Building Trust in AI: Ethics and Governance

By

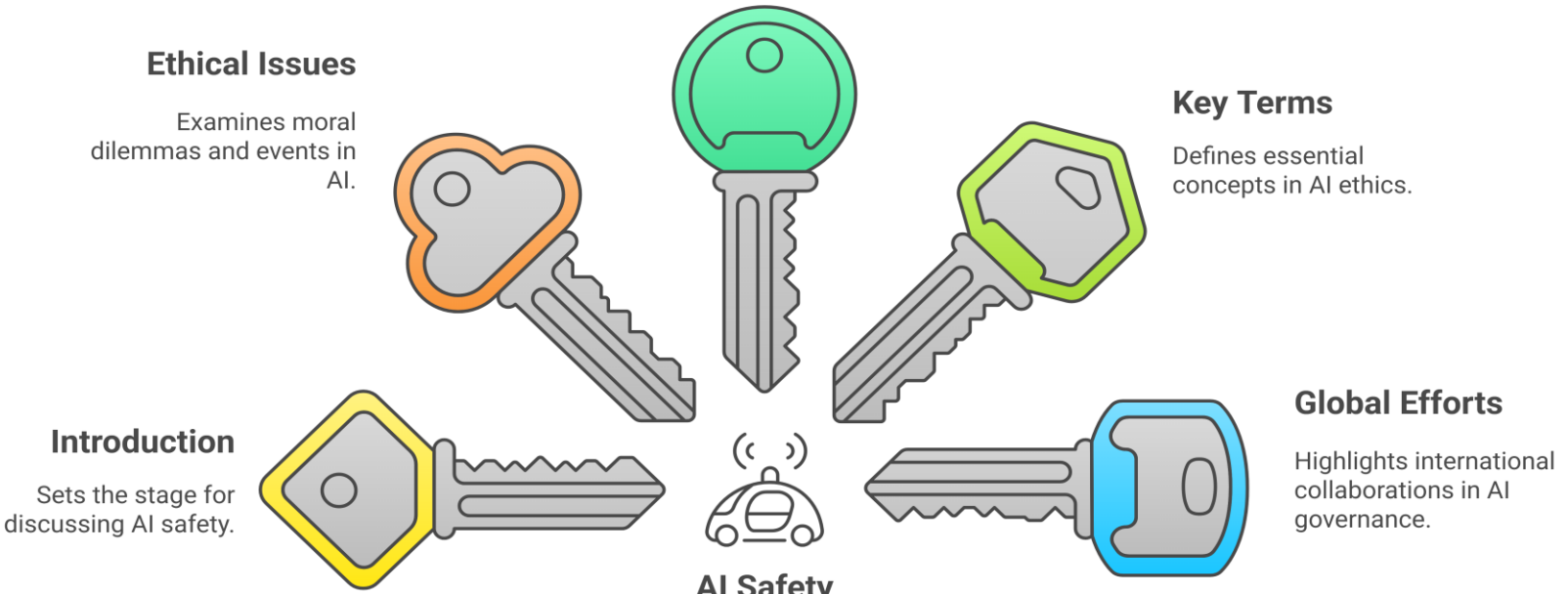
Dr.Hamad Alawad

MOI- KSA

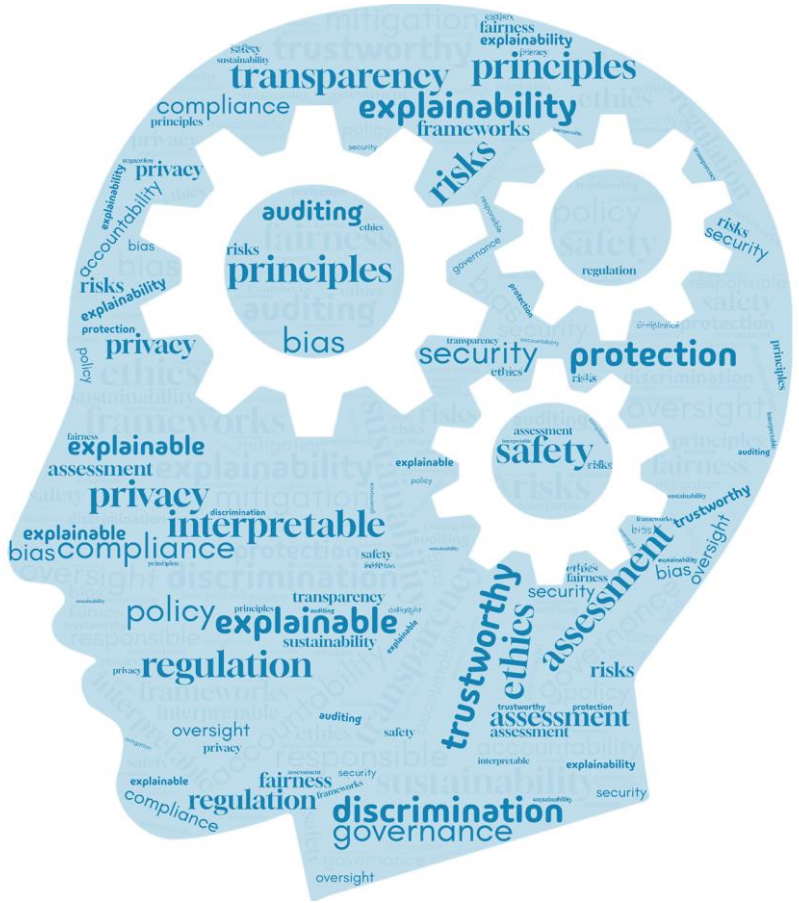
February 2025

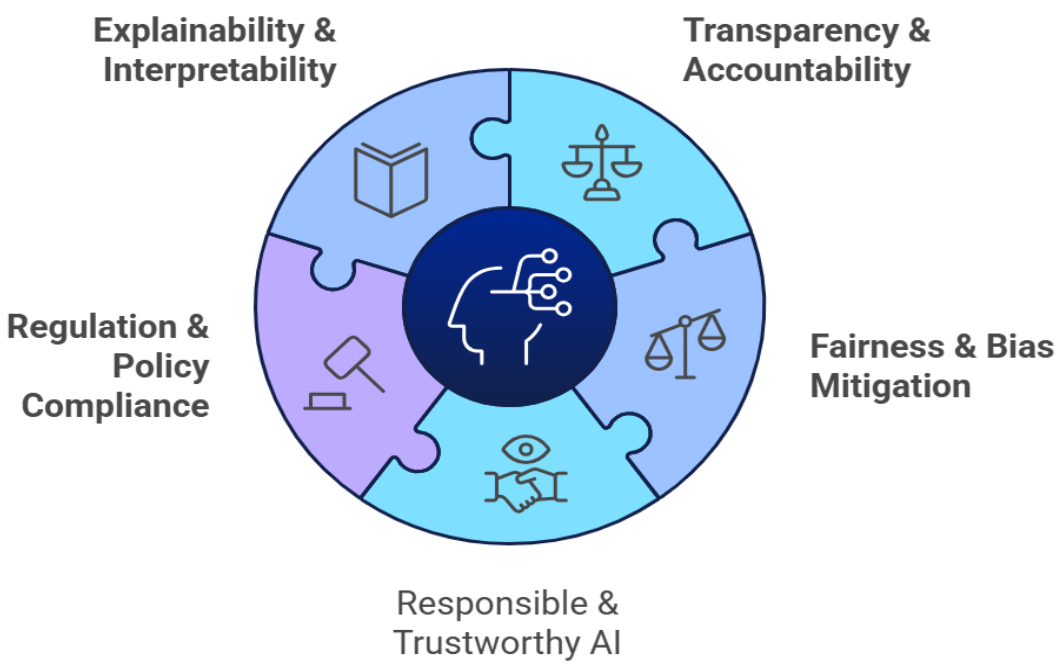
AGENDA

Navigating AI Safety



Key Terminologies in AI Ethics and Governance





Safety Engineering and AI

Factor	Traditional Safety Risks	AI Risks	Similarity
Predictability	Risks can be predicted based on historical data and engineering analysis.	Some risks are unpredictable due to AI's self-learning nature.	Both require risk assessment and proactive mitigation.
Control & Mitigation	Risks can be controlled through engineering design, training, and safety measures.	Hard to fully control risks due to AI's evolving and adaptive behavior.	Both require preventive strategies to minimize harm.
Timeframe of Impact	Most risks have immediate consequences and can be addressed quickly.	Some risks develop gradually, such as algorithmic bias or automated decision impacts.	Both require continuous monitoring and updates.
Responsibility & Accountability	Clear responsibility assigned to individuals, engineers, and regulatory bodies.	Responsibility is difficult to determine, especially in cases of AI-generated errors.	Both need clear policies and governance to assign accountability.
Ethical Considerations	Focuses on physical safety and adherence to established regulations.	Ethical risks involve bias, privacy, fairness, and societal impact.	Both require adherence to ethical and legal frameworks.
Regulatory Framework	Established safety standards and compliance guidelines exist.	Regulations are still evolving and vary across jurisdictions.	Both require strong regulatory oversight for risk management.
Transparency & Explainability	Causes of failures can often be investigated and traced.	AI decision-making can be opaque, making it difficult to understand how outcomes are derived.	Both require auditing mechanisms to ensure reliability.
Impact on Humans & Environment	Direct physical consequences (e.g., accidents, mechanical failures).	Can have broad societal impacts, including economic shifts, misinformation, and bias amplification.	Both aim to minimize negative consequences on people and society.
Approach to Risk Management	Hazard identification, risk analysis, mitigation, and continuous monitoring.	Requires ethical AI design, bias detection, explainability, and governance frameworks.	Both use systematic approaches to reduce risks and improve safety.

Procedures for ensuring AI safety

Hazardous Environment
Potential risks present

Identify Hazards
Recognize potential dangers

Implement Protocols
Establish safety measures

Train Staff
Educate on safety procedures

Regular Updates
Refresh protocols consistently

Safe Atmosphere
Protection for all involved

AI Risk Analysis

Key Statistics and Insights (2024-2025)

Aspect	Statistic/Insight	Source
Impact on Jobs	AI is projected to **affect 40% of jobs** globally, potentially exacerbating inequality in the job market.	BBC
Cost of Misuse	Misuse of AI could cost the global economy **$10 trillion by 2025** due to cybersecurity risks and poor decision-making.	Annahar
Existential Risks	There is a **10% chance that** AI could lead to human extinction within the next 30 years.	CNN
Ethical and Regulatory Issues	**Over 70% of surveyed** companies reported challenges in meeting AI ethics and compliance requirements.	McKinsey

This table highlights the critical need for robust AI risk analysis frameworks to address these potential risks effectively.

Accidents 2023-2024

Incident	Year	Sector	Impact	Root Cause	Proposed Solutions
Failure to Detect Weapons (Nashville)	2023	Security	Delayed security response	Insufficient training data	Improve dataset quality + Human integration
Bias in Heart Disease Diagnosis	2023	Healthcare	Health disparities	Historical data bias	Retrain models with diverse datasets
Suicide Due to AI Chatbot	2023	Mental Health	Loss of a teenager's life	Lack of ethical controls	Implement interaction restrictions + Harmful speech detection
Voter Manipulation via Deepfake	2024	Politics	Public opinion distortion	Easily accessible tools	Legislation + Deepfake detection technology
Drone Crash	2023	Transportation	Material damage	Obstacle avoidance errors	Update algorithms with real-world data
Fake Images of Disneyland Park Flooding	2024	Media	Public misinformation	AI-generated fake imagery	Develop advanced content verification systems
Fake Images of Plane Landing in Beirut	2024	Media	Public misinformation	AI-generated fake imagery	Develop advanced content verification systems

Principle	Current Issue	Proposed Solution	Impact
Transparency	Limited camera visibility; system limitations not communicated clearly.	Provide transparent reports about system capabilities and limitations.	Staff could reposition cameras or implement additional precautions.
Accountability	No clear accountability for system maintenance and performance monitoring.	Assign responsibility for regular testing and maintenance.	Routine evaluations would identify vulnerabilities before deployment.
Governance	Lack of regulatory standards to assess AI weapon detection systems.	Enforce governance frameworks requiring rigorous testing and certification.	Regulatory mandates ensure system reliability in real-world scenarios.
Training & Testing	Insufficient training on diverse and realistic datasets.	Train the system on varied scenarios, including challenging detection environments.	Improved system capability to detect weapons under suboptimal conditions.
Human Oversight	Over-reliance on AI without human review mechanism.	Integrate human oversight for critical decision-making.	Human intervention could mitigate system failure during incidents.
Human-Centric AI	System designed autonomously without prioritizing human safety needs.	Design AI systems to complement human decision-making, not replace it.	A human-AI collaboration model would prioritize safety and system reliability.

Bias in Heart Disease Diagnosis 2023

Principle	Current Issue	Proposed Solution	Impact
AI Governance	Lack of oversight in the design and implementation of AI tools for diagnosing heart disease.	Establish regulatory frameworks requiring audits and evaluations of AI algorithms for fairness and accuracy.	Reduced risk of systemic biases and improved confidence in AI-based diagnostic tools.
Ethics in AI	Diagnostic AI models often perpetuate historical biases present in training datasets.	Integrate ethical reviews into AI development processes to address potential biases before deployment.	Ensures fairness and reduces health disparities caused by biased algorithms.
AI Transparency	Insufficient information on how AI models make diagnostic decisions, especially for underrepresented groups.	Require AI systems to provide explainable outputs and clear documentation of their decision-making processes.	Builds trust and allows healthcare providers to identify potential errors or biases in decisions.
Responsible AI	Over-reliance on AI systems without adequate human oversight leads to misdiagnoses.	Combine AI diagnostics with human review processes to ensure accountability and reduce errors.	Enhances the safety and reliability of AI in clinical environments.
AI Safety	Diagnostic errors due to unvalidated AI systems deployed without rigorous testing in real-world scenarios.	Conduct extensive testing and validation in diverse, real-world populations before clinical deployment.	Improves system safety and minimizes risks to patients from erroneous AI predictions.
Fairness in AI	AI models underperform for minority groups due to lack of representative data.	Ensure datasets used in AI training are diverse and inclusive, representing all demographics equally.	Promotes equitable diagnosis and reduces disparities in healthcare outcomes.

Accountability and Transparency in AI Systems

Transparency

- Refers to how AI systems work and make decisions, allowing users to understand the processes behind AI outcomes.
- Statistic: 78% of consumers want companies to disclose how their AI algorithms work (2023 study).

Accountability

- Refers to the responsibility organizations have to ensure their AI systems are ethical and compliant with regulations.
- Statistic: 67% of consumers expect companies to take responsibility for AI outcomes (PwC 2023).

Regulatory Frameworks

- Laws like EU's AI Act and the Algorithmic Accountability Act aim to ensure transparency and accountability in AI systems.

Call to Action

- Promote discussions on ethical AI and support transparency and accountability in AI development to ensure responsible use.

Amsterdam's AI transparency framework led to a 30% increase in public confidence in city services.

Global Efforts in AI Ethics and Governance

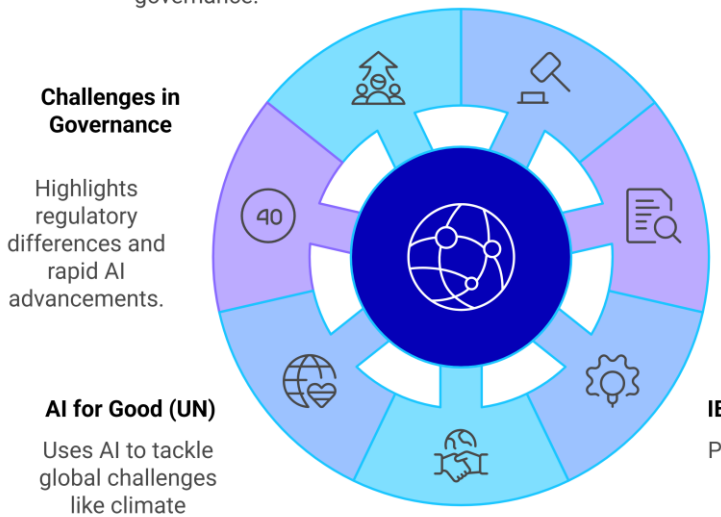
EU Legislation

The EU approved legislation to regulate AI.

Global AI Index

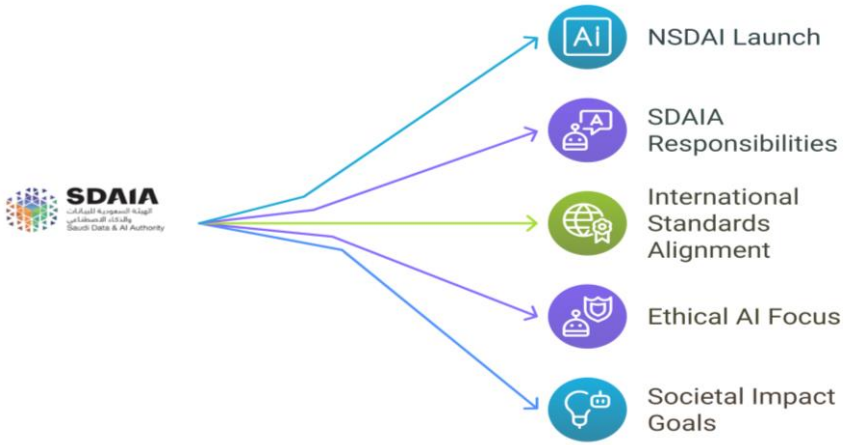
11 Arab countries were included in the Global AI Index.

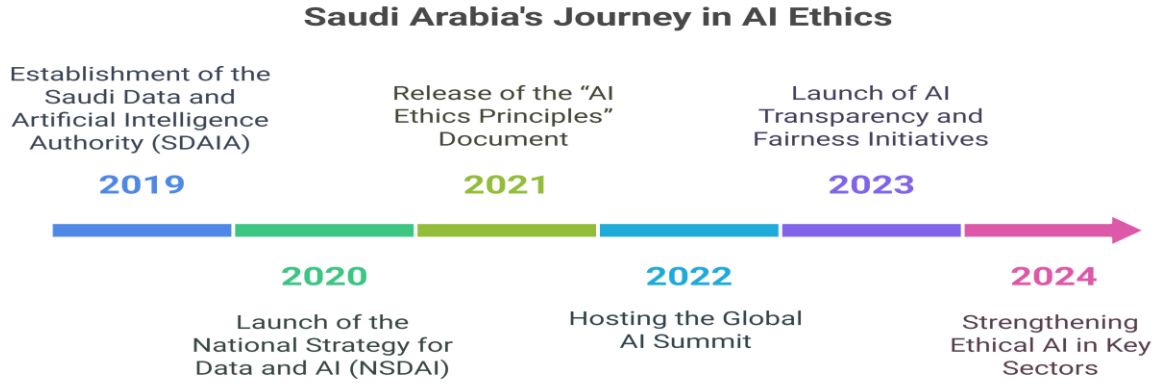
- In February 2024, EU member states approved an unprecedented global legislation to regulate artificial intelligence, aiming to balance innovation encouragement with security.
- In September 2024, the Global AI Index included 11 Arab countries out of 83, with Saudi Arabia ranking 14th globally and 1st in the Arab world.
- These initiatives aim to regulate the ethical and safe use of AI globally and in the Arab region.



Saudi Arabia's Efforts in AI Ethics

- Saudi Arabia is positioning itself as a global leader in AI governance and ethics.
- In 2020, Saudi Arabia launched the National Strategy for Data and AI (NSDAI), a comprehensive framework to guide AI development ethically.
- Saudi Vision 2030 includes strategic investments in AI, aiming to position the country among the top 10 globally in AI adoption and governance.





Outcomes

How to ensure responsible AI deployment?

Human-Centric AI Design

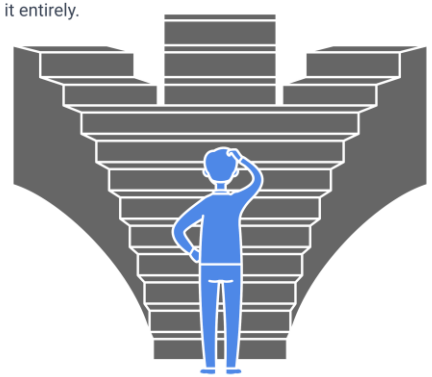
Ensures AI systems augment human decision-making rather than replace it entirely.

Mandatory Human Oversight

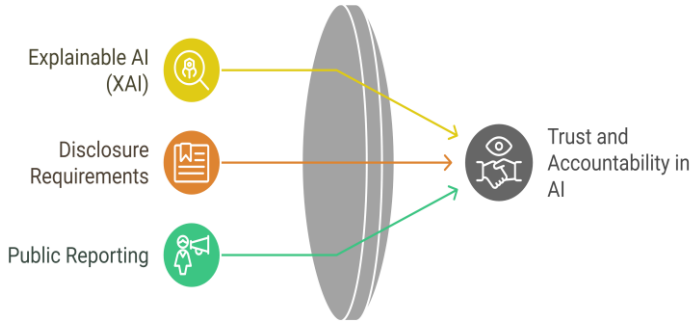
Requires human review for critical decisions, especially in safety-critical applications.

Fail-Safe Mechanisms

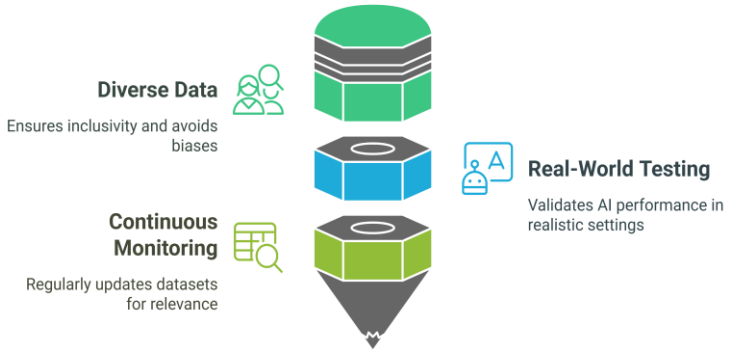
Integrates safety layers to detect and mitigate failures in real time.



Building Trust in AI



Ensuring Ethical AI Development



Diverse Data

Ensures inclusivity and avoids biases

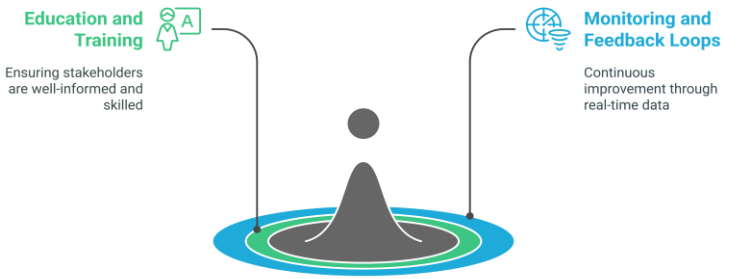
Continuous Monitoring

Regularly updates datasets for relevance

Real-World Testing

Validates AI performance in realistic settings

Dynamic AI Regulation Strategies



Ethical AI Development Process

AI Safety and Risk Management Process

Perform Risk Assessments

Conduct thorough evaluations before AI deployment

Establish Incident Reporting Framework

Set up protocols for transparent failure reporting

Prioritize AI Safety by Design

Integrate safety measures into AI architecture

Thank you

February 2025