Trust, Ethics in AI

By-
Pankaj Kumar Madhukar
(GM-AI)
VFS Global

Division of AI

01 **Artificial Intelligence**
It involves techniques that equip machine to emulate human behaviour, **recognize patterns, solve complex problem**.

02 **Machine Learning**
ML is subset of AI, uses algorithm **to detect patterns in large data sets**. It uses supervised/unsupervised learning methods

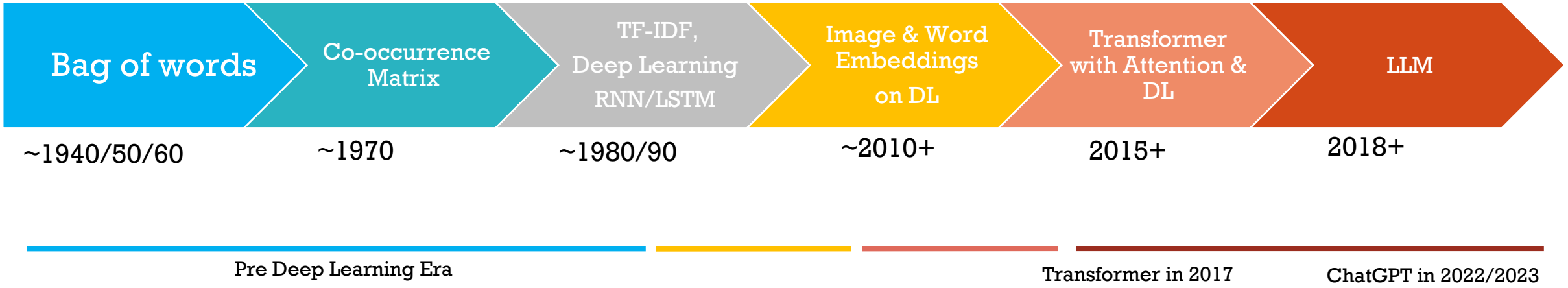
03 **Deep Learning**
DL is subset of ML, **uses neural networks** for in-depth data processing.

04 **Generative AI**
It is **subset of DL** models

What is GEN AI

Neural Network based Models

GEN AI evolution



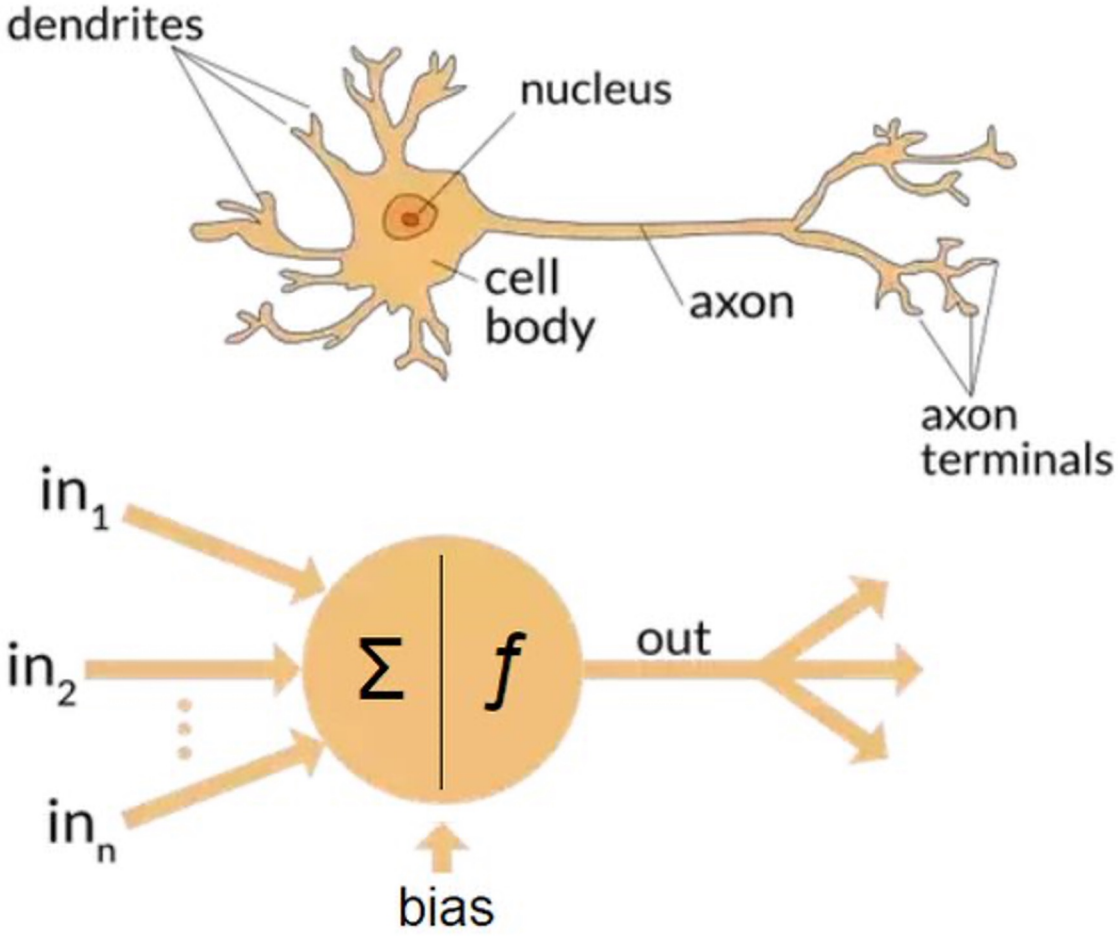
What is LLM?

- Large language models (LLM) are very **large deep learning models** that are pre-trained on vast amounts of data.
- The underlying **transformer is a set of neural networks** that consist of an encoder and a decoder with self-attention capabilities.
- During training, the model learns to **predict the next word or sequence of words** based on the input it receives.

LLM capabilities

- ★ Text Processing (NLU, NLG) & Summarization
- ★ Conversational Chat
  - ○ Email Reply/ Social Media comment reply, Q&A
- ★ Image Feature Identification & Generation
- ★ Voice Processing
  - ○ Call Recording Processing
  - ○ Voice recognition (dialect handling) and synthesis (vernacular)
  - ○ Neural Voice (human voice)
- ★ Multi Language with vernacular support
  - ○ Translation – Hinglish, English, Hindi, Marathi, Spanish etc
- ★ Code Generation
- ★ Copilot for productivity
- ★ Compliance
  - ○ Customer said “Not Interested”
  - ○ Object identification

Human Brain vs. Deep Learning



Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

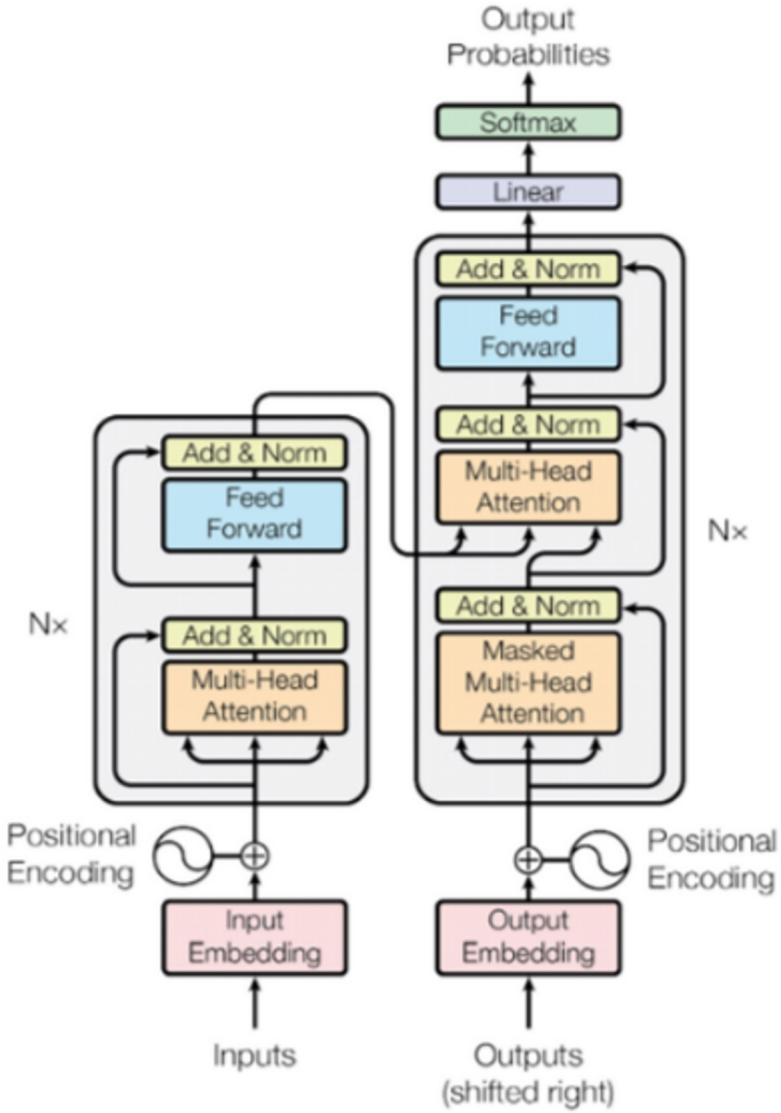
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



Google Map

Privacy Concerns: Google Maps collects vast amounts of location data, which raises issues regarding user privacy and data security. Users may be unaware of how their data is used or shared.

Bias in Routing: AI algorithms may favor certain routes based on historical traffic data.

OLA

Surge Pricing: OLA employs dynamic pricing models.

Data Privacy: OLA collects sensitive data (e.g., location, payment information), raising concerns about how this data is stored, used, and shared.

Zomato

Algorithmic Bias: The recommendation system may favor certain restaurants based on previous user behavior, which can marginalize small, local eateries, leading to reduced visibility and sales.

Labor Rights of Delivery Personnel: The AI-driven performance metrics used to assess delivery personnel can pressure workers, impacting their income and job security, often leading to exploitative working conditions.

Amazon

Algorithmic Bias in Recommendations: Amazon's recommendation system may inadvertently reinforce biases, promoting products that align with historical purchasing data, potentially disadvantaging lesser-known brands.

IVR

User Frustration: Poorly designed IVR systems can lead to user frustration

Bias in Speech Recognition: AI models used in IVR may struggle with diverse accents and dialects.

A father and son are in a horrible car crash that kills the dad. The son is rushed to the hospital; just as he's about to go under the knife,
the surgeon says, "I can't operate — that boy is my son!"

Who is the Surgeon?

Who is the Surgeon?

The surgeon is the boy's mother

Trust in AI

Trust depends on

- Ethics
- Explainability

Ethics in AI

Ethics in AI addresses the **moral** and **societal implications** of AI technologies.

These principles ensure that AI systems are designed and used in ways that respect human rights, promote fairness, and avoid harm.

Ethical Principles

1. Fairness

AI systems **should be fair** and **not discriminate against individuals or groups** based on characteristics such as race, gender, age, ethnicity, or socioeconomic status.

- **Why it Matters**: AI can perpetuate or exacerbate existing biases if trained on biased data. Ensuring fairness prevents discrimination and promotes equality.

- **Example**: In AI-based hiring systems, fairness ensures that all candidates are evaluated equally, regardless of gender or ethnicity.

Ethical Principles

2. Transparency

AI systems **should operate in a transparent** and **explainable manner**, providing insights into how decisions are made.

- **Why it Matters**: Transparency helps build trust in AI systems, especially in high-stakes areas like healthcare or criminal justice.
- **Example**: An AI medical diagnosis system should explain the reasoning behind its recommendation, helping doctors and patients understand the factors influencing its decision.

Ethical Principles

3. Accountability

There should be **clear accountability for the actions** and **outcomes** of AI systems. Developers, users, or organizations must be responsible for ensuring that AI behaves ethically.

- **Why it Matters**: If an AI system causes harm or makes incorrect decisions, someone needs to be held accountable for rectifying the issue.

- **Example**: If a self-driving car causes an accident, there should be a clear process to determine responsibility (e.g., the car manufacturer, software developer, or operator).

Ethical Principles

4. Privacy

AI systems must respect the **privacy of individuals** by safeguarding personal data and using it only in ethical and authorized ways.

- **Why it Matters**: AI systems often rely on vast amounts of personal data, raising concerns about unauthorized data use or surveillance.

- **Example**: AI in healthcare should ensure that patient data is anonymized and used only for legitimate medical purposes, respecting patient consent.

Ethical Principles

5. Beneficence and Non-Maleficence

AI should be used to **benefit society and individuals**, while avoiding harm or negative consequences.

- **Why it Matters**: AI systems can cause unintended harm if not carefully designed, so developers must ensure that systems promote good and minimize risks.
- **Example**: In autonomous weapons, ethical AI design would prioritize preventing unnecessary harm to civilians and ensuring compliance with international humanitarian law.

Ethical Principles

6. Autonomy

AI systems should **respect human autonomy**, giving people control over how AI affects their lives. Humans should be able to make informed decisions about interacting with AI systems.

- **Why it Matters**: AI should enhance human decision-making, not replace it or make decisions without human oversight.

- **Example**: In healthcare, AI should assist doctors by providing recommendations, but the final decision on treatment should rest with the physician and patient.

Ethical Principles

7. Robustness and Safety

AI systems **must be reliable, safe, and secure**. They should be designed to handle unexpected situations without causing harm or malfunctioning.

- **Why it Matters**: Faulty or insecure AI systems can lead to dangerous consequences, especially in critical areas like autonomous driving, healthcare, or finance.

- **Example**: An autonomous drone should be robust enough to avoid crashes or failures due to environmental changes or technical malfunctions.

Ethical Principles

8. Inclusiveness

AI should be developed and used in ways that **promote inclusivity** and ensure that all individuals and communities can benefit from technological advancements.

- **Why it Matters**: AI should not only cater to certain groups but should be designed to be inclusive of different demographics, languages, and cultural backgrounds.

- **Example**: AI-driven education platforms should accommodate students with diverse learning styles, abilities, and needs to ensure equitable access to education.

Ethical Principles

9. Sustainability

AI development and usage **should consider the environmental** and **social impact**, ensuring that it supports sustainable development goals and minimizes harm to the environment.

- **Why it Matters**: AI systems, especially large-scale models, consume significant energy, contributing to environmental degradation.

- **Example**: AI systems should be optimized for energy efficiency, reducing the carbon footprint associated with training large neural networks.

India's Economic Survey – Clause 6.8

"Even as developed nations prepare to impose a carbon tax at the border on imports coming into their countries laden with carbon, they are ramping up energy demand like never before, thanks to their obsession with letting Artificial Intelligence (AI) guide, take over and dominate natural intelligence**. One of the leading global technology companies promised to achieve Net Zero by 2030** at the turn of the decade. But, the race to dominate the emerging technology of Artificial Intelligence has caused its emissions to be higher by 30 per cent by 2023. "

Ethical Principles

10. Justice and Equity

AI should promote justice and equity, ensuring that its **benefits are shared broadly across society** and do not disproportionately benefit or harm certain groups.

- **Why it Matters**: AI technologies should not exacerbate existing inequalities or create new forms of injustice.
- **Example**: In financial services, AI algorithms should ensure equitable access to credit, not reinforcing economic disparities through biased credit scoring models.

Responsible AI

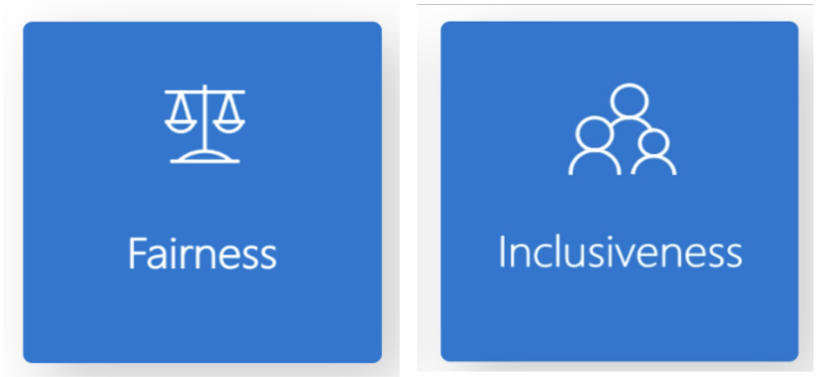
Responsible AI aims to **embed such ethical principles into AI applications** and **workflows** to **mitigate risks and negative outcomes** associated with the use of AI, while maximizing positive outcomes.

Responsible AI serves as the **foundation** of the trust by addressing critical concerns.

Responsible AI is an **approach** to **developing**, **assessing**, and **deploying** AI systems in a safe, trustworthy, and ethical way.

Responsible AI Principles

Responsible AI Principles



- Treat everyone fairly
- Avoid bias
- Eg. Loan Application

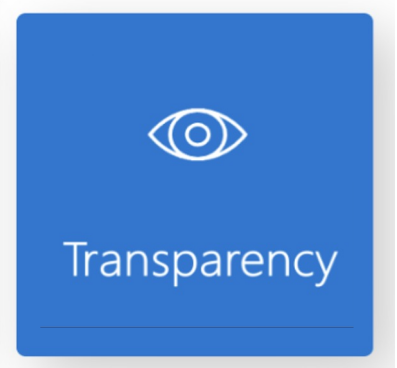
- Bias Detection in Data
- Fairness Metrics
- Regular Audits and Updates
- User-Centric Design

Responsible AI Principles

- AI systems should operate reliably, safely, and consistently
- Respond safely to unanticipated conditions
- Resist harmful manipulation

- Continuous Monitoring
- Security Measures
- Robust System Design
- Interdisciplinary Collaboration

Responsible AI Principles



- AI systems must take inform decisions
- Interpretability: The useful explanation of the behavior of AI systems and their components

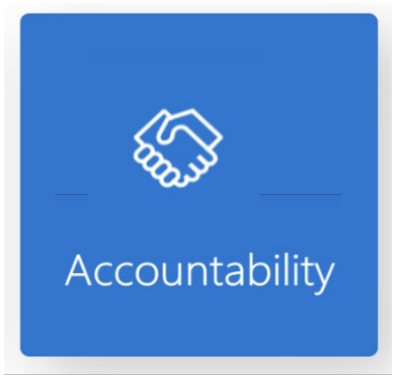
- Clear Documentation
- Explainable AI (XAI) Techniques
- Feedback Mechanisms
- Audit Trails and Logging

Responsible AI Principles

- Privacy protection
- Securing personal and business information
- Require transparency about the collection, use, and storage of data

- Restrict access to resources and operations by user account or group.
- Restrict incoming and outgoing network communications.
- Encrypt data in transit and at rest along with Data masking.
- Scan for vulnerabilities.
- Apply and audit configuration policies.

Responsible AI Principles



- The people who design and deploy AI systems must be accountable

- Clear Roles and Responsibilities in the Team
- Feedback and Reporting Mechanisms
- Impact Assessments
- Audit Trails and Logging

Relationship Between Ethical AI and Responsible AI

- **Ethical AI** lays the **theoretical foundation** by defining the moral and philosophical guidelines AI systems should follow.

- **Responsible AI** focuses on the **practical execution** of these ethical guidelines by putting in place mechanisms, policies, and accountability measures to ensure that AI operates according to those principles in the real world.

Trust and Ethics in Financial Systems:

A Real-World Application of AI and the Role of Explainable AI

Application

Scenario:

Ethical Issues ?

How Explainable AI Helps ?

Next Steps for Trust and Ethics?

Credit Scoring and Loan Approvals

Scenario: You have developed an AI model which automate the process of evaluating loan applications. The model used large datasets, including credit history, income levels, employment records, and spending habits, to assess an applicant's creditworthiness.

Model deny a loan application of the applicant who lives in a neighborhood.

Ethical Issues: Bias, Lack of Transparency, Accountability

How Explainable AI Helps: It can help ensure **fair and transparent decision-making** by providing clear insights into how and **why a particular decision was made**. For example, if a loan application is denied, It can highlight the specific factors (e.g., credit score, income) that led to the denial, allowing the applicant to understand the decision. Moreover, it can help developers and regulators detect unintended biases in the AI model and take corrective action

Next Steps for Trust and Ethics: Bias Auditing, Regulatory Compliance, User Education

Fraud Detection

Scenario: You have developed an AI system to monitor transactions in real-time, flagging unusual activities that may indicate fraud.

For instance, if an individual's credit card is used in two different countries within a short time span, an AI system might automatically block the transaction.

Ethical Issues: Lack of Transparency, Responsibility, Trust Erosion

How Explainable AI Helps: It can provide clarity about **why certain transactions are flagged** as fraudulent. By offering insights into the specific factors (e.g., geographic location, transaction size, or frequency) that triggered the fraud alert, customers can understand why their transaction was blocked. Moreover, It can help banks fine-tune their models by analyzing false positives and adjusting the system to minimize these errors.

Next Steps for Trust and Ethics: Customer Communication, Human Oversight, Ongoing Model Refinement

Customer Service Chatbots for Financial Advice

Scenario: You have developed AI-powered chatbot to offer financial advice or answer customer queries

Chatbot involves sensitive decisions like investment strategies or loan recommendations based on his profile.

Ethical Issues: Transparency, Accountability, Misleading Advice

How Explainable AI Helps: Explainable AI can clarify the **rationale behind the advice provided** by AI-driven chatbots. By explaining the factors considered in giving a specific recommendation, It can help users understand the basis of the advice, whether it is based on the customer's financial history, market conditions, or other relevant factors. This builds trust and allows users to make more informed decisions.

Next Steps for Trust and Ethics: Human Oversight, Ethical AI Use

Facial Recognition Systems

Scenario: Facial recognition technology is widely used for security purposes, such as unlocking smartphones or identifying suspects in public surveillance.

Facial recognition developed by you often misidentify people of color and women at higher rates.

Ethical Issues: Bias and Discrimination, Privacy, Trust

How Explainable AI Helps: It can help **identify why a facial recognition system is making mistakes** or biased predictions by revealing **which facial features are being weighted heavily in identification**. This allows developers to detect and address biases, ensuring that the system performs fairly across different demographic groups.

Next Steps for Trust and Ethics: Algorithmic Auditing, Regulatory Oversight, Bias Mitigation

Responsible AI in Action

INDIA

National AI Strategy (NITI Aayog 2018):

- **Key Focus**: Responsible AI for inclusive growth. Focuses on ethical AI use in key sectors like agriculture, healthcare, and education to address social challenges.
- **Guiding Principles**: Fairness, transparency, and explainability in AI systems, with emphasis on solving local challenges.
- **Applicability**: Aimed at both public and private sectors, with a focus on scaling ethical AI practices in a large, diverse population.

European Union

AI Act (proposed in 2021):

- **Key Focus**: A risk-based regulatory approach classifying AI applications into high-risk, limited-risk, and minimal-risk categories.
- **High-Risk AI**: AI in sectors like healthcare, finance, and law enforcement must comply with strict requirements for transparency, accountability, and fairness.
- **Applicability**: Across all EU member states; applies to any AI system impacting European citizens, even if developed outside the EU.

USA

AI Bill of Rights (Blueprint released in 2022):

- **Key Focus**: Outlines five principles, including data privacy, algorithmic discrimination protections, and ensuring transparency and explainability.
- **Industry-Specific**: No overarching federal AI law, but regulatory bodies like the FTC and NIST play a role in shaping responsible AI practices, particularly around consumer protection and cybersecurity.
- **Applicability**: Sector-based (e.g., finance, healthcare); different states (like California) are developing their own AI-specific regulations.

AUSTRALIA

AI Ethics Framework (launched in 2019):

- **Key Focus**: Eight core ethical principles including transparency, fairness, privacy, and accountability in AI systems.
- **Human Rights and Fairness**: Strong emphasis on aligning AI with human rights principles and minimizing bias.
- **Applicability**: Public and private sectors, with applications in sectors like defense, healthcare, and energy.

India's Economic Survey –Insights for AI

In Clause 4.66

According to research by NASSCOM, **India is an attractive destination for AI investment** due to its relatively low operating costs and the world's second-largest pool of highly skilled AI, machine learning, and big data workers.

In Clause 11.2

Studies suggest that the application of Artificial Intelligence (AI) is likely to restrain the growth opportunities for business services progressively and, therefore, poses a challenge to long-term sustainability and job creation. Thus, **focusing on human capital to take advantage of the agglomeration effects of large**, well-functioning cities is critical for the growth of services, especially those with global market potential.

In Clause 2.133

The next big step in the coming years is likely to be towards Artificial Intelligence/ Machine Learning (AI/ML), Decentralised Finance, Internet of Things (IoT), etc., which have a vast potential to disrupt the digital payments ecosystem. Further, **the vision is for India to evolve as a 'fintech nation'** with the highest number of fintech firms and the highest fintech adoption rate by incumbents fuelled by digital public infrastructure.

QnA

Thank You