and Machine Learning at UCL, where he did his dissertation on Bias in Facial Recognition. He works closely with industry leaders, business executives, and technical teams to drive fair, transparent, and impactful AI innovation.

Holistic AI

Holistic AI is an AI Governance company that aims to empower enterprises to adopt and scale AI with confidence.

Holistic AI Platform

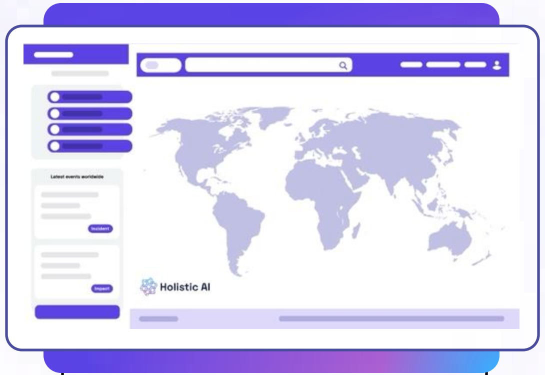


Hi, I'm John Doe Name . I was born on April 12, 1980 Date . I live at 123 Fake Street, Springfield, OH, 45503 Address . My phone number is 555-555-5555 Phone Number and my email address is johndoe@example.com Email Address . My social security number is 123-45-6789 SSN .

Hallucination based in Dialog History
Question Answering Generation Hallucination
Summarization Data Hallucination
General Data Generation Hallucination

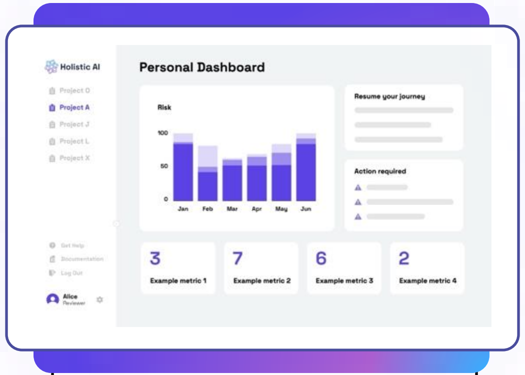
Safeguard

Governance and monitoring of Large Language Models.



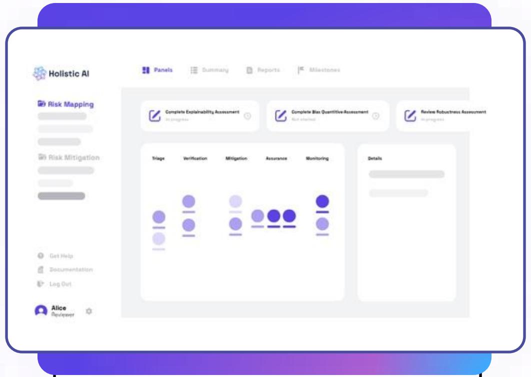
Tracker

All-in-one repository of all things related to AI regulations and harms.



AI Governance Platform

Enterprise-ready application for safe AI adoption and usage.



Bespoke AI Auditing

Third-party quantitative assessment and reporting for specific AI risks and use cases.

Agenda

- Setting the Stage for AI Governance
- Risks Associated with AI
- Bias and Fairness in AI
- Governing AI Bias

What is Artificial Intelligence?

Regulatory Definitions

Holistic AI

UK ICO definition:

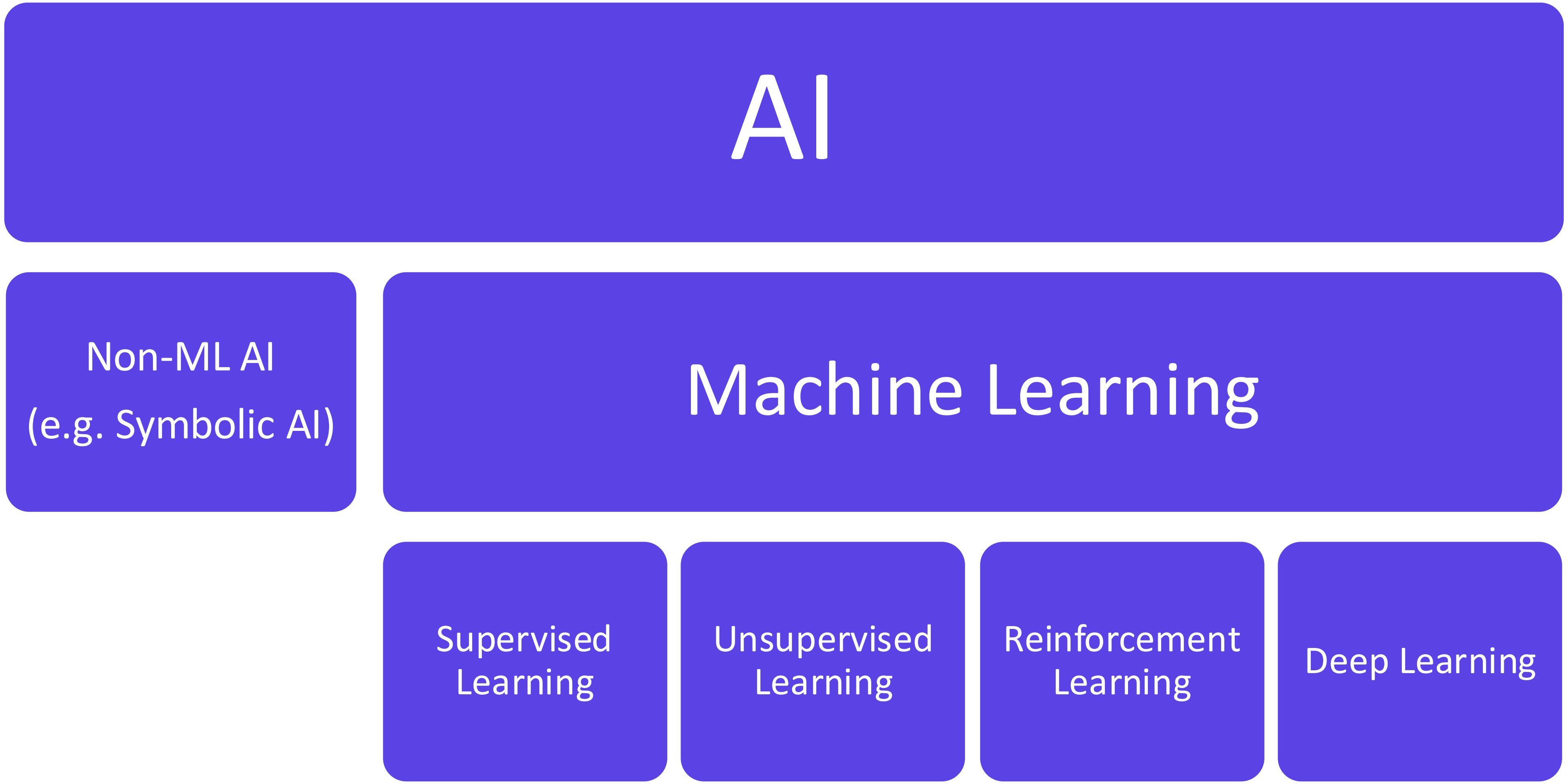
“AI is an umbrella term for a range of technologies and approaches that often attempt to mimic human thought to solve complex tasks”

Definition of AI systems

EU AI Act: “a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments”

ISO/IEC 22989: “engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives”

https://www.holisticai.com/blog/comparing-definitions-of-ai



AI

Non-ML AI

(e.g. Symbolic AI)

Machine Learning

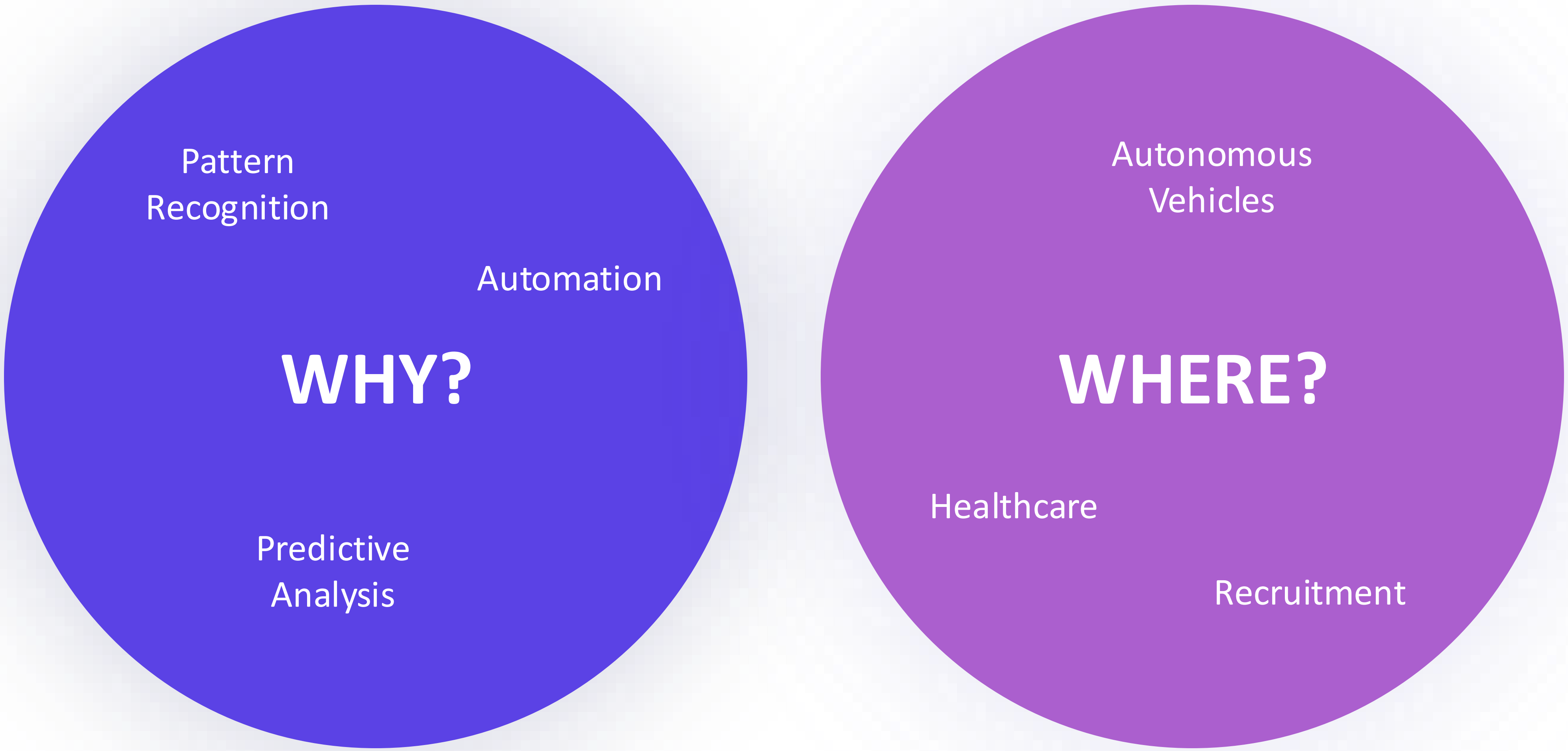
Supervised Learning

Unsupervised Learning

Reinforcement Learning

Deep Learning

Why and where is AI used?



WHY?

Pattern Recognition

Automation

Predictive Analysis

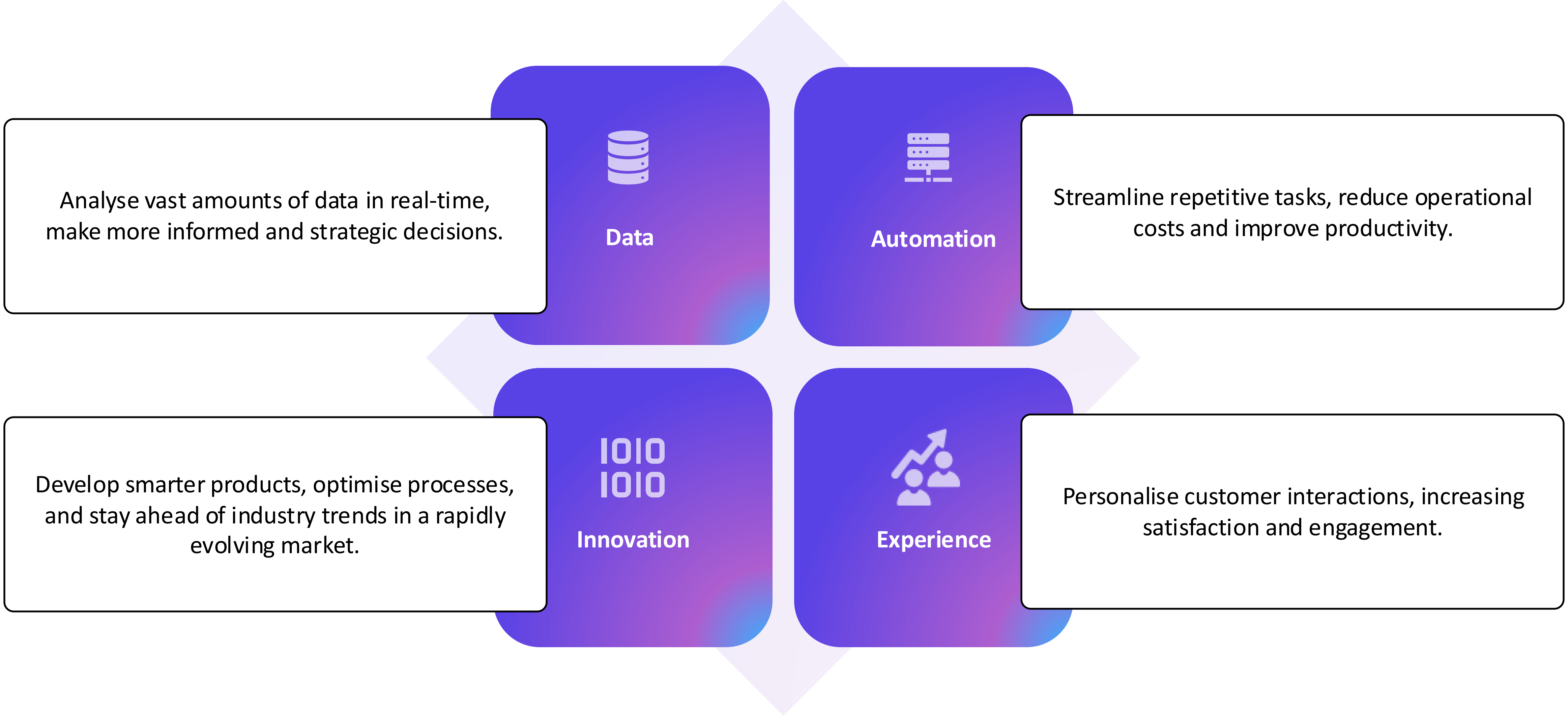
WHERE?

Autonomous Vehicles

Healthcare

Recruitment

Why are businesses adopting AI?



Data

Analyse vast amounts of data in real-time, make more informed and strategic decisions.

Automation

Streamline repetitive tasks, reduce operational costs and improve productivity.

Innovation

Develop smarter products, optimise processes, and stay ahead of industry trends in a rapidly evolving market.

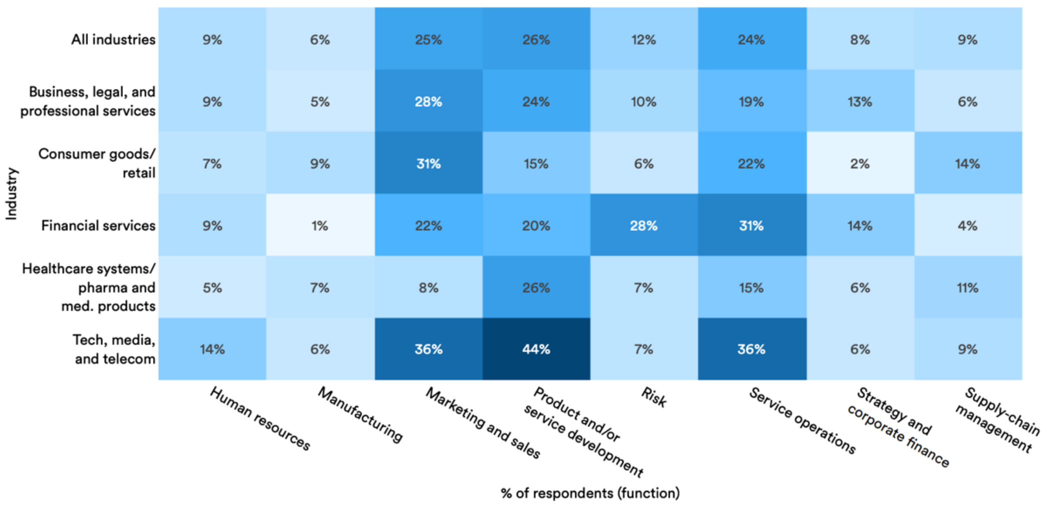
Experience

Personalise customer interactions, increasing satisfaction and engagement.

What is happening?

AI adoption by industry and function, 2023

Source: McKinsey & Company Survey, 2023 | Chart: 2024 AI Index report



Industry	Human resources	Manufacturing	Marketing and sales	Product and/or service development	Risk	Service operations	Strategy and corporate finance	Supply-chain management
All industries	9%	6%	25%	26%	12%	24%	8%	9%
Business, legal, and professional services	9%	5%	28%	24%	10%	19%	13%	6%
Consumer goods/ retail	7%	9%	31%	15%	6%	22%	2%	14%
Financial services	9%	1%	22%	20%	28%	31%	14%	4%
Healthcare systems/ pharma and med. products	5%	7%	8%	26%	7%	15%	6%	11%
Tech, media, and telecom	14%	6%	36%	44%	7%	36%	6%	9%

Emerging Regulations and Frameworks

Holistic AI

UK AI Regulation White Paper

EU AI Act

NIST AI RMF

ISO 42001

The Future

UK AI Market

is predicted to grow to over $1 trillion by 2035

Global GDP

is estimated to grow by $15.7 trillion by 2030

Explosion in AI vendors

OpenAI | ANTHROPIC | deepseek

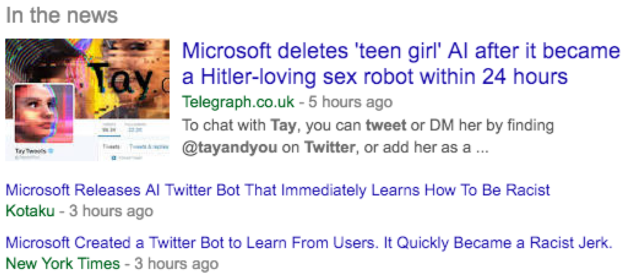
Risks Associated with AI

General set of AI risks

Holistic AI

Security and Privacy	Data minimization principles as well as adopt privacy-enhancing techniques to mitigate personal or critical data leakage.
Bias and Discrimination	Avoid unfair treatment of individuals given their protected characteristics.
Interpretability and Explainability	Provide decisions or suggestions that can be understood by their users and developers.
Performance and Robustness	Be safe and secure, not vulnerable to tampering or compromising of the data they are trained on.

To avoid these cases

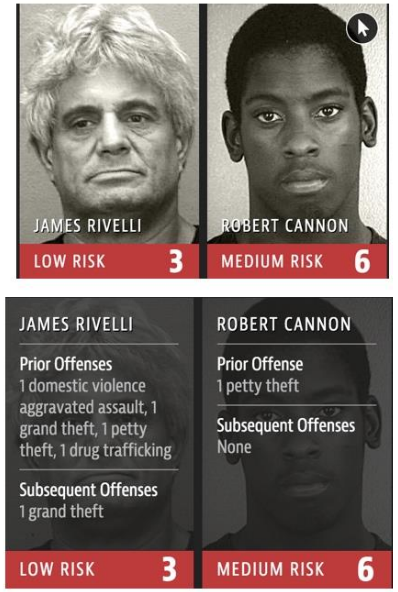


In the news

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours
Telegraph.co.uk - 5 hours ago
To chat with Tay, you can tweet or DM her by finding @tayandyou on Twitter, or add her as a ...

Microsoft Releases AI Twitter Bot That Immediately Learns How To Be Racist
Kotaku - 3 hours ago

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.
New York Times - 3 hours ago



JAMES RIVELLI
LOW RISK 3

ROBERT CANNON
MEDIUM RISK 6

JAMES RIVELLI
Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking
Subsequent Offenses
1 grand theft
LOW RISK 3

ROBERT CANNON
Prior Offense
1 petty theft
Subsequent Offenses
None
MEDIUM RISK 6

Why Bias is the Key Risk?

Reputational damage

Instances of biased AI garner significant public attention and can lead to public backlash against the technology, as well as increased scrutiny and regulation.

Challenging risk to deal with

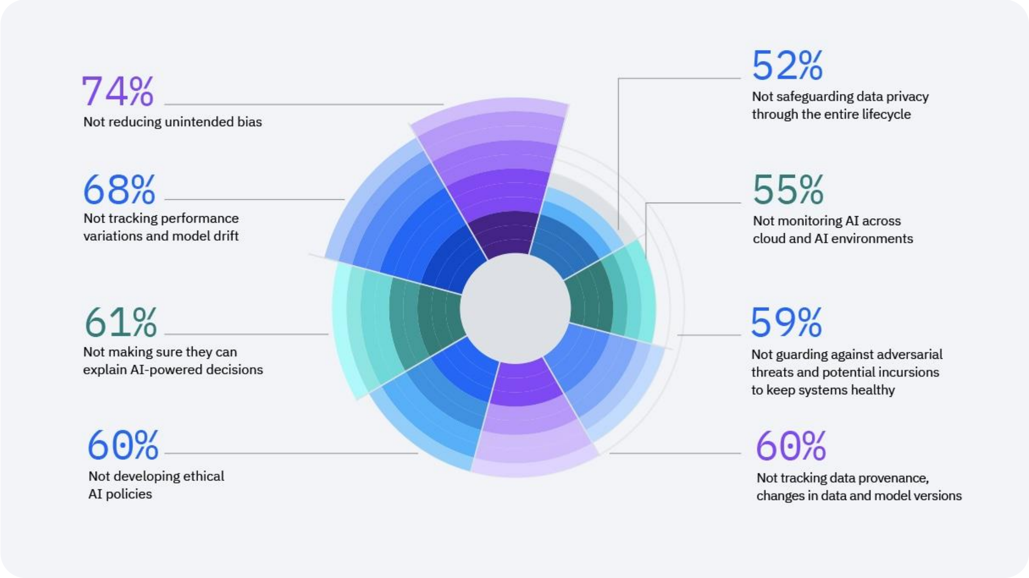
Compared to other key risks, Bias is undoubtedly the area where data scientists and other professionals are least prepared to deal

Regulatory requirements

Most laws and regulations are targeted towards bias, for example anti-discrimination laws.

General awareness

Anecdotally, Privacy is for Data what Bias is for AI



https://www.ibm.com/watson/resources/ai-adoption

Societal Implications of Bias in AI

Holistic AI

Historical Discrimination

Society's historical bias reflected in the data that we use to train algorithmic systems. will be perpetuated by the decision-making system.

Impact on Life Chances

Algorithmic systems are being used for decision making in a wide range of areas such as criminal justice, healthcare, education, employment, consumer lending, etc.

Responsibility

When developing decision-making systems, we are responsible for minimizing harm.

Business Implications of Bias in AI

Holistic AI

TIME

Aug 2022

“A Factual Error by Bard AI Chatbot Just Cost Google $100 Billion”

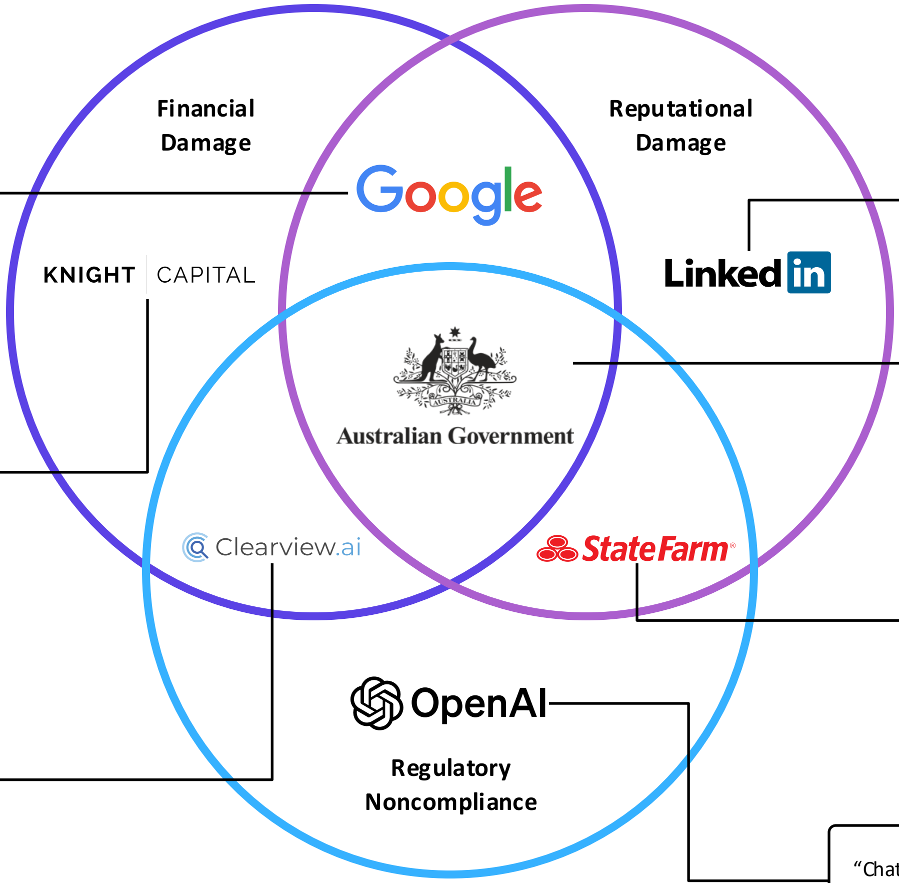
Aug 2022

“Knight Capital says Trading Glitch Cost It $440 Million”

F

May 2022

“Clearview AI Fined $9.4 Million in U.K. For Illegal Facial Recognition Database”



MIT Technology Review

Jun 2021

“LinkedIn’s Job-Matching AI was Biased. The Company’s Solution? More AI”

REUTERS

Aug 2022

“Australia Forced to Pay $1.2bn in Compensation for Unlawful Automated Debt Recovery Scheme”

Dec 2022

“State Farm Hit with Massive Racial Discrimination Lawsuit Automated System that Allegedly Treats Black Homeowners with Extra Suspicion”

BBC

Apr 2023

“ChatGPT Banned in Italy over Privacy Concerns”

Risks and issues with HR technologies

With the pervasive usage, side-effects such as bias, transparency, cheating will increase

Holistic AI

Tutoring firm settles US agency's first bias lawsuit involving AI software

EEOC Settles First AI-Bias Case, but Many More Are Likely

Amazon scrapped 'sexiest AI' tool

Lawsuit Claims Workday's AI And Screening Tools Discriminate Against Those Black, Disabled Or Over Age 40

HireVue drops facial monitoring amid A.I. algorithm audit

Candidates using ChatGPT to cheat their way into a job

Real-World Examples of Bias in AI

ML can be biased

Amazon Recruitment

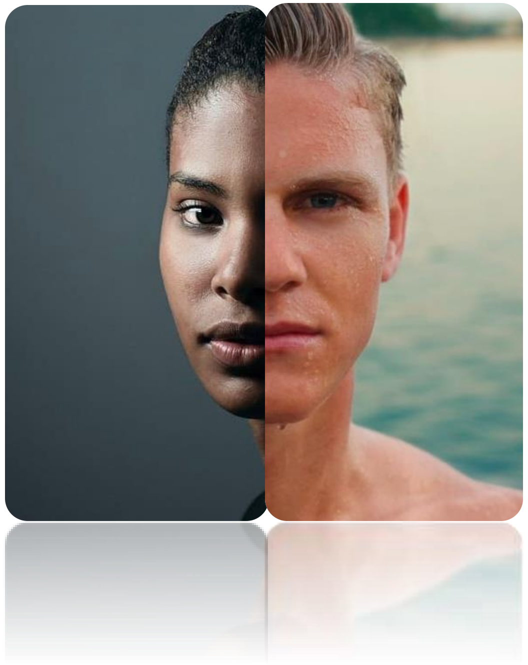
RESUME

2023: AWESOME

2020: WOMEN’S COLLEGE

Gender Recognition (“Gender Shades”)

Up to ~35% error rate



Up to ~1% error rate

The COMPAS Example

- Used in US court to predict risk of recidivism

- ProPublica study in 2016:

	White	Black
Labelled higher risk – but didn’t re-offend	23%	45%
Labelled low risk – did re-offend	48%	28%

- Northpointe defence: accuracy for white and black is the same (~60 %)

GenAI

- UNESCO study reveal that popular LLMs like GPT3.5 and Llama 2 exhibit bias against women in its outputs.
  - Richer narratives in stories about men
  - Homophobic attitudes and racial stereotyping

Bias and Fairness in AI

What is Bias?

Global definition

Bias is a tendency to **favour one group or idea over others.** It can be based on a person’s race, gender, religion, etc.

Machine Learning

Systematic errors in the decision-making process of an algorithm that lead to inaccurate or unfair outcomes.

What is Fairness?

Individual Fairness

Individual Fairness: Seeks for similar individuals to be treated similarly

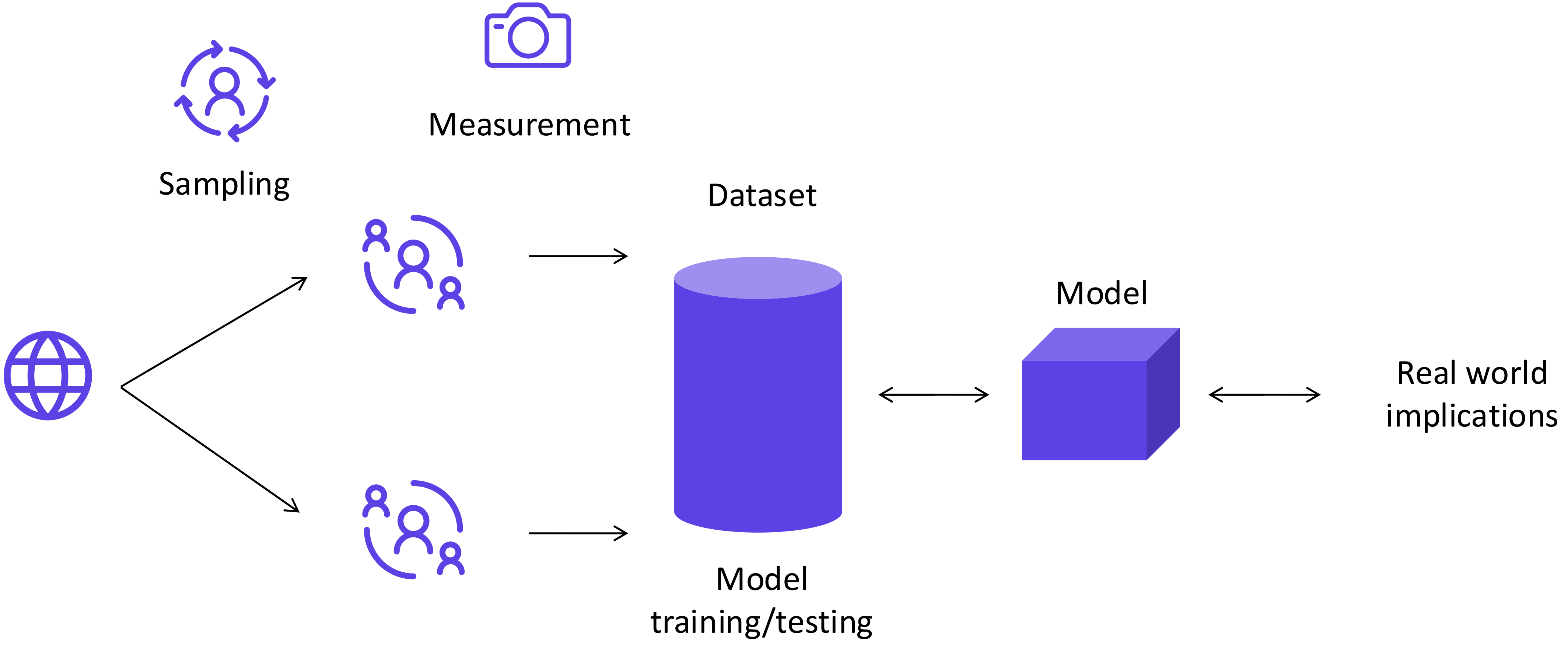
- “Am I as an individual being treated as another with similar qualifications?”
  - Fairness through awareness vs. Fairness through unawareness
  - Counterfactual fairness

Group Fairness

Group Fairness: Split a population into groups defined by protected attributes (e.g. male, female) and seeks for some measure to be as equal as possible across groups

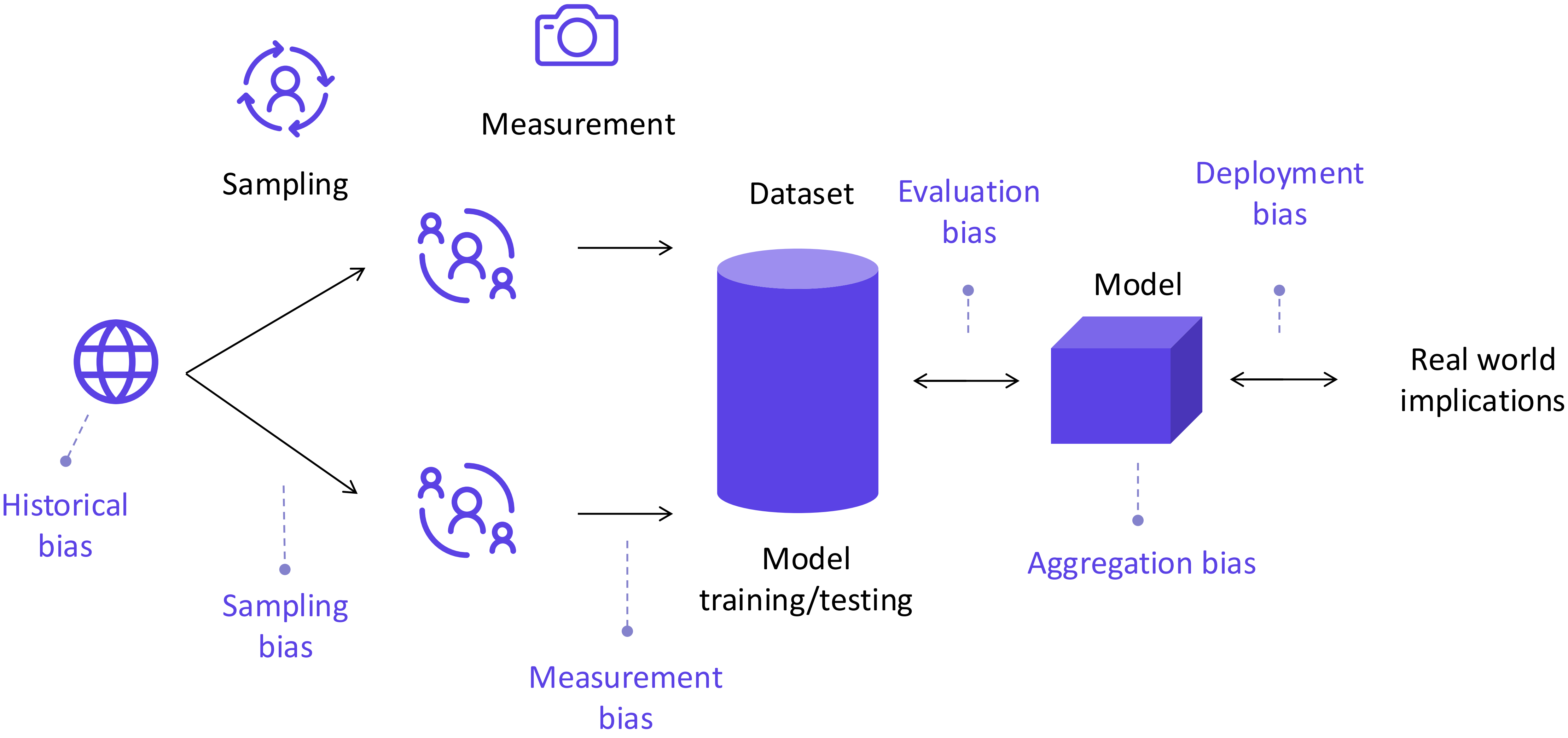
- “Is the minority group to which I belong being treated as the majority group?”
  - Equality of Outcome: Distribution of success across all groups should be the same
  - Equality of Opportunity: Distribution of success across groups with similar qualifications should be the same

Machine Learning lifecycle



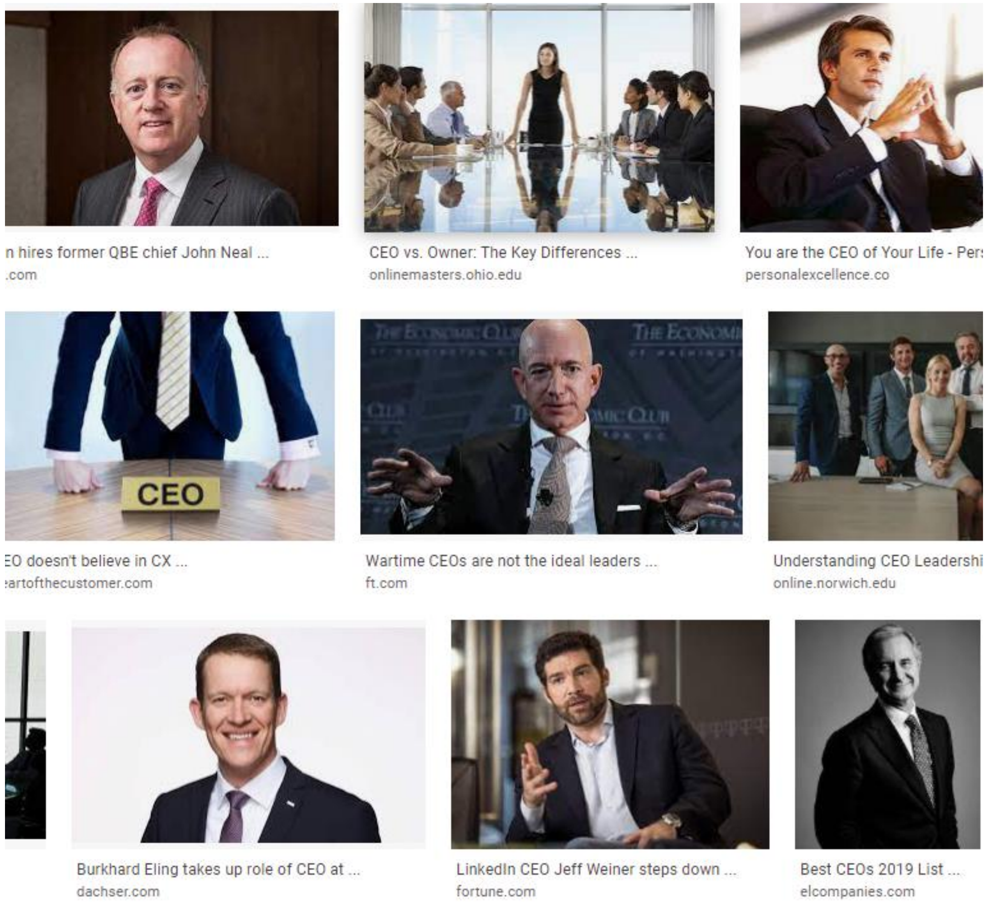
Sources of Bias

Holistic AI



Historical Bias

- Pre-existing bias reflected in the data
- Example: In 2018, 5% of Fortune 500 CEOs were women.
- What should be the results of an image search for ‘CEO’? Google recently changed image search result to include higher proportion of women



n hires former QBE chief John Neal ...
.com
CEO vs. Owner: The Key Differences ...
onlinemasters.ohio.edu
You are the CEO of Your Life - Per
personalexcellence.co
CEO
EO doesn't believe in CX ...
eartofthecustomer.com
Wartime CEOs are not the ideal leaders ...
ft.com
Understanding CEO Leadershi
online.norwich.edu
Burkhard Eling takes up role of CEO at ...
dachser.com
LinkedIn CEO Jeff Weiner steps down ...
fortune.com
Best CEOs 2019 List ...
elcompanies.com

Deployment bias

- Model was trained and tested for use case X, but is used for application Y
- Example: Model was trained to identify good candidates for investment banking jobs, and now being used to identify junior level lawyers

Governing AI Bias

AI Governance

- Refers to the rules, policies, and frameworks that guide the development and use of AI, ensuring it aligns with ethical, legal, and societal standards.
    - Risk Management
    - Ensuring Trust and Accountability

Why is AI Governance Critical?

Managing Risks

Allows for the measuring, mitigating, and monitoring, of risks that are most relevant to the AI use case.

Trust and Accountability

Ensures trustworthy and responsible adoption, use, and scaling of AI within organisations.

Global AI Regulations and Governance Frameworks

Holistic AI

EU AI Act

Regulates AI based on risk, with strict requirements for high-risk AI systems, ensuring safety, transparency, and accountability. Designed to enhance trust in AI technologies and promote innovation while safeguarding fundamental rights.

ISO/IEC 42001

A standard for AI Management Systems that provides guidelines on managing AI systems' lifecycle, focusing on transparency, accountability, and ethical use. Promotes risk management and ensuring AI systems operate safely and in alignment with societal values.

NIST AI RMF

A voluntary guideline aimed at fostering trustworthiness in AI systems. Provides a framework for organizations to manage risks related to AI, focusing on fairness, transparency, privacy, and robustness.

UK AI Strategy

A flexible, principles-based approach to AI regulation, emphasizing innovation while ensuring that AI systems are aligned with ethical standards. Focuses on safety, transparency, and accountability, and aims to support both innovation and consumer protection.

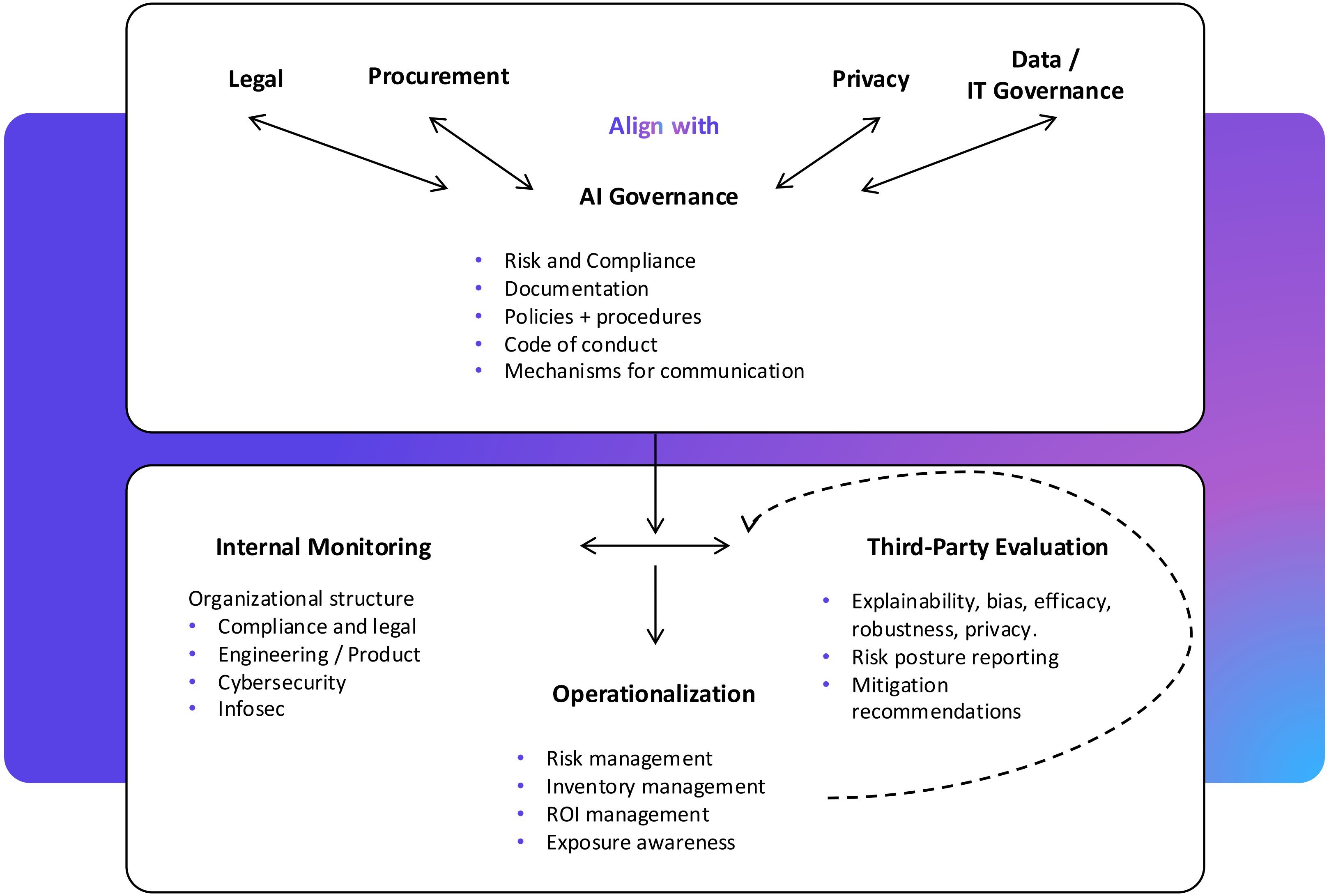
The UK vs. EU vs. US

Holistic AI

	UK	EU	US
Regulatory Approach	Principles-based, flexible, risk-based approach. Focus on fostering innovation while ensuring ethical AI.	Comprehensive, legally binding framework with specific risk categories.	Decentralized, industry-driven, with voluntary guidelines and sector-specific regulations.
Key Frameworks/Legislation	• UK AI Strategy (2021) • AI Regulatory Framework (work in progress) • AI and Data Bill (proposed)	• EU AI Act (AIA) • General Data Protection Regulation (GDPR) • Digital Services Act (DSA)	• National AI Initiative Act (2020) • Algorithmic Accountability Act (proposed)
Risk Classification	Risk-based but not as formalized as the EU's approach. Focus on high-risk AI sectors like healthcare and finance.	Clear classification of AI systems into high-risk, limited-risk, and minimal-risk categories.	Risk-based approach in some proposals (e.g., Algorithmic Accountability Act) but generally lacks formalized risk categories at the federal level.
Focus Areas	Safety, transparency, ethical AI, fostering innovation, and sector-specific regulation.	Safety, transparency, accountability, protecting fundamental rights, human oversight, and addressing AI's social impact.	Ethical AI, transparency, fairness, privacy, consumer protection, and AI's societal impacts.
Legal Enforceability	Not fully enforced yet, but proposals for legal regulation are underway.	Legally binding with penalties for non-compliance, particularly for high-risk AI systems.	Currently voluntary frameworks like NIST AI RMF; no comprehensive AI law yet at the federal level.

Adopting and Scaling AI Responsibly

Holistic AI



Legal

Procurement

Align with

Privacy

Data / IT Governance

AI Governance

- Risk and Compliance
- Documentation
- Policies + procedures
- Code of conduct
- Mechanisms for communication

Internal Monitoring

Organizational structure

- Compliance and legal
- Engineering / Product
- Cybersecurity
- Infosec

Third-Party Evaluation

- Explainability, bias, efficacy, robustness, privacy.
- Risk posture reporting
- Mitigation recommendations

Operationalization

- Risk management
- Inventory management
- ROI management
- Exposure awareness

AI Risk Appetite and Posture

AI Risk Appetite

The level of risk your business is willing to accept in adopting AI technologies.

- Ensures trustworthy use and responsible scaling of AI within organisations.
    - *What level of risk are we willing to accept in adopting AI technologies to drive innovation and competitive advantage?*

AI Risk Posture

The strategic approach a business takes in managing and responding to risks.

- Allows for the measuring, mitigating, and monitoring, of risks.
    - *What strategies and resources do we need to put in place to manage AI risks effectively and ensure long-term sustainability?*

Considerations and Trade-Offs

- **Innovation:** Balancing rapid AI development with the need to manage risks like bias and security.
  - Faster innovation at the risk of ethical or security issues.
- **Adoption:** Fast-tracking AI adoption vs. ensuring compliance with ethical and regulatory guidelines.
  - Slower AI adoption with reduced exposure to risks and compliance issues.
- **Data:** Using personal data to improve AI performance while protecting individual privacy.
  - Better AI outcomes vs. privacy concerns.
- **Flexibility:** Meeting regulations while maintaining flexibility for AI innovation.
  - Enhanced risk control at the expense of operational speed.

Trade-off Example

Holistic AI

A company is introducing an AI-driven recruitment platform to streamline its hiring process. The platform is designed to assess resumes, conduct initial candidate screenings, and even provide interview recommendations based on historical hiring data. While the system is meant to enhance efficiency and reduce human bias, concerns arise about whether it inadvertently perpetuates biases, particularly around gender, race, and socioeconomic background.

Trade-off Example

- **Innovation vs. Bias Management**
  - Faster adoption improves efficiency but risks reinforcing historical biases.
  - Slower deployment allows for thorough bias auditing and mitigation but delays benefits.
- **Faster Innovation vs. Ethical Issues**
  - Rapid rollout achieves quick results and cost savings but may raise ethical concerns like discrimination.
  - A slower rollout ensures compliance with laws but may lead to missed opportunities or delays.
- **Personal Data vs. Privacy Concerns**
  - More data improves model accuracy but increases privacy and consent risks.
  - Using limited, publicly available datasets protects privacy but may reduce performance.
- **Regulatory Compliance vs. Flexibility for Innovation**
  - Strict compliance minimises legal risks but slows innovation.
  - Greater flexibility enables faster progress but increases regulatory exposure.

Holistic AI

🌐 holisticai.com ✉ we@holisticai.com

The contents of this presentation do not serve as legal counsel or an authoritative viewpoint. They should not be employed as a solitary source for settling legal disputes or managing litigations. While entities engaged in ongoing disputes may find value in comprehending the discussions and their legal backdrop, this information is not a replacement for legal advice. The provided material is strictly for informational purposes. All rights reserved. Unauthorized copying, reproduction, distribution, transmission, modification, creation of derivative works, or any other form of exploitation of the copyrighted material is prohibited without prior written consent from Holistic AI.