


An abstract graphic on the left side of the slide. It features a vertical color gradient bar transitioning from blue at the top to purple in the middle to magenta at the bottom. Overlaid on this bar and extending into the black background are numerous thin, concentric circles in various colors including blue, green, yellow, and orange, creating a sense of motion or a ripple effect.

Moving Away From the Trolley Problem

A paradigm shift in ethics of
autonomous vehicles (AVs)



The NeuroComputational Ethics Research Group at NCSU

Group aim: To evaluate the ethics of emerging technologies and to utilize technology for the purposes of empirical research.

Current Research:

- VR simulations of moral decision making for AVs
- Studying adolescent morality
- Reviewing the harms and benefits of Transcranial Magnetic Stimulation
- Assessing the usefulness of repurposed non-stimulate ADHD medications for cognitive enhancement.

Issue / Assertion

Current moral psychology methodologies are insufficient for guiding the creation of ethical AVs.

Agent-Deed-Consequence (ADC) operationalized virtual stimuli are a better tool for informing the moral actions of AVs.



Presentation Flow

AV Background

- Categorization of AVs
- Software Architecture

Why are Ethically Acting AVs Necessary?

- Collisions are unavoidable
- Ethics Settings in AVs make fully autonomous versions trustworthy

Sacrificial Dilemmas and their Flaws

- Overview of Sacrificial Dilemmas
- The Moral Machine Experiment (MME) Description
- Binary Choice
- External Validity

Our Solution

- ADC model of moral judgment
- Virtual Reality
- Experimental Design
- Content Display

Conclusion

The background features a complex, abstract pattern of overlapping geometric shapes, primarily triangles and lines, in shades of dark red and green. A vertical color gradient bar is positioned on the left side, transitioning from blue at the top to purple and then to a dark red at the bottom. The text "AV Background" is centered in the lower half of the image, set against a solid black rectangular background.

AV Background

Categorization of AVs

SYNOPSYS®

LEVELS OF DRIVING AUTOMATION



0

NO AUTOMATION

Manual control. The human performs all driving tasks (steering, acceleration, braking, etc.).



1

DRIVER ASSISTANCE

The vehicle features a single automated system (e.g. it monitors speed through cruise control).



2

PARTIAL AUTOMATION

ADAS. The vehicle can perform steering and acceleration. The human still monitors all tasks and can take control at any time.



3

CONDITIONAL AUTOMATION

Environmental detection capabilities. The vehicle can perform most driving tasks, but human override is still required.



4

HIGH AUTOMATION

The vehicle performs all driving tasks under specific circumstances. Geofencing is required. Human override is still an option.



5

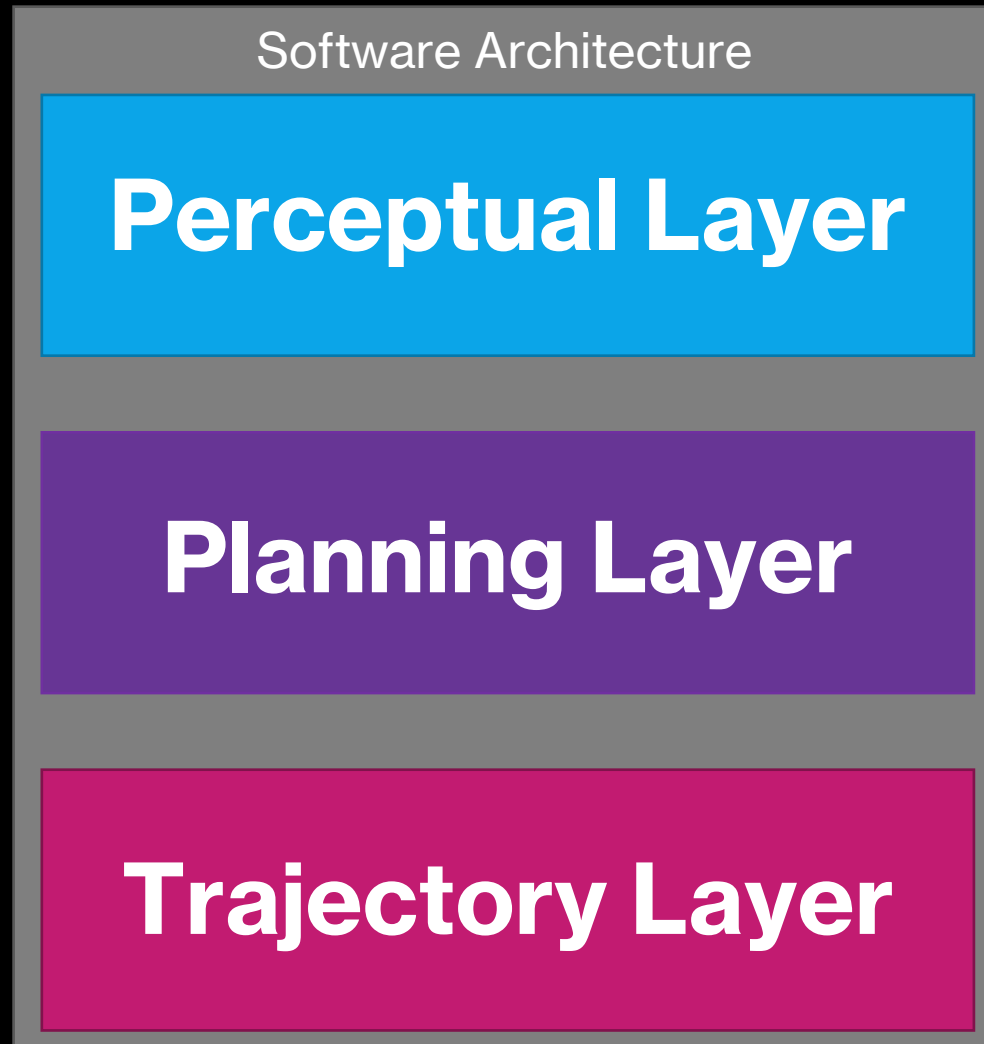
FULL AUTOMATION

The vehicle performs all driving tasks under all conditions. Zero human attention or interaction is required.

THE HUMAN MONITORS THE DRIVING ENVIRONMENT

THE AUTOMATED SYSTEM MONITORS THE DRIVING ENVIRONMENT

AV Software Architecture



Structured like a nervous system.

← Sensory equipment

← Integration of information

Where embedded Ethics Settings (ESs) would be located.

← Motion control



Why are Ethical AVs Necessary?

Collisions

- Collisions are unavoidable part of driving for both non-autonomous and AVs.
 - NHTSA recognized nearly 400 crashes involving level 2 systems (July 2021-December 2022).
- While AV malfunctions are possible, this technologically is likely safer than non autonomous vehicles.
- Therefore, when accidents happen, it is likely that prioritization between unfavorable choices will be required.

Non-automated vehicle crashes (2015)

Table 1. Driver-, Vehicle-, and Environment-Related Critical Reasons

Critical Reason Attributed to	Estimated	
	Number	Percentage* ± 95% conf. limits
Drivers	2,046,000	94% ±2.2%
Vehicles	44,000	2% ±0.7%
Environment	52,000	2% ±1.3%
Unknown Critical Reasons	47,000	2% ±1.4%
Total	2,189,000	100%

*Percentages are based on unrounded estimated frequencies
(Data Source: NMVCCS 2005–2007)

Table 2. Driver-Related Critical Reasons

Critical Reason	Estimated (Based on 94% of the NMVCCS crashes)	
	Number	Percentage* ± 95% conf. limits
Recognition Error	845,000	41% ±2.2%
Decision Error	684,000	33% ±3.7%
Performance Error	210,000	11% ±2.7%
Non-Performance Error (sleep, etc.)	145,000	7% ±1.0%
Other	162,000	8% ±1.9%
Total	2,046,000	100%

*Percentages are based on unrounded estimated frequencies
(Data Source: NMVCCS 2005–2007)

2015. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. A Brief Statistical Summary, Washington, D.C.: National Highway Traffic Safety Administration.

Trust

ABI Model of Perceived Trustworthiness



Created in **BioRender.com**

Trust is a crucial determinate of technological adoption by the public.

A 'correct' set of ESs could drastically increase benevolence and integrity.

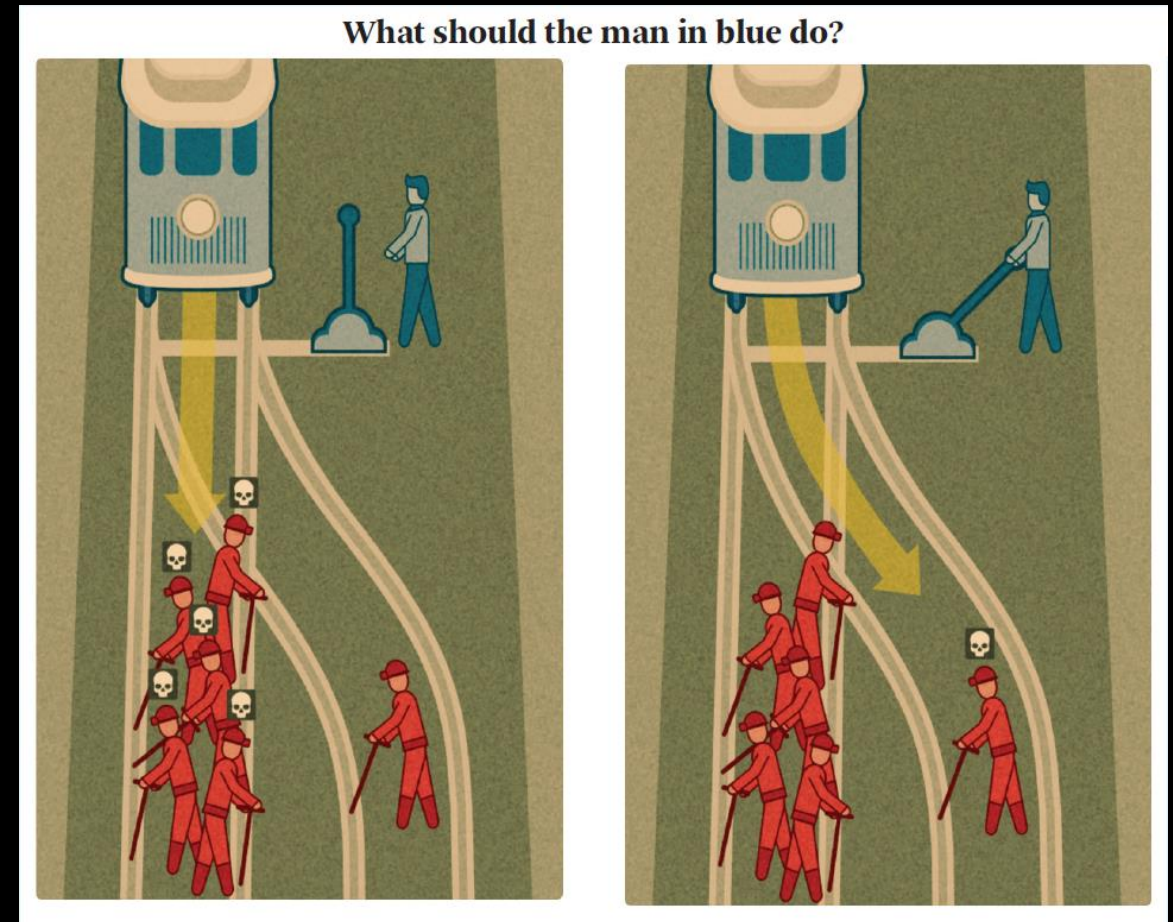
AVs informed by human moral intuition are more likely to be accepted by the public and produce trustworthy actions.



Sacrificial Dilemmas and their Flaws

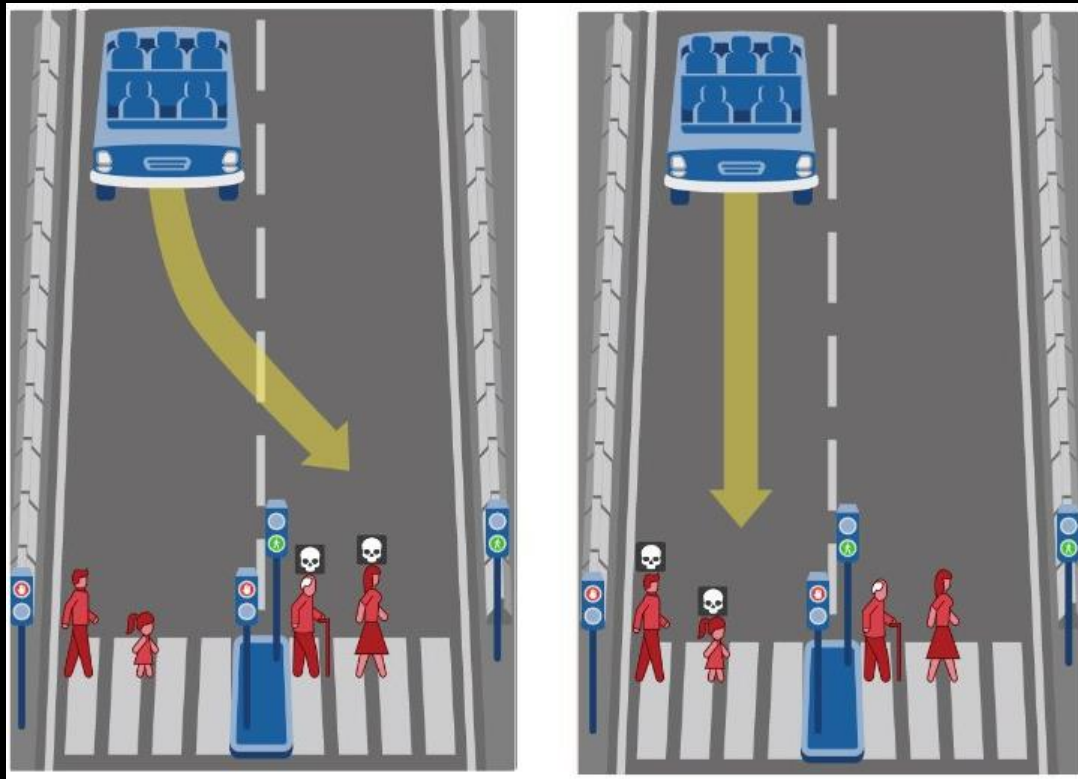
Sacrificial Dilemmas

- The Trolley Problem
 - Proposed in 1967 by Philippa Foot.
 - Popularized by Judith Jarvis Thomson 1985.
- Dilemma's Core Theme
 - Is killing a single person to save more people permissible?
- Pits consequentialist (consequences) and deontological (rules) intuitions against each other.



Moral Machine Experiment

This experiment presents subjects with 13 scenarios varied over 9 factors.



Strongest Preferences Observed

- | | |
|---------------------|----------------------------|
| 1. Humans over pets | 4. Lawful |
| 2. More lives | 5. Perceived social status |
| 3. The young | 6. Physically fit |

Pros

- 39.6 million data points
- 130 countries with at least 100 respondents

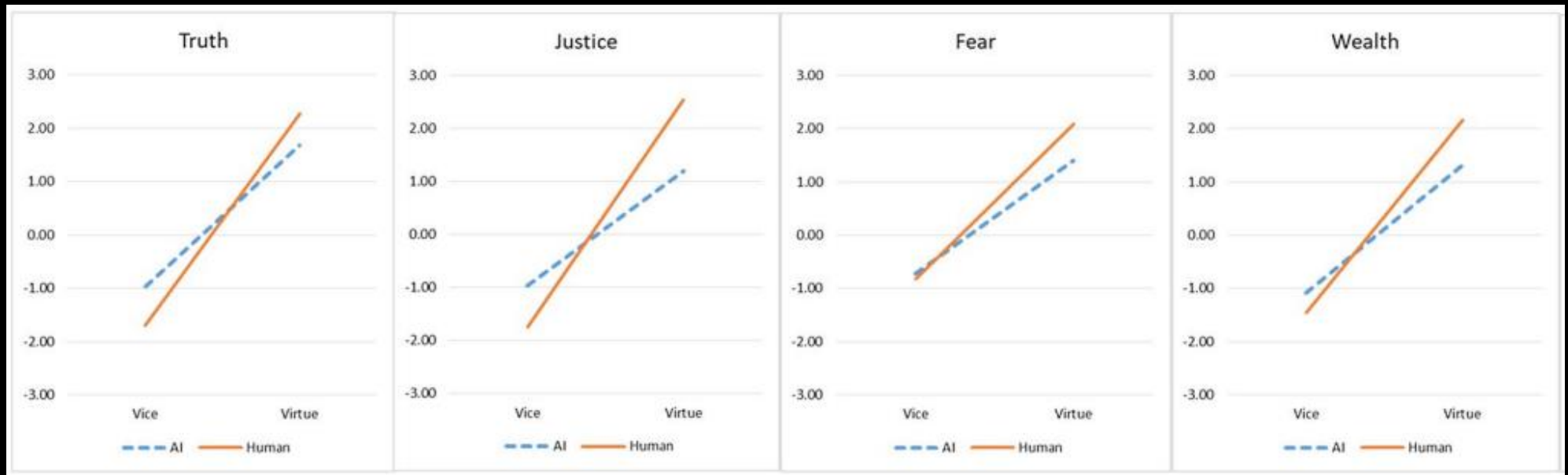
Cons

- Binary
- Lacks external validity (realism)
 1. Experimental
 2. Mundane
 3. Psychological

Binary Choice

Typically, this framework is a good experimental design, but in the sacrificial dilemma paradigm it fails to incorporate other moral theories.

Virtue ethics is a critical component in the formation of the moral evaluations of humans and recent studies suggest people attribute moral traits to AI systems in a similar fashion.

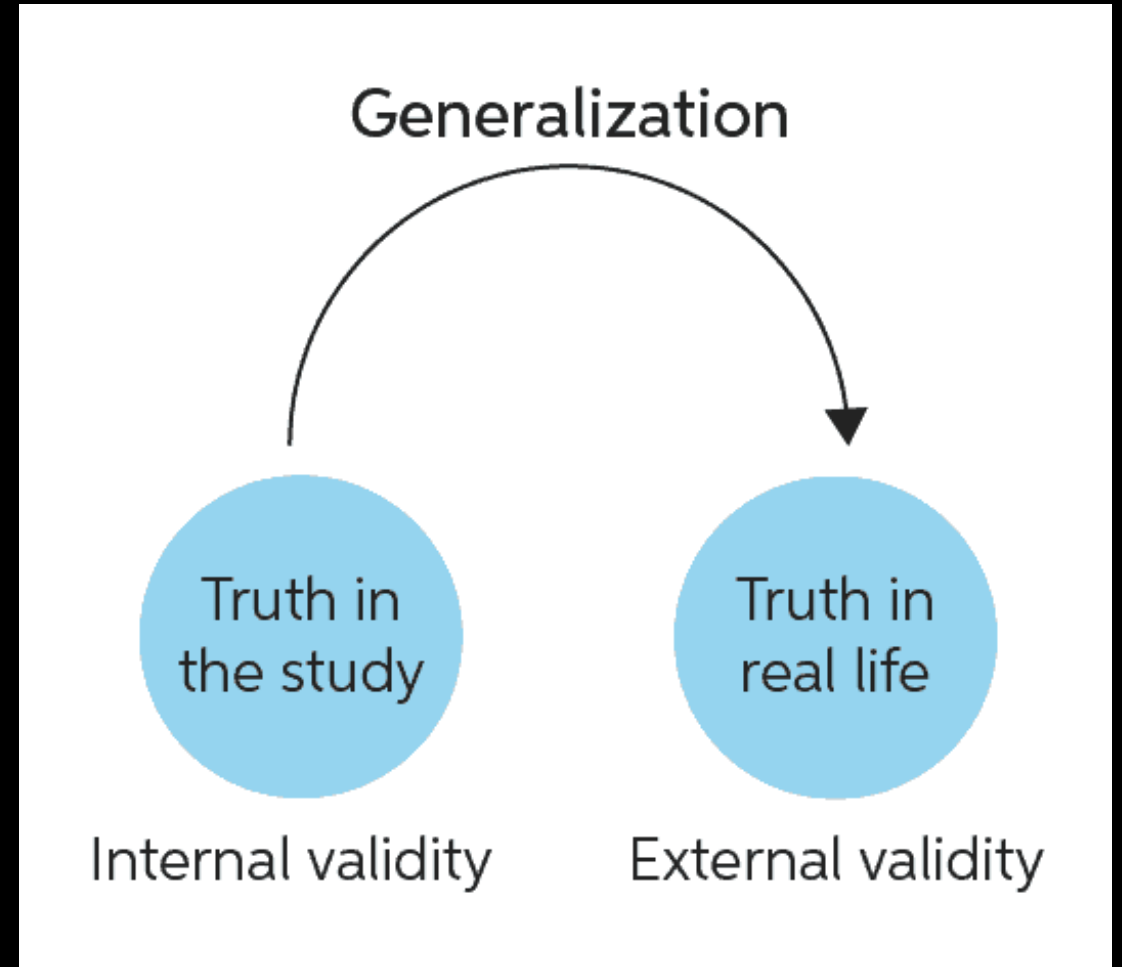


External Validity

Defined as the generalizability of a study's results to other situations.

Sacrificial dilemmas fail to display:

1. Experimental realism - Sometimes elicit a humorous response.
2. Mundane realism - Only includes high stakes scenarios.
3. Psychological realism - Likely do not active the psychological processes as real-world moral judgments.



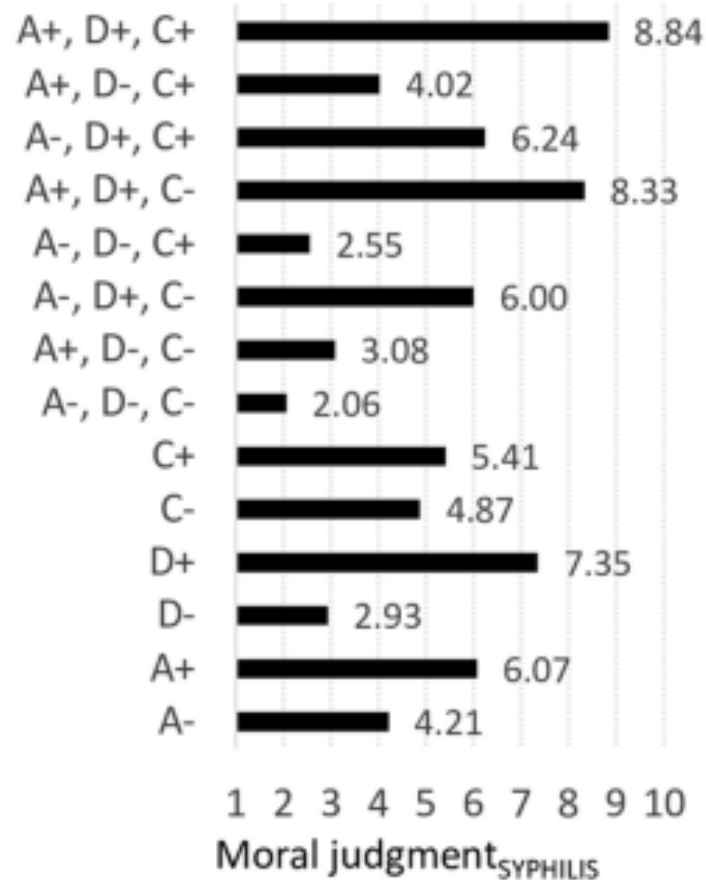
Our Original Solution



ADC Model of Moral Judgment

- Incorporates three central pillars of ethical theory into one model.
 - Virtue Ethics - (Cowardice vs. Bravery)
 - Deontology - (Abiding by rule sets)
 - Utilitarianism - (Greater good outcomes)
- Varies each model aspects over positive / negative axis.
- Allows each aspect to compete equally in the domain of moral evaluation.

Panel 1: Low-stakes
(Experiment 1)



Dubljević, Veljko, Sebastian Sattler, and Eric Racine. 2018. "Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment." *PLOS ONE*

Virtual Reality

Identical task environments except for manipulated stimuli.



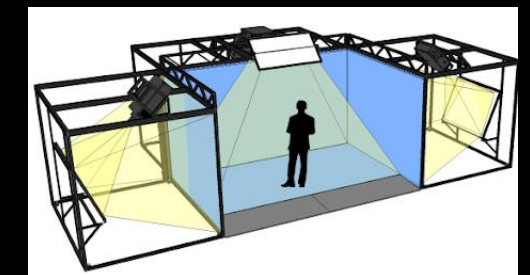
Can safely expose to subjects to a range of situations.



Facilitates the presentation of 3 dimensional audio-visual stimuli.



The same scenario can be displayed across a range of devices.



Experimental Design

We have created six virtual vignettes each consisting of six versions of the same situation.

All versions follow the same structural pattern:

1. Agent displays + / - character trait.
2. Obeys / disobeys some traffic rule.
3. Experiences some vehicle related + / - consequence.

Participants are asked to observe these events and give a moral acceptability rating at the end.

Experimental Design			
Not included +A +D +C	Scenario 1 +A -D +C	Scenario 2 -A +D +C	Scenario 3 -A -D +C
Scenario 4 -A -D +C	Scenario 5 -A +D -C	Scenario 6 +A -D -C	Not included -A -D -C

How morally acceptable was this situation?

Please select an option: 1 (least acceptable) to 10 (most acceptable)

1 2 3 4 5 6 7 8 9 10

If you would like to replay the simulation, then select the button labeled replay to the right

Replay Simulation
Replay

Content Display





Conclusions

- Given the inevitability of collisions, for fully autonomous vehicles to be a trusted part of the traffic community, functional ethics settings are necessary.
- Sacrificial dilemma paradigms need to be abandoned for more realistic methodologies.
- ADC model operationalized virtual stimuli:
 1. Includes virtue ethics (removing binary choice)
 2. Offers realism (increasing experimental and mundane realism)
 3. Utilizes VR (increasing psychological realism while still enabling big data collection)