Overview

What is AI ethics?

Origins AI ethics

Transition to Maturity AI ethics: 2019

Organized AI ethics

Principles versus dilemmas/tensions in AI ethics

What is AI ethics

What is AI?

Knowledge production from pattern recognition
Opposed to human knowledge production from linear reasoning, cause/effect
AI: Rain → Clouds, Human: Clouds → Rain

What is Ethics?

Translating experience into language of human values
(Opposed to Physics, Economics)

Evaluation, what is worth doing/having on human level

Origins AI ethics

Medical ethics +
business ethics +
general applied ethics

1985, Beauchamp and Childress, *Principles of Biomedical Ethics*	2019, Transition to AI, European Union: *Ethics Guidelines for Trustworthy AI*
Autonomy	Autonomy
Non-maleficence	Prevention of harm
Justice	Fairness (Justice)
Beneficence	Explicability

*https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

*Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Framework for Trustworthy AI

INTRODUCTION

Lawful AI

(not dealt with in this document)

Ethical AI

Robust AI

CHAPTER I

Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

4 Ethical Principles

Acknowledge and address tensions between them

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

CHAPTER II

Realisation of Trustworthy AI

Implement the key requirements

7 Key Requirements

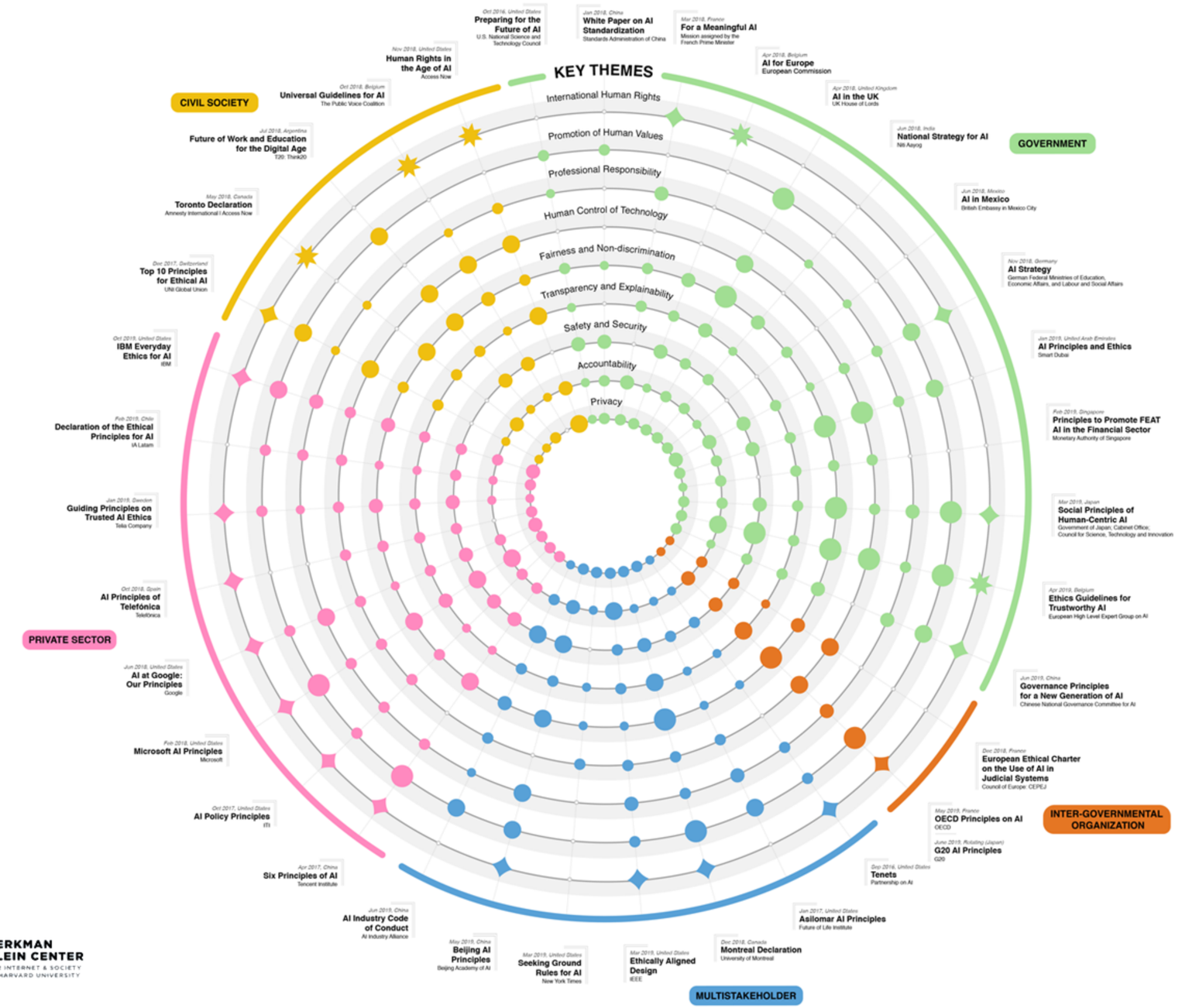
Evaluate and address these continuously throughout the AI system's life cycle via

Technical Methods

Non-Technical Methods

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

Proliferation of AI ethics principles…



BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

	Consolidated	Ethics Guidelines for Trustworthy AI (EC)	Survey of Ethics Principles (Jobin et al 2019)
Individual	Autonomy	Human agency & oversight	Freedom & autonomy
	Dignity		Dignity
	Privacy	Privacy & data governance	Privacy
Social	Fairness	Diversity, non-discrimination, fairness	Justice & Fairness
	Solidarity		Solidarity
	Social Wellbeing	Societal & environmental wellbeing	Sustainability
			Beneficence & non-maleficence
Technical	Performance		Trust
	Safety	Technical robustness & safety	
	Explainability & accountability	Transparency	Transparency
		Accountability	Responsibility

Dilemmas in AI ethics: True, Practical

True: Privacy versus Fairness
Perfectly safeguarding users' personal information prohibits assessments of AI functioning across distinct demographic groups.

Practical: Accuracy versus explainability
The most accurate AI results can emerge from an internal logic that developers or users do not fully understand. Still, perfect accuracy can coexist with full explainability.