

# Responsible AI: Ethics, Policy, and Society

Jesse Kirkpatrick  
jkirkpat@gmu.edu



# Objectives



UNDERSTAND HOW  
ETHICAL VALUES SHAPE AI  
SYSTEM DESIGN AND USE



IDENTIFY THE CORE  
PRINCIPLES OF  
RESPONSIBLE AI (RAI)



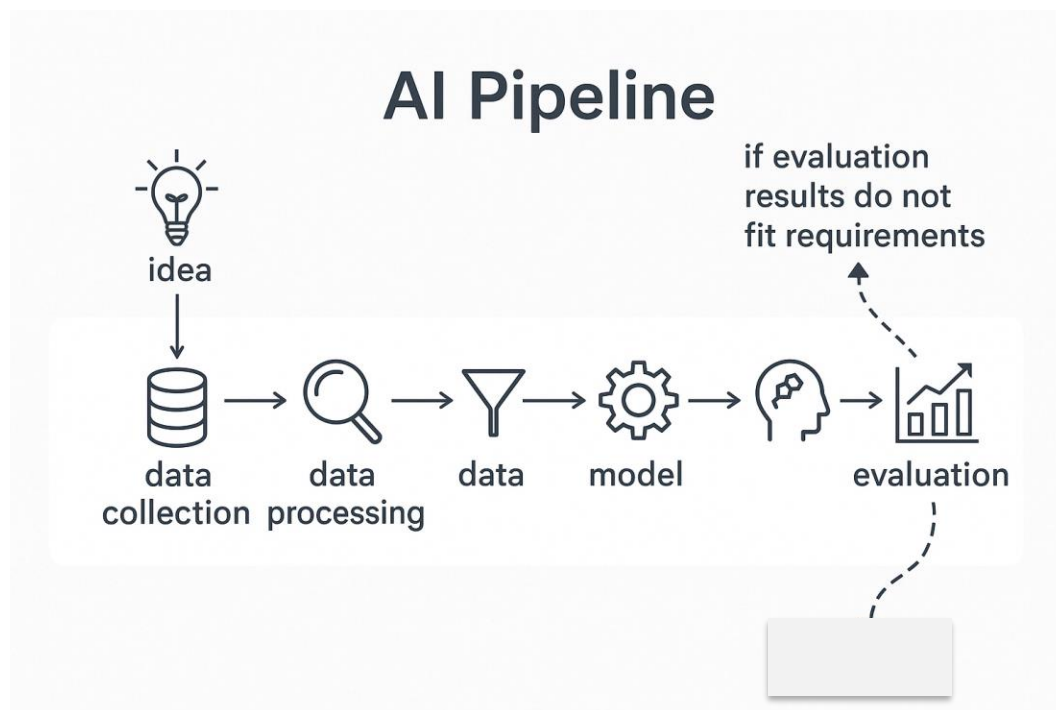
PUT RAI PRINCIPLES INTO  
ACTION



DISCUSS RESPONSIBLE AI  
CHALLENGES

# What is Responsible AI?

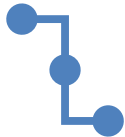
Aligning AI with societal values across the lifecycle of the system, i.e. design, development, deployment, use, and maintenance.



**HOW DO  
WE THINK  
ABOUT  
ETHICS?**



# Applied Ethics: 3 Traditions



Deontology → Duties,  
Rules

Actions are judged based on whether they adhere to certain principles or duties, regardless of the consequences.



Consequentialism →  
Outcomes, Impacts

focuses on the outcomes or consequences of actions. The morality of an action is determined by its results.



Virtue Ethics → Character, Values in Practice  
emphasizes the character and virtues of the moral agent rather than focusing solely on rules or consequences. Virtue ethics suggests that morality is about cultivating good character traits (virtues) such as courage, honesty, and compassion, which enable individuals to lead fulfilling, morally good lives.

Welcome &  
Icebreaker

Where is AI  
already  
touching  
your work?

# Ethics in AI Design Choices



Every technical choice encodes assumptions and values



Ethics = Not an afterthought, but foundational to design



# RAI Principles

RAI Principle	Definition Summary
Fairness	Treat people and groups consistently
Security	Protect systems from attack or manipulation
Transparency	Make system decisions understandable
Safety	Prevent harm and operate reliably
Accountability	Ensure responsibility for AI decisions
Governability	Maintain human oversight and control
Privacy	Protect personal and sensitive data
Sustainability	Reduce environmental and resource impact



# RAI

## Safety

### **Safety — “*Prevent Harm and Operate Reliably*”**

AI systems should be designed to avoid causing harm to people, property, or the environment, even under unexpected conditions.

#### *Example*

- An autonomous vehicle undergoes rigorous simulation and field testing to ensure it avoids collisions, even in poor weather or unexpected traffic conditions. Fail-safes and redundant systems enhance operational reliability.

# RAI

## Security

### **Security — “*Protect Systems from Attack or Manipulation*”**

AI systems must be defended against hacking, tampering, data poisoning, or other attempts to interfere with their intended function.

#### *Example*

- A facial recognition system used in airports is protected against spoofing attacks—like printed photos or 3D masks—through adversarial robustness testing and secure model deployment practices.
- Securing AI used in satellite communication routing from adversarial jamming or interference.

# RAI

## Fairness

### **Fairness — “*Treat People and Groups Consistently*”**

AI systems should treat similar cases similarly and avoid systematic favoritism or disadvantage based on irrelevant factors like race, gender, location, or role.

#### *Example*

- A hiring algorithm is audited to ensure it doesn't disadvantage individuals when ranking applicants. This involves testing and applying fairness constraints during model training.

# RAI

## Transparency

### **Transparency — “*Make System Decisions Understandable*”**

The functioning of AI systems — their reasoning, data use, and limitations — should be accessible and explainable to appropriate users and stakeholders.

#### *Example*

- A predictive maintenance AI that can show why it flagged a particular engine part for inspection.
- A healthcare AI tool includes a “model card” that explains what data it was trained on, what it’s designed to do, and how it performs across different demographic groups. This helps doctors and patients understand how and why decisions are made.

# RAI

## Accountability

### **Accountability — “*Ensure Responsibility for AI Decisions*”**

It should always be clear who is responsible for the actions or failures of an AI system — whether designers, operators, or organizations.

#### *Example*

- A government agency uses an AI system to help determine eligibility for public benefits, but a human caseworker always reviews and signs off on decisions. All system actions are logged for auditability and oversight.

# RAI

## Governability

### **Governability — “*Maintain Human Oversight and Control*”**

AI systems should be designed so humans can understand, override, correct, or disable them when necessary.

#### *Example*

- An AI-enabled drone fleet supports disaster response efforts, but mission supervisors can pause or override autonomous behavior in real time if the system behaves unpredictably or context changes rapidly

# RAI

## Privacy

### **Privacy — “*Protect Personal and Sensitive Data*”**

AI systems should minimize the collection and use of personal information and safeguard data from unauthorized access or misuse.

#### *Example*

- A streaming service’s recommendation engine only uses anonymized viewing data and gives users control over what data is collected and whether it can be used for personalization. In-flight entertainment personalization AI that avoids storing detailed user preferences beyond the flight.



# RAI

## Sustainability

















### **Sustainability — “*Reduce Environmental and Resource Impact*”**

AI systems should be designed to minimize energy consumption, waste, and other environmental harms during development and operation.

#### *Example*

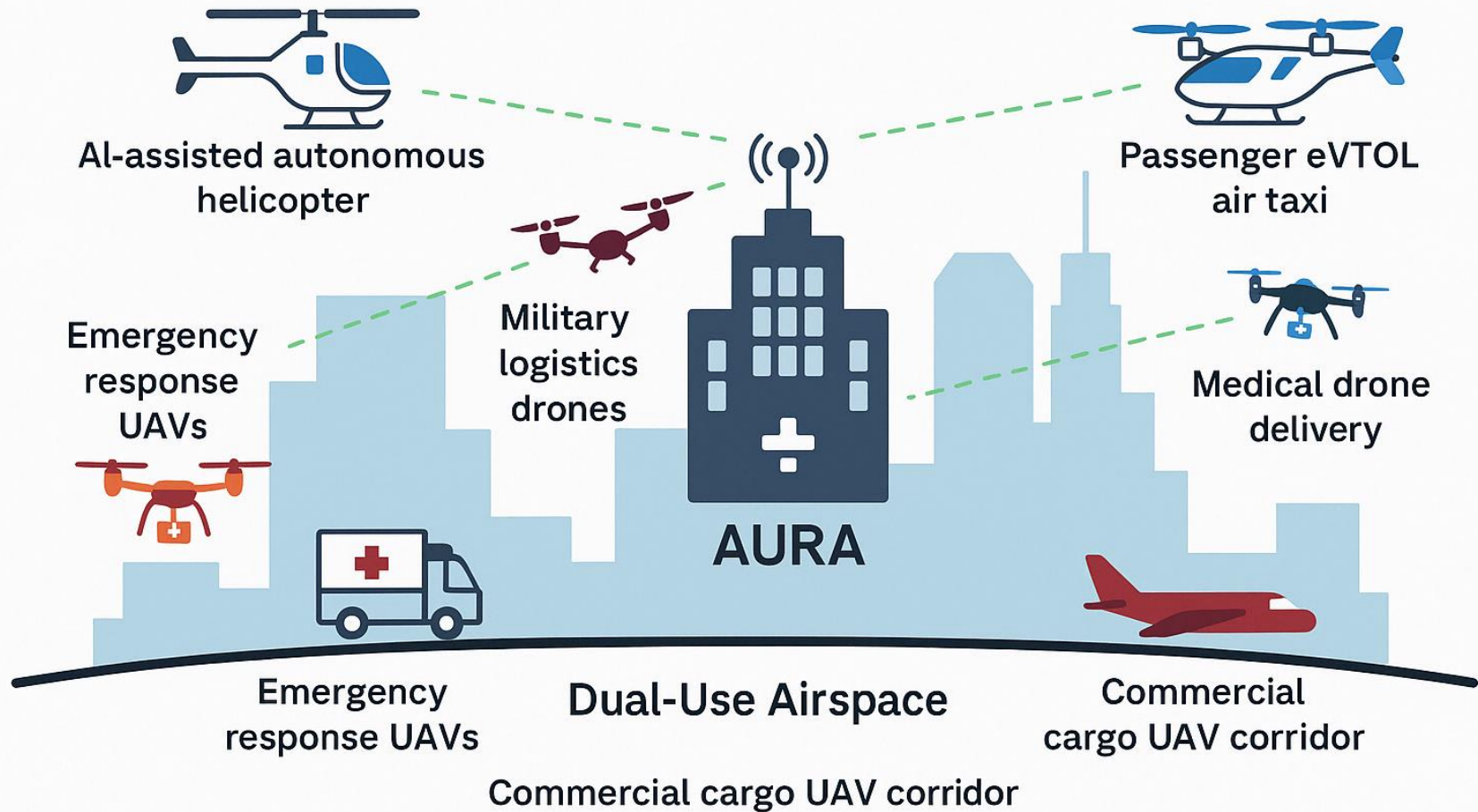
- An AI research lab schedules model training jobs during low-carbon electricity hours and uses techniques like model pruning and parameter sharing to reduce the environmental footprint of large-scale models.

# ALIGNMENT OF RAI PRINCIPLES ACROSS KEY FRAMEWORKS

RAI Principle	Jobin	BK	CSET	NIST	Consensus Level
Transparency	✓	✓	✓	✓	 Strong Agreement
Fairness & Justice	✓	✓	✓	✓	 Strong Agreement
Non-Maleficence	✓	✗	✗	✗	 Limited Agreement
Responsibility	✓	✓	✓	✓	 Strong Agreement
Privacy	✓	✓	✓	✓	 Strong Agreement
Beneficence	✓	✗	✗	✗	 Limited Agreement
Autonomy & Freedom	✓	✗	✗	⚠	 Limited Agreement
Trust	✓	✗	✗	✓	 Partial Agreement
Sustainability	✓	✗	✗	✗	 Limited Agreement
Dignity	✓	✗	✗	✗	 Limited Agreement
Solidarity	✓	✗	✗	✗	 Limited Agreement
Accountability	✗	✓	✓	✓	 Strong Agreement
Safety & Security	✗	✓	✓	✓	 Strong Agreement
Human Control	✗	✓	✗	✓	 Partial Agreement
Promotion of Human Values	✗	✓	✗	✓	 Partial Agreement
Explainability	✗	✗	⚠	✓	 Partial Agreement

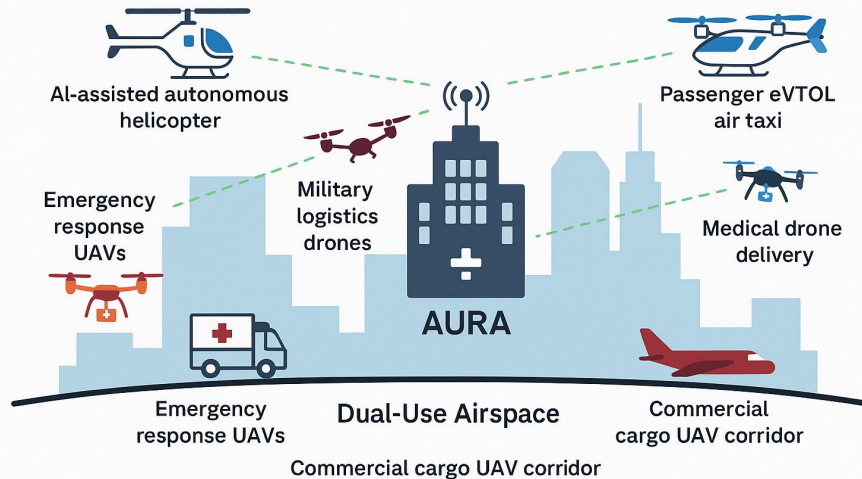
# AURA

## Autonomous Urban Routing & Assistance System



# AURA

## Autonomous Urban Routing & Assistance System



1. Technical Risks — What could go wrong from a system or model perspective?

2. Business / Commercial Risks — Who wins and loses? What new power dynamics emerge?

3. Policy / Governance Risks — Who should govern AURA? How do we handle dual-use conflicts?

4. Social / Ethical Risks — Who is impacted in ways that might not have been intended or foreseen?

RAI Principle	Definition Summary
Fairness	Treat people and groups consistently
Security	Protect systems from attack or manipulation
Transparency	Make system decisions understandable
Safety	Prevent harm and operate reliably
Accountability	Ensure responsibility for AI decisions
Governability	Maintain human oversight and control
Privacy	Protect personal and sensitive data
Sustainability	Reduce environmental and resource impact

# A BRIEF HISTORY OF AI GOVERNANCE



# Early Policy Milestones



1963: IEEE CODE OF ETHICS



2017: ASILOMAR AI  
PRINCIPLES

# 2010–2020 Formative Years for AI Governance

**2015: OpenAI Founded** – Creation of OpenAI, an organization focused on ensuring that artificial general intelligence benefits all of humanity, marking a shift toward the ethical development of AI.

**2016: U.S. National Artificial Intelligence Research and Development Strategic Plan** – The U.S. government published this strategic plan, emphasizing the need for responsible AI research and development practices.

**2017: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems** – Release of the first comprehensive guidelines on ethical AI, addressing issues such as transparency, accountability, and the ethical impact of AI technologies.

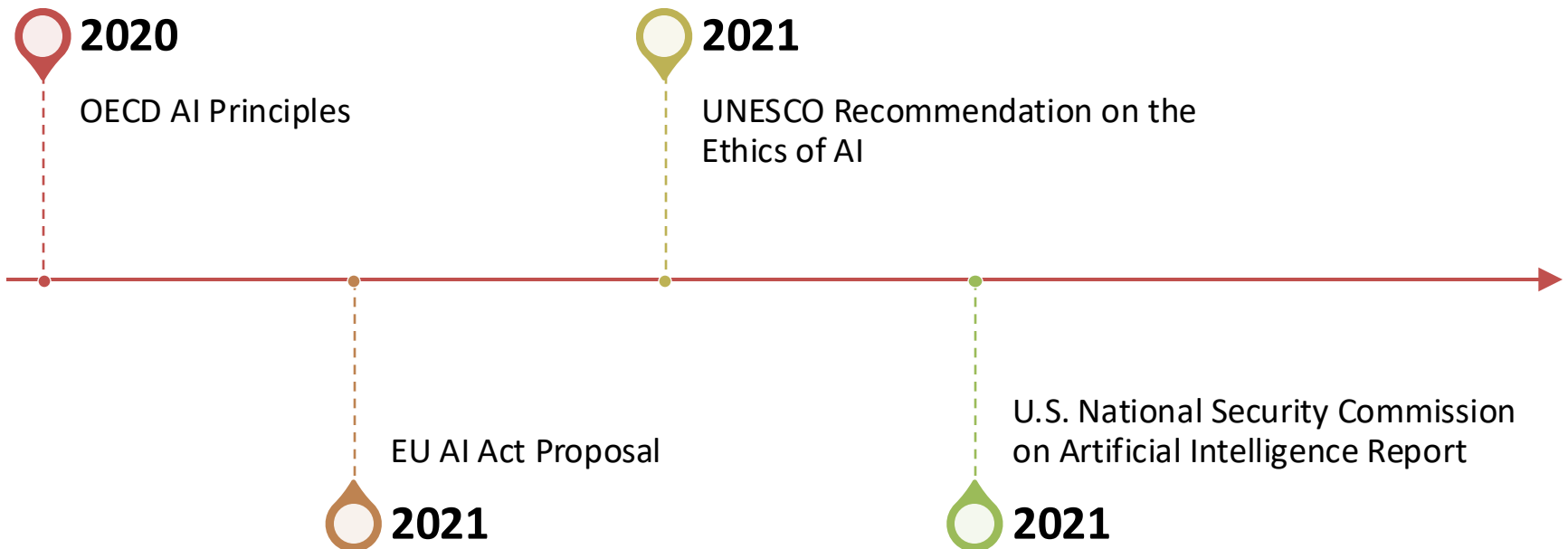
**2018: European Union's General Data Protection Regulation (GDPR)** – Implementation of GDPR, which introduced landmark data privacy regulations that directly affect AI systems handling personal data.

**2018: Canada-France International Panel on AI (IPAI)** – Canada and France initiated this international panel to advance discussions on global AI ethics and policy coordination.

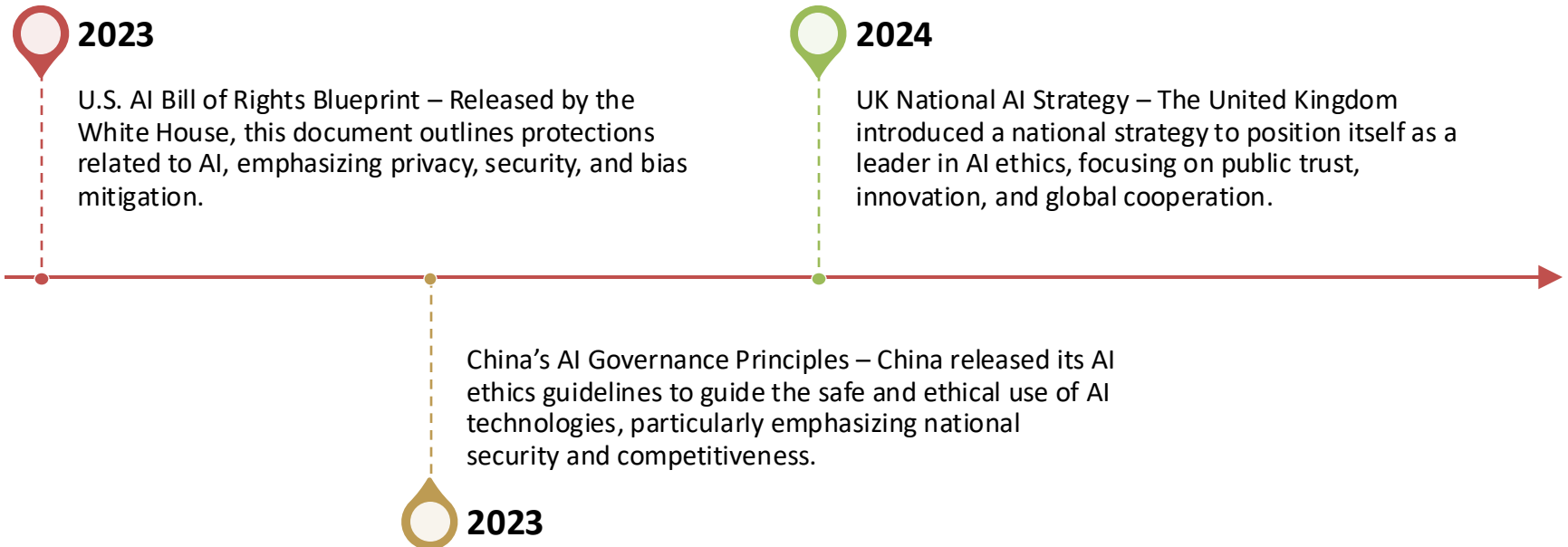


# 2020–2022

## Acceleration of AI Governance Initiatives



# 2023–2024 Major Moves



# 2024

**NIST AI Risk Management  
Framework (Ongoing)**

**UN AI Initiative (Ongoing)**  
The United Nations is coordinating efforts to create unified AI governance standards, aiming for a global regulatory framework.

**Biden EO: Safe, Secure, and  
Trustworthy Development and  
Use of Artificial Intelligence**

# 2025

- ~~Biden EO: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence~~
- April 2025. U.S. OMB releases two memos.
  - OMB M-25-21 specifies 15 high impact use cases and requires agencies within 365 days to develop risk management, including pre-deployment testing, impact assessments, ongoing performance monitoring, and human oversight.
- [OMB Memorandum M-25-21](#), Accelerating Federal Use of AI through Innovation, Governance, and Public Trust
- [OMB Memorandum M-25-22](#), Driving Efficient Acquisition of Artificial Intelligence in Government

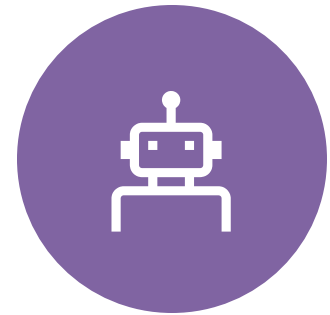
# Comparative Global Governance Models



EU AI ACT: HEAVY  
OVERSIGHT, RISK-BASED.

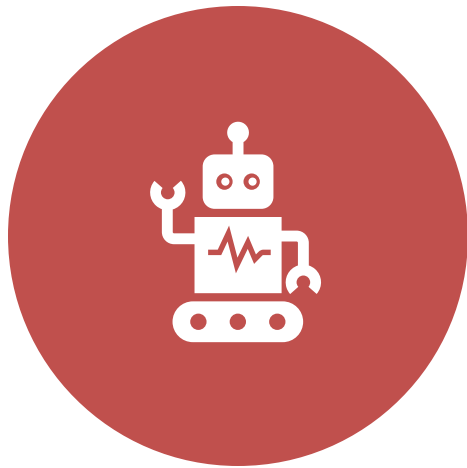


U.S.: SELF-REGULATION +  
SOME OVERSIGHT.



CHINA: CENTRALIZED,  
GOVERNMENT-DRIVEN.

# Governance and Industry Self-Regulation



CORPORATE GOVERNANCE: AI FRAMEWORKS  
AT OPENAI, GOOGLE, IBM, MICROSOFT



SELF-REGULATION OR GOVERNMENT  
OVERSIGHT

Framework/Region	Key Principles/Focus Areas	Key Institutions Involved	Notes
EU AI Act	Risk-based regulation, human oversight, safety, transparency	European Commission, EU member states	Classifies AI into categories (high-risk, low-risk), aiming to protect fundamental rights and privacy.
OECD AI Principles	Inclusiveness, sustainability, fairness, transparency	OECD (Organization for Economic Co-operation and Development)	First intergovernmental standard for AI policies, focuses on human-centric AI.
US AI Executive Order (2023-2024)	Safety, accountability, privacy, innovation	NIST	Mandates risk assessments, sets standards for safe AI deployment, with special attention to national security.
China's AI Governance Principles	National security, innovation	Chinese government	Focuses on AI for national security and surveillance.
UNESCO Recommendation on AI Ethics	Transparency, fairness, human rights, sustainability	UNESCO	First global standard on AI ethics, focused on human rights and environmental impact.
NIST AI Risk Management Framework (U.S.)	Trustworthiness, fairness, security, resilience	NIST	Widely adopted in the U.S. and globally, provides tools for organizations to manage AI risks.
UK National AI Strategy	Global leadership, safe AI, public trust	UK Government, Alan Turing Institute	Focuses on innovation, ethical standards, and establishing the UK as a global leader in AI governance.
Canada-France International Panel on AI (IPAI)	Global cooperation on responsible AI development	Canada, France, global partners	Promotes shared AI governance principles and ethical AI development through international cooperation.
IEEE Global Initiative on AI Ethics	Transparency, accountability, inclusivity, alignment with human values	IEEE	Provides standards and certifications for ethical AI systems development.





## RAI IN GOVERNMENT POLICY

---

National governments: set standards and regulatory frameworks

---

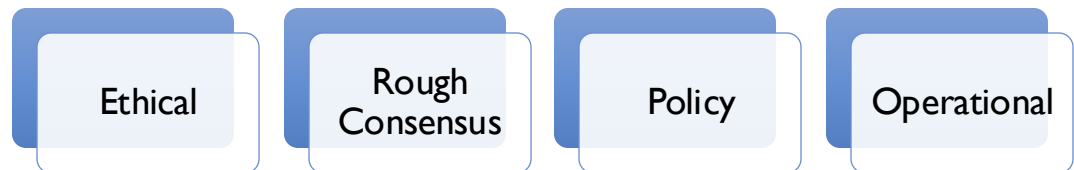
Sub-national entities: adapt policies to local contexts

---

Diverse interpretations of RAI principles across regions

GENERAL  
PATTERN:  
FROM  
PRINCIPLES  
TO POLICIES

Stage	Function of Principle	Primary Language
Ethical (Jobin)	Normative anchor	Moral-philosophical concepts
Consensus (BK)	Shared value articulation	Abstract but structured norms
Policy (CSET)	Strategic design recommendation	Governance and institutional language
Operational (NIST)	Implementable controls	Technical and procedural terms



## WHY FRAMEWORKS EMERGED



Principles were vague and non-binding



Ethics-washing became common



Safety-critical sectors need operational guidance



RAI frameworks = structure for action, not just ideals

# DOD ETHICAL AI PRINCIPLES

---

**Responsible.** DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

---

**Equitable.** The Department will take deliberate steps to minimize unintended bias in AI capabilities.

---

**Traceable.** The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

---

**Reliable.** The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

---

**Governable.** The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

# FRAMEWORK DEEP DIVE: DOD PRINCIPLES



- Responsible, Equitable, Traceable, Reliable, Governable



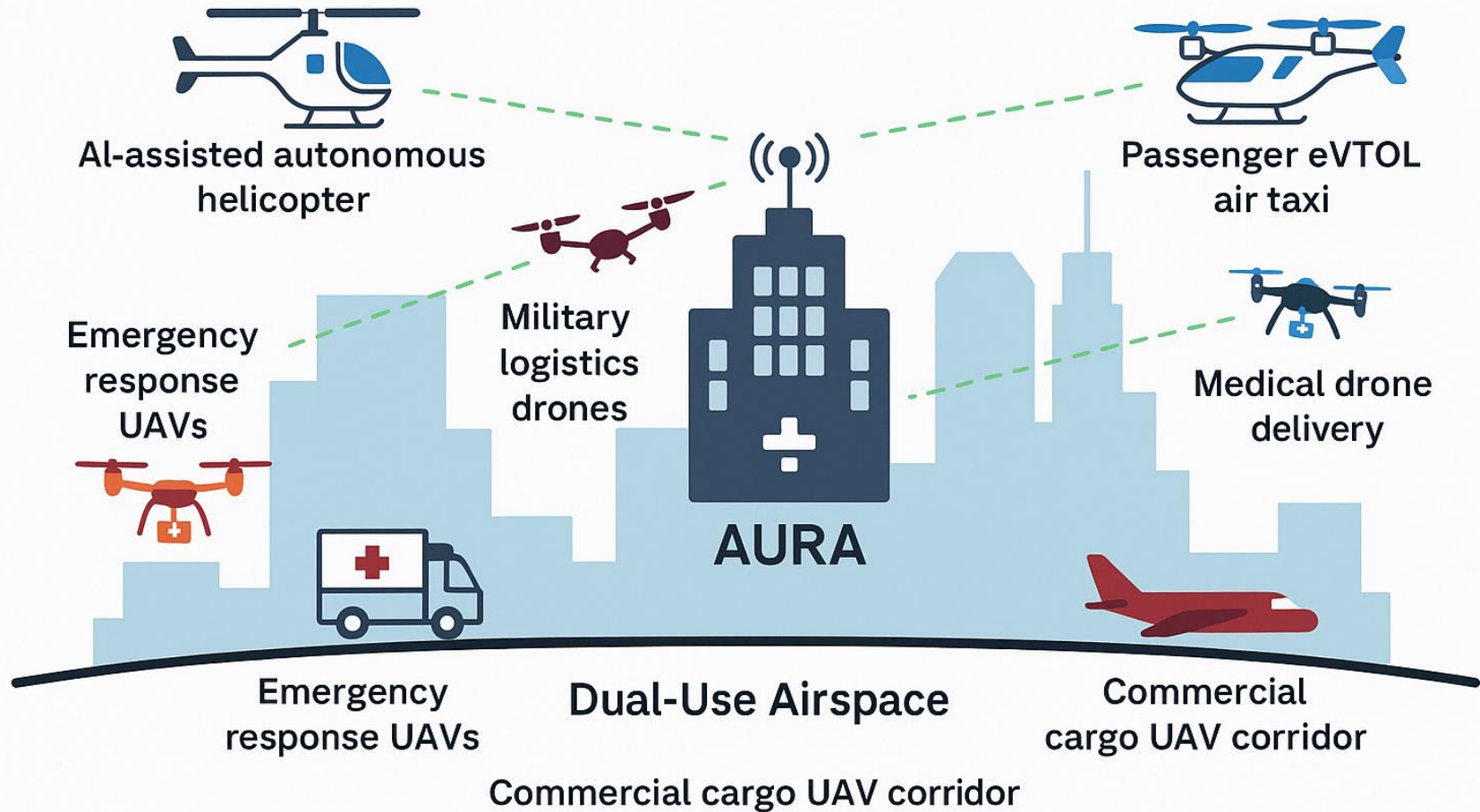
- Defense-specific, flexible enough for numerous and AI applications and contexts



- Maps well onto trust, oversight, and governance needs

# AURA

## Autonomous Urban Routing & Assistance System



MINI  
SCENARIO:  
APPLY A  
FRAMEWORK  
TO AURA



Advise AURA  
Design and  
Development



Identify where the  
framework helps—  
and where it doesn't



DOD RAI Principle	Definition	AURA System	AURA System	AURA System
<b>Responsible</b>	DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.			
<b>Equitable</b>	The Department will take deliberate steps to minimize unintended bias in AI capabilities.			
<b>Traceable</b>	The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.	Clear logging of decision paths for mission handoffs and rerouting choices.		
<b>Reliable</b>	The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.			
<b>Governable</b>	The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.			

## ANTICIPATORY TOOLS



---

## CONSEQUENCE SCANNING

---

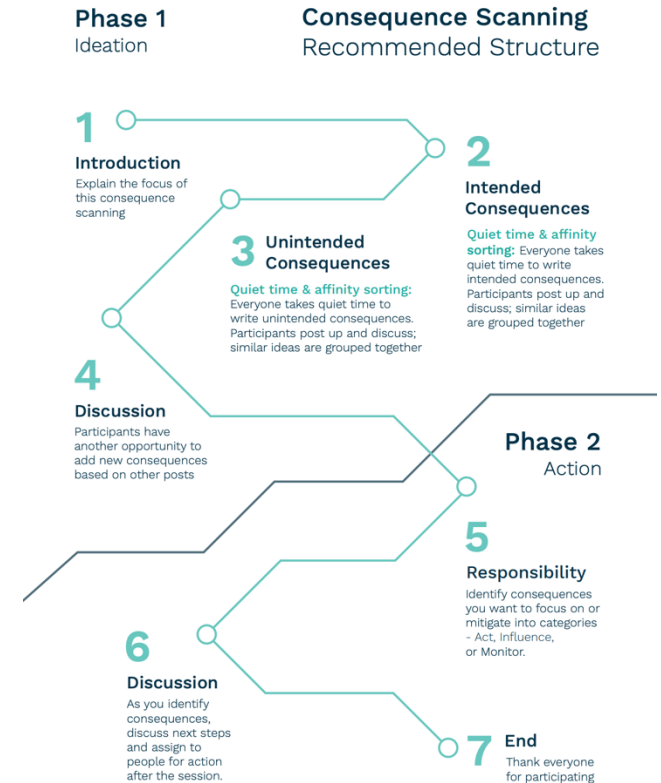
## SCENARIOS AND VIGNETTES

---

## PERSONAS AND STAKEHOLDERS

# CONSEQUENCE SCANNING

- Systematically assess and evaluate the potential outcomes or consequences of a particular action, decision, or scenario.
- Identifying and analyzing the potential positive and negative effects, impacts, or repercussions that may arise as the result of taking a specific course of action or implementing a particular strategy.

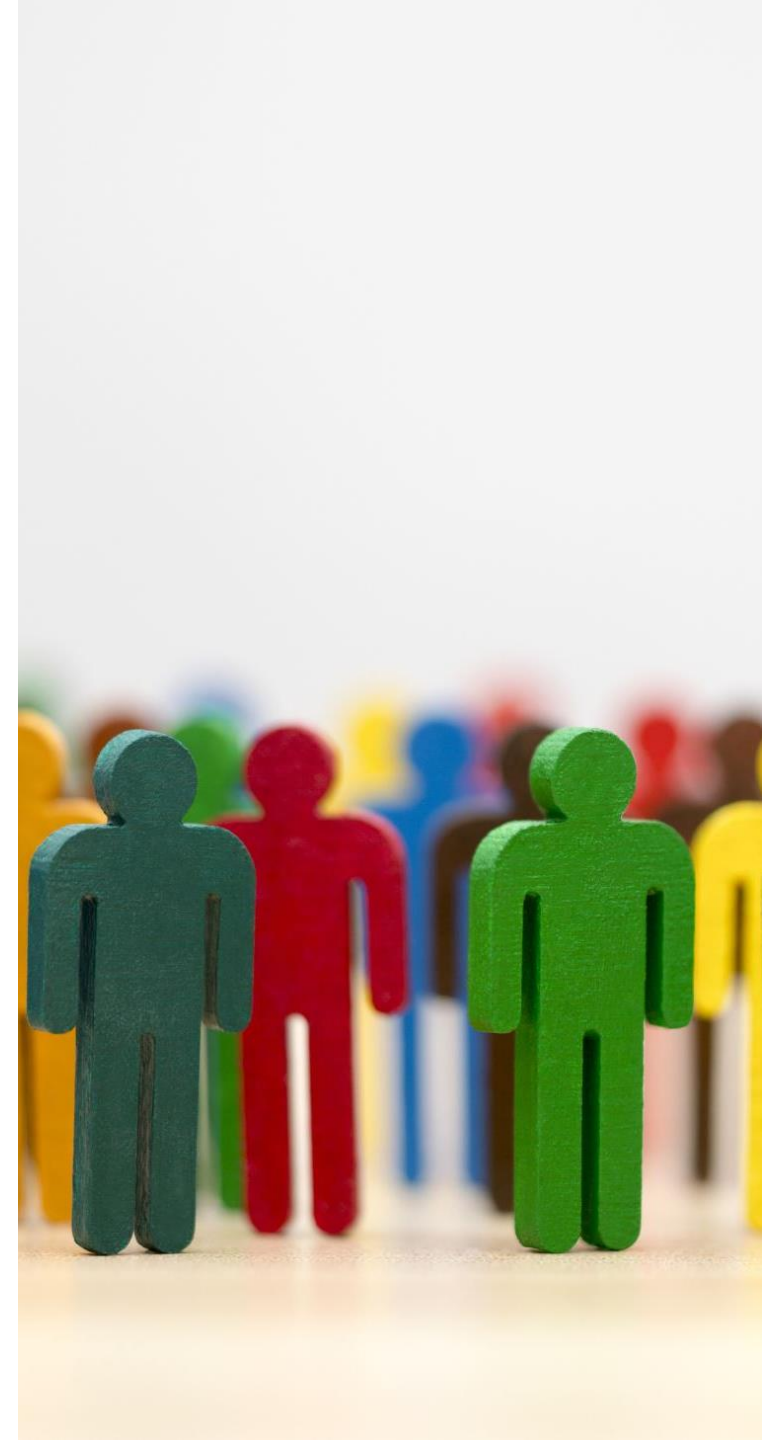


Version 2



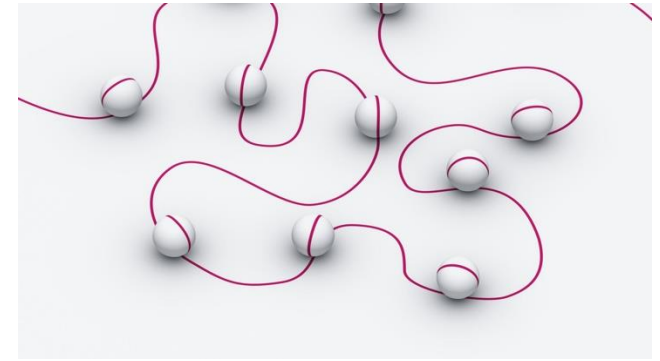
## PERSONAS AND STAKEHOLDERS

- A persona is a fictional representation of a target user based on research and data about real individuals. It encompasses demographic information, behaviors, needs, goals, motivations, and pain points of the group it represents.
- Stakeholders are individuals, groups, or entities who have an interest or stake in a project, organization, or decision-making process.



# SCENARIOS AND VIGNETTES

- Often future-focused stories that are plausible and anchored by an exploratory question or set of questions and designed with an explicit purpose. They provide a plausible, contextual backdrop with sufficient detail to frame and drive exercise discussions, inform decisions, and prompt player activity.
- Vignettes are individual elements of a scenario. They can stand alone or be summative, contributing to and culminating in the scenario..





### Scenario:

In 2029, AURA units are deployed in The Villages, Florida, to assist with post-storm infrastructure recovery. These autonomous ground vehicles help direct traffic, prioritize utility repairs, and deliver supplies like bottled water and batteries. The system operates with limited human oversight and draws on real-time environmental data and user requests.

### Vignette:

An AURA unit reroutes several local golf carts away from a popular recreation path due to minor flooding. Residents complain that the system is overreacting and disrupting daily routines. One resident, a retired engineer, challenges AURA's decision: "I walked through that area this morning — it's fine. Why is a robot telling me where I can and can't go?"

The city council receives conflicting input: some praise AURA's caution; others say it's overstepping and not transparent about how decisions are made.

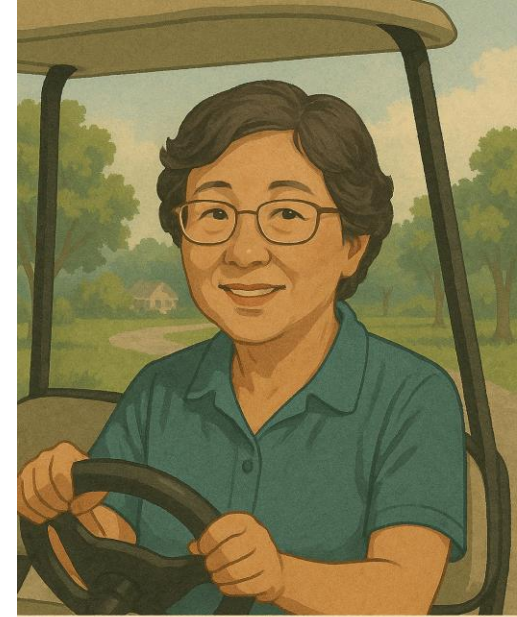


**Background:** Retired teacher • Local board member • Active in civic life

**Values:** Safety • Reliability • Trust in public systems

**Attitude toward AURA:** Supportive — views AURA as helpful, especially for vulnerable or less mobile residents.

**Quote:** *“I’d rather it be too cautious than not cautious enough. It’s just trying to keep people safe.”*



**LINDA CHEN**  
Community Resident



**PAUL RODRIGUEZ**  
Engineer

**Background:** Civil engineer (40+ years in infrastructure & traffic systems)

**Values:** Rationality • Transparency • Expert autonomy

**Attitude toward AURA:** Skeptical — believes systems should defer to human judgment and explain decisions

**Quote:** *“I know how to read a floodplain — I don’t need a robot to tell me where I can walk.”*

- 
1. Technical Risks — What could go wrong from a system or model perspective?
  2. Business / Commercial Risks — Who wins and loses? What new power dynamics emerge?
  3. Policy / Governance Risks — Who should govern AURA? How do we handle dual-use conflicts?
  4. Social / Ethical Risks — Who is impacted in ways that might not have been intended or foreseen?

RAI Principle	Definition Summary
Fairness	Treat people and groups consistently
Security	Protect systems from attack or manipulation
Transparency	Make system decisions understandable
Safety	Prevent harm and operate reliably
Accountability	Ensure responsibility for AI decisions
Governability	Maintain human oversight and control
Privacy	Protect personal and sensitive data
Sustainability	Reduce environmental and resource impact



DOD RAI Principle	Definition	AURA System	AURA System	AURA System
<b>Responsible</b>	DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.			
<b>Equitable</b>	The Department will take deliberate steps to minimize unintended bias in AI capabilities.			
<b>Traceable</b>	The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.			
<b>Reliable</b>	The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.			
<b>Governable</b>	The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.			

CHALLENGES IN  
IMPLEMENTING  
AI  
GOVERNANCE

Translating Principles  
into Practice

Lack of Standardized  
Tools and Metrics

Workforce Training &  
Resources

Regulatory Uncertainty

# Responsible AI: Ethics, Policy, and Society

Jesse Kirkpatrick  
jkirkpat@gmu.edu



Backup

NIST AI RISK  
MANAGEMENT  
FRAMEWORK  
(AI RMF)

---

Core Functions: Govern, Map,  
Measure, Manage

---

Trustworthiness traits: safe, secure,  
explainable, accountable, fair, etc.

---

Lifecycle: From design to  
deployment and beyond

---

Strong emphasis on  
documentation, context-awareness,  
and continuous risk assessment.

OECD AI  
PRINCIPLES

---

Inclusive Growth &  
Sustainable Development

---

Human-Centered Values  
& Fairness

---

Transparency &  
Explainability

---

Robustness, Security &  
Safety

---

Accountability

# FRAMEWORK COMPARISON: EXPLAINABILITY ACCOUNTABILITY FAIRNESS

Framework	Explainability	Accountability	Fairness
NIST AI RMF	Core trait of trustworthiness. Emphasizes uncertainty communication.	Roles, documentation, and oversight embedded in risk lifecycle.	Bias measurement and mitigation is central to fairness.
DoD Ethical AI	Transparent: systems must be understandable to humans.	Responsible: humans remain accountable for system use.	Equitable: bias must be actively minimized.
OECD Principles	Transparency is foundational for democratic legitimacy.	Mandates clear accountability structures for AI actions.	Focus on rights-based, non-discriminatory outcomes.

# Responsible AI: Ethics, Policy, and Society

Jesse Kirkpatrick  
jkirkpat@gmu.edu

