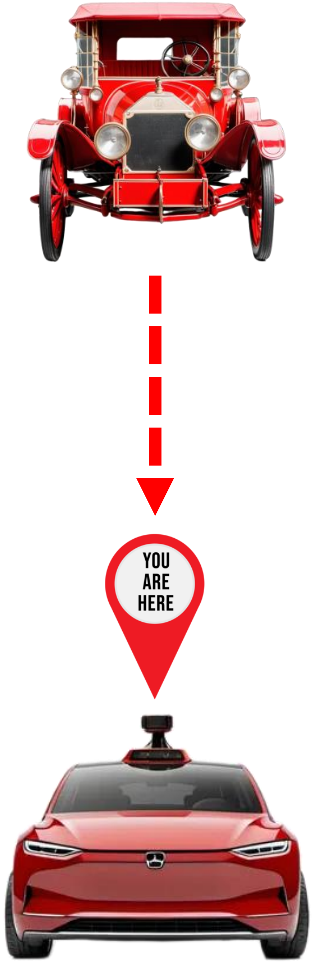
A Brief History of Automated Vehicle Ethics:

Where We've Been & Where We're Headed

Dr. Katie Evans

Future Networked Car Symposium – AI4Good Summit – July 11th, 2025

Overview: The Evolution of AV Ethics



Old School AV Ethics (2014 – 2020)

- The Problem with the Trolley Problem
- Characteristics of Old School AV Ethics

What Happened In Between

- AI Ethics & Governance Through Time
- Current & Emerging AI Use Cases in Automotive

New School AV Ethics (2022 – current)

- Characteristics of New School AV Ethics
- Two New School AV Ethics Problems

Important Notice

All views expressed are personal and do not reflect the formal position of IEEE.
IEEE SA Standards Board Bylaws 5.2.1.6

KEY TAKEAWAY:

STOP TALKING ABOUT THE TROLLEY PROBLEM.

AV ethics isn't just about decision-making in critical scenarios.

Old School AV Ethics

(2014 – 2020)

The Problem With The Trolley Problem

The Trolley Premise:

Fully Autonomous Vehicles (FAVs) could reduce traffic fatalities by up to 90%. Nonetheless, ‘driverless accidents’ are still possible, if rare.

The Trolley Problem:

How should autonomous vehicles be programmed to crash when a collision is unavoidable, or where every action option results in harm? How should we decide which lives to save?

Some Proposed Trolley Solutions:

Academia: Moral theories (e.g. utilitarianism), or empirical research into public acceptability (e.g. MIT Moral Machine Experiment) could be used to guide ethical decisions in unavoidable accidents.

Public Sector: Decisions which involve the use of certain subjective characteristics (e.g. gender), or involve trade-offs across human beings, are not permitted.

Engineering: such accidents will not occur with robust design, or if they do, the time to collision should be spent optimizing trajectories, decelerations, and interactions with road users to mitigate a crash.

Characteristics of Old School AV Ethics

Focus on Critical Decision-Making:

Most research focuses on the Trolley Dilemma*, the Molly Problem**, or generally lethal decision-making in level 5 AVs in mixed fleet traffic scenarios.

+

Focus on Abstract Ethical Theories:

Most research into the design and development of ethical decision-making in AVs makes use of classical theories in moral philosophy (e.g. utilitarianism).

+

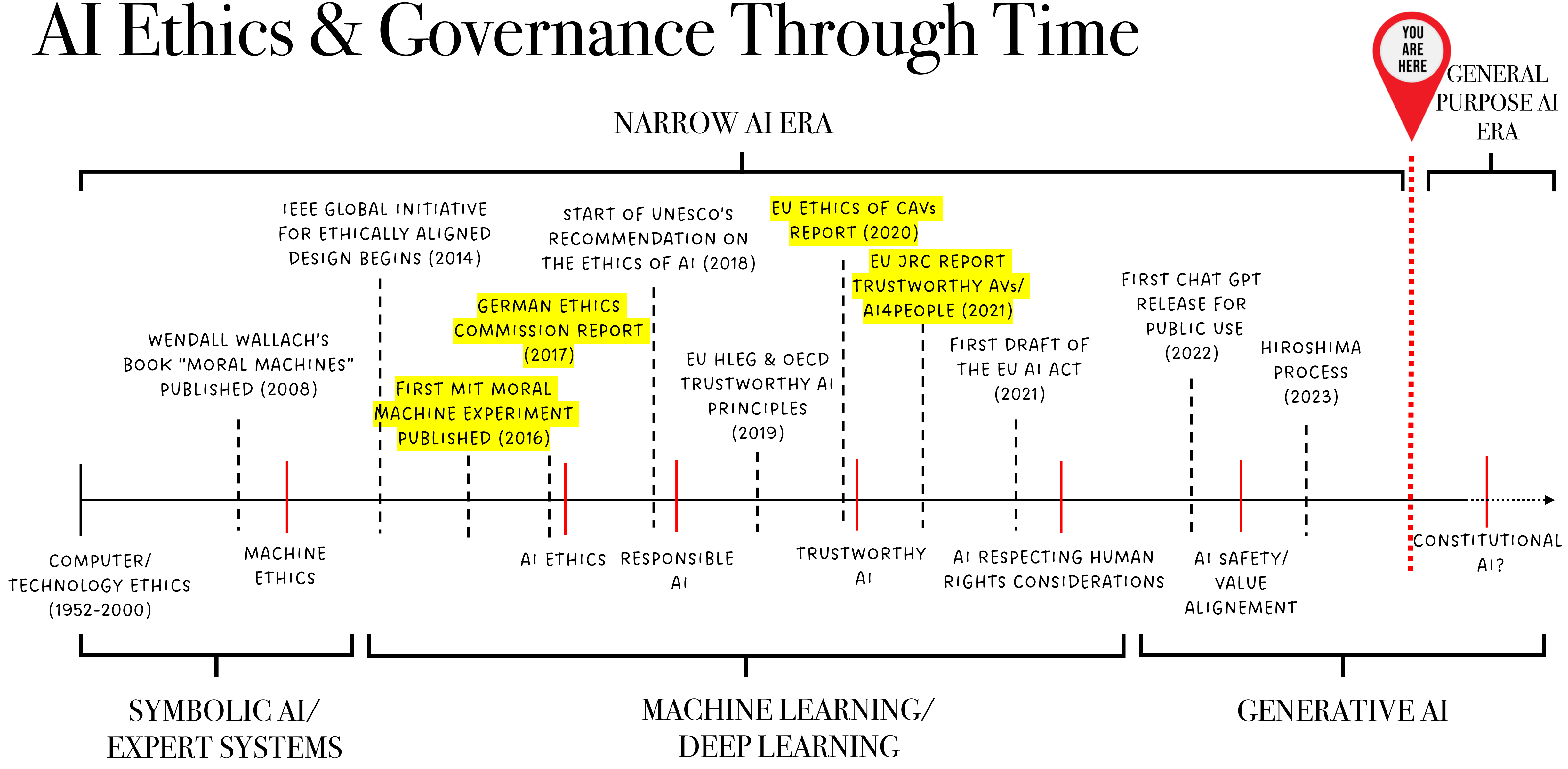
Focus on Public Acceptability:

Most research into the validation of ethical behaviour in AVs relies on public acceptability: what is ethical is whatever behaviour (or ethical theory) people empirically prefer, *c.f.* MIT Moral Machine Experiment***.

*LIN (2014);
**ITU-T FGAI4AD-02 (2021)
***BONNEFON ET AL.,(2016)

What Happened In Between…

AI Ethics & Governance Through Time



YOU ARE HERE

NARROW AI ERA

GENERAL PURPOSE AI ERA

WENDALL WALLACH'S BOOK "MORAL MACHINES" PUBLISHED (2008)

IEEE GLOBAL INITIATIVE FOR ETHICALLY ALIGNED DESIGN BEGINS (2014)

FIRST MIT MORAL MACHINE EXPERIMENT PUBLISHED (2016)

GERMAN ETHICS COMMISSION REPORT (2017)

START OF UNESCO'S RECOMMENDATION ON THE ETHICS OF AI (2018)

EU HLEG & OECD TRUSTWORTHY AI PRINCIPLES (2019)

EU ETHICS OF CAVs REPORT (2020)

EU JRC REPORT TRUSTWORTHY AVs/ AI4PEOPLE (2021)

FIRST DRAFT OF THE EU AI ACT (2021)

FIRST CHAT GPT RELEASE FOR PUBLIC USE (2022)

HIROSHIMA PROCESS (2023)

COMPUTER/ TECHNOLOGY ETHICS (1952-2000)

MACHINE ETHICS

AI ETHICS

RESPONSIBLE AI

TRUSTWORTHY AI

AI RESPECTING HUMAN RIGHTS CONSIDERATIONS

AI SAFETY/ VALUE ALIGNEMENT

CONSTITUTIONAL AI?

SYMBOLIC AI/ EXPERT SYSTEMS

MACHINE LEARNING/ DEEP LEARNING

GENERATIVE AI

Type of AI	Non-Safety Functions	Safety Functions			Non-Driving Functions
		Perception	Planning	Actuation	
Supervised Learning (SL)	GESTURE CONTROL VOICE RECOGNITION	DETECTION OF OTHER ROAD USERS DETECTION OF ROAD INFRASTRUCTURE	TRAJECTORY PREDICTION	N/A	DETECTION OF DRIVER'S EYE GAZE PREDICTIVE MAINTENANCE
Unsupervised Learning (UL)	N/A	EXTRACTING SCENARIOS FOR REAL WORLD DATA FOR VALIDATION GENERATION OF SYNTHETIC DATA	TRAJECTORY PREDICTION (e.g. KALMAN FILTERS, GAUSSIAN PROCESS ARCHITECTURES)	N/A	FAULT DETECTION
Semi-Supervised Learning (SSL)	N/A	STREAMLINING DATA LABELLING PROCESSES FOR LESS SAFETY-CRITICAL SYSTEMS	SHADOW MODE USED IN DEVELOPMENT FOR TRAINING OF CONTROL ALGORITHMS	N/A	
Reinforcement Learning (RL)	N/A	PERCEPTION (EMERGENT)	LANE CENTERING OR ACC SYSTEMS (EMERGENT)	N/A	PREDICTIVE MAINTENANCE

*SOURCE: ECE/TRANS/WP.29/1182, CONSIDERATIONS ON AI IN ROAD VEHICLES, ANNEX II (2024)

Type of AI	Non-Safety Functions	Safety Functions			Non-Driving Functions
		Perception	Planning	Actuation	
Supervised Learning (SL)	[redacted] ROLE OF GENERAL PURPOSE/ GENERATIVE AI?	DETECTION OF OTHER ROAD USERS DETECTION OF ROAD INFRASTRUCTURE	TRAJECTORY PREDICTION	N/A	[redacted] ROLE OF "ONLINE LEARNING"? FROZEN V. UNFROZEN?
Unsupervised Learning (UL)	N/A	EXTRACTING SCENARIOS FOR REAL WORLD DATA FOR VALIDATION GENERATION OF SYNTHETIC DATA	TRAJECTORY PREDICTION (e.g. KALMAN FILTERS, GAUSSIAN PROCESS ARCHITECTURES)	N/A	FAULT DETECTION
Semi-Supervised Learning (SSL)	N/A	STREAMLINING DATA LABELLING PROCESSES FOR LESS SAFETY-CRITICAL SYSTEMS	SHADOW MODE USED IN DEVELOPMENT FOR TRAINING OF CONTROL ALGORITHMS	N/A	
Reinforcement Learning (RL)	N/A	PERCEPTION (EMERGENT)	LANE CENTERING OR ACC SYSTEMS (EMERGENT)	N/A	PREDICTIVE MAINTENANCE

*SOURCE: ECE/TRANS/WP.29/1182, CONSIDERATIONS ON AI IN ROAD VEHICLES, ANNEX II (2024)

New School AV Ethics

(2022 – Present)

Characteristics of New School AV Ethics

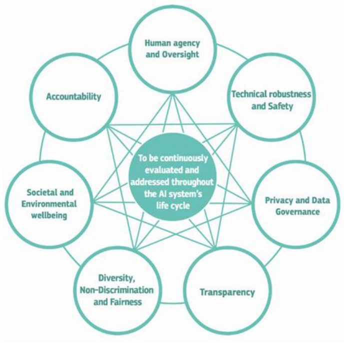
Focus on Use of AI Across Whole Vehicle Lifecycle:

The scope of modern AV ethics covers the use of AI not only in DDT performance (mundane, critical scenarios), but also its role in vehicle system design, development, deployment and use.

+

Focus on Trustworthy & Responsible AI Principles:

Modern AV ethics is informed by the emerging consensus in the horizontal AI ecosystem on the norms and principles that constitute best practice, and is further applied to the specifics of the AV use case.


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

+

Focus on Societal Impact, Trust & Respect of Rights:

The validation of ethical design, development and behaviour in AVs aligns with emerging risk-based AI regulation: it should minimise adverse societal and (human/ fundamental) rights impacts, and promote trust with users and broader society.

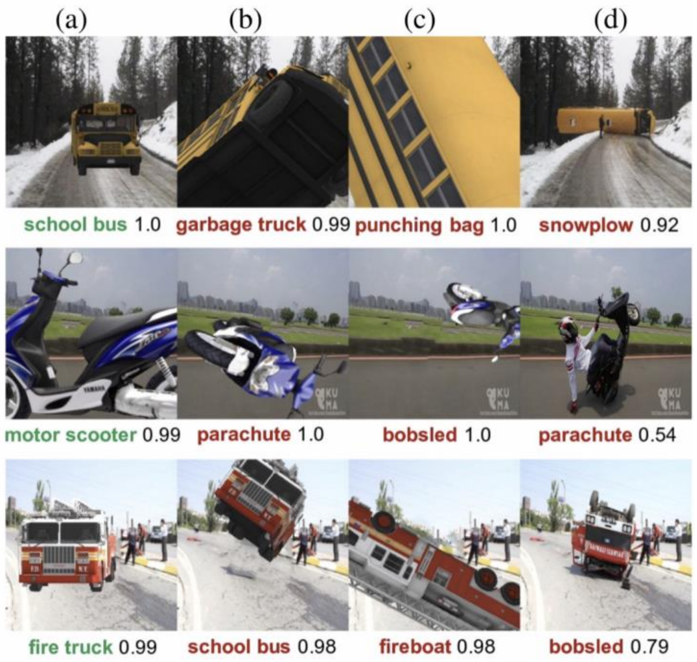
Image sources: <a href='https://pngtree.com/freepng/scales-of-justice-isolated-on-transparent-background_20182018.html'>png image from pngtree.com/</a>;
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

Two New School AV Ethics Problems

The Bias/ Robustness Problem:

What features can we reliably and robustly detect in an autonomous vehicle’s ODD under (even) nominal conditions, and can we robustly specify the target operating domain?

What (mundane) behaviour and movement patterns pose unexpected risks for road users*?



*Strike (with) a Pose: Neural Networks Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al., CVPR 2019.

The Privacy Problem:

How much data (including subjective or personal characteristics of road users) does an autonomous vehicle need to collect to ensure safe operation *generally*, including for:

- Remote operation
- In-service monitoring
- Event Data Recorders (accident reconstruction)
- Vehicle and Device communication (V2V, V2X)
- Passenger surveillance (e.g. attentiveness)
- Data collection for training, scenario definition
- Object & Event Detection & Response

Mozilla Report- Privacy CAVs

How much does a vehicle need to know v. how much should it know to be privacy respecting?

Thank you!