Stony Brook University

AV ETHICS & The Plurality of Moral Theories

A talk about a problem that emerges when humans build powerful things.

My case in point are AVs (automated vehicles).

Stony Brook University Physics and Astronomy Colloquium, November 14, 2023

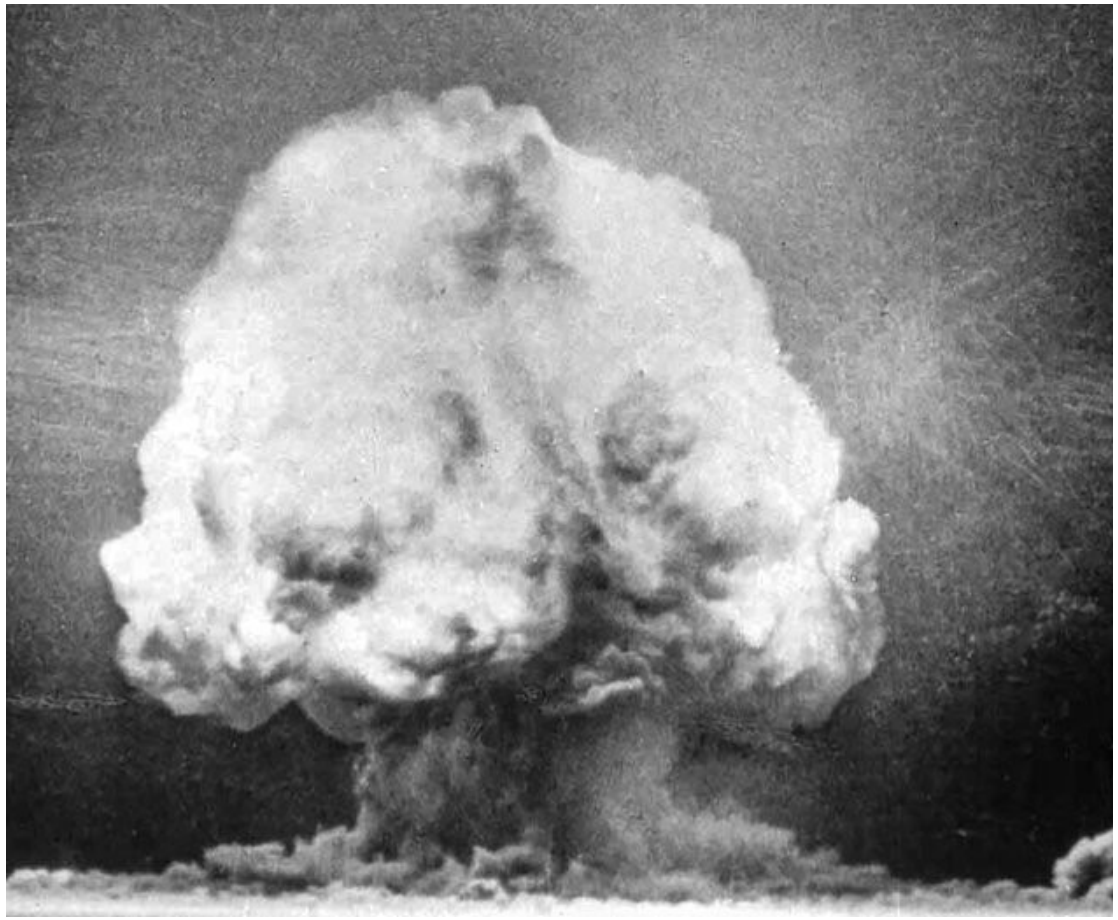
Wolf Schäfer, CEAS Department of Technology and Society

Powerful Things

Since physicists succeeded in building atom bombs, science has become “much too important to be left to the scientists” (Conant).

Contemporary examples of this challenge include genome editing with CRISPR in biotechnology and generative artificial intelligence (AI) with large language models in computer science.

My case in point is the emergence of self-driving cars – far less frightening than nuclear warheads, yet a far-reaching invention for science, technology, and society, nevertheless.



Stony Brook University

Electric Fiction

One wonders how the domino tiles are kept in place? Magnets perhaps.

However, if you look for an iconic image of the 1950s – here you have it. A perfect American family, father and mother, son and daughter, all white, cheerfully playing dominoes under the protective, transparent dome of a slick sports car.

This is a self-driving vehicle, obviously, and it is moving its passengers smoothly toward the relaxed future of automobility.

Who had that vision? A car company? No. A lobby for the US electric light and power companies predicted AVs.

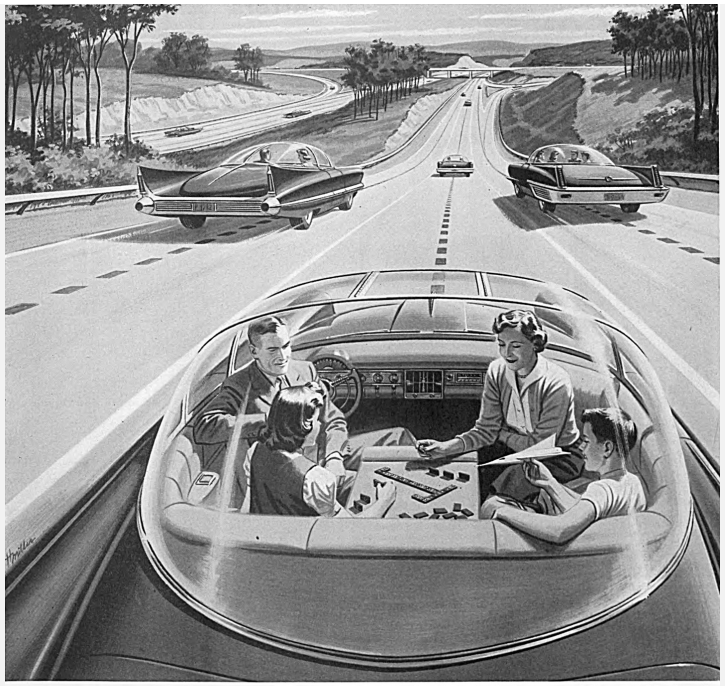
Stony Brook University

e-Irony

A consortium of more than 400 electric power companies placed this ad in the *Saturday Evening Post* in 1956. The subtitle told the future they saw:

- ELECTRICITY MAY BE THE DRIVER. One day your car may speed along an electric super-highway, its speed and steering automatically controlled by electronic devices in the road. Travel will be more enjoyable. Highways will be made safe – by electricity! No traffic jams … no collisions … no driver fatigue.

Wonderful! But ironic too when you realize that electronic road-guidance was featured, and electric vehicles were not. The cars of the ad's "new electric age" were still powered by old-fashioned internal combustion engines.



ELECTRICITY MAY BE THE DRIVER. One day your car may speed along an electric super-highway, its speed and steering automatically controlled by electronic devices embedded in the road. Travel will be more enjoyable. Highways will be made safe—by electricity! No traffic jams . . . no collisions . . . no driver fatigue.

POWER COMPANIES BUILD FOR YOUR NEW ELECTRIC LIVING

Your air conditioner, television and other appliances are just the beginning of a new electric age.

Your food will cook in seconds instead of hours. Electricity will close your windows at the first drop of rain. Lamps will cut on and off automatically to fit the lighting needs in your rooms. Television "screens" will hang on the walls. An electric heat pump will use outside air to cool your house in summer, heat it in winter.

You will need and have much more electricity than you have today. Right now America's more than 400 independent electric light and power companies are planning and building to have twice as much electricity for you by 1965. These companies can have this power ready when you need it because they don't have to wait for an act of Congress—or for a cent of tax money—to build the plants.

The same experience, imagination and enterprise that electrified the nation in a single lifetime are at work shaping your electric future. That's why in the years to come, as in the past, you will benefit *most* when you are served by independent companies like the ones bringing you this message—America's Electric Light and Power Companies*.

No traffic jams…no collisions…no driver fatigue

The irony is resolving now. Not smart roads but smart cars are entering the traffic mix. New Tesla-like vehicles – fully electrical cars augmented with automated driving systems (ADS) – are likely to displace the old ICE population over the next few decades.

The general engineering goal of ***no collisions*** is driving this evolution. Other benefits play a role, but the expectation of a dramatic reduction in road traffic accidents after the transition to AVs was correct in 1956 and is well-founded in 2023.

My students and I assume that car accidents will decline, but still happen, and that societal scrutiny of robot-car fatalities will increase, especially in **edge cases**, where the automotive AI will compute alternative outcomes and decide an action based on incompatible moral theories, such as Utilitarianism and Kantianism (more about these theories later).

Stony Brook University

A Global Burden

The World Health Organization (WHO) – the lead agency for road safety within the UN – released its fourth report on road safety in 2018 with data from 175 countries.* Its Findings:

- Every year the lives of approximately 1.3 million people are cut short because of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability owing to their injury.
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.
- Speeding, driving under the influence of alcohol and other psychoactive substances, distracted driving, unsafe vehicles, and nonuse of motorcycle helmets, seat-belts, and child restraints are key factors of traffic accidents.

The US is not exempt from this burden. The annual data of the American National Highway Traffic Safety Administration (NHTSA) show that.** In 2021,

- there were an estimated 6,102,936 police-reported traffic crashes in which 42,939 people were killed and an estimated 2,497,657 people were injured.
- Approximately 1 person was killed every 12 minutes, and an estimated 5 people injured every minute.
- Compared to 2020 this was a 10-percent increase in the number of traffic fatalities, and a 9.4-percent increase in the estimated number of people injured.

Robot-cars do not speed, drive under the influence, or get distracted. Hence, smart cars are a big step forward.

FAR BEYOND

* See https://www.who.int/publications/i/item/9789241565684. A fifth report will come out in late 2023. See also https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

** Summary of Motor Vehicle Traffic Crashes at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813515.

Stony Brook University

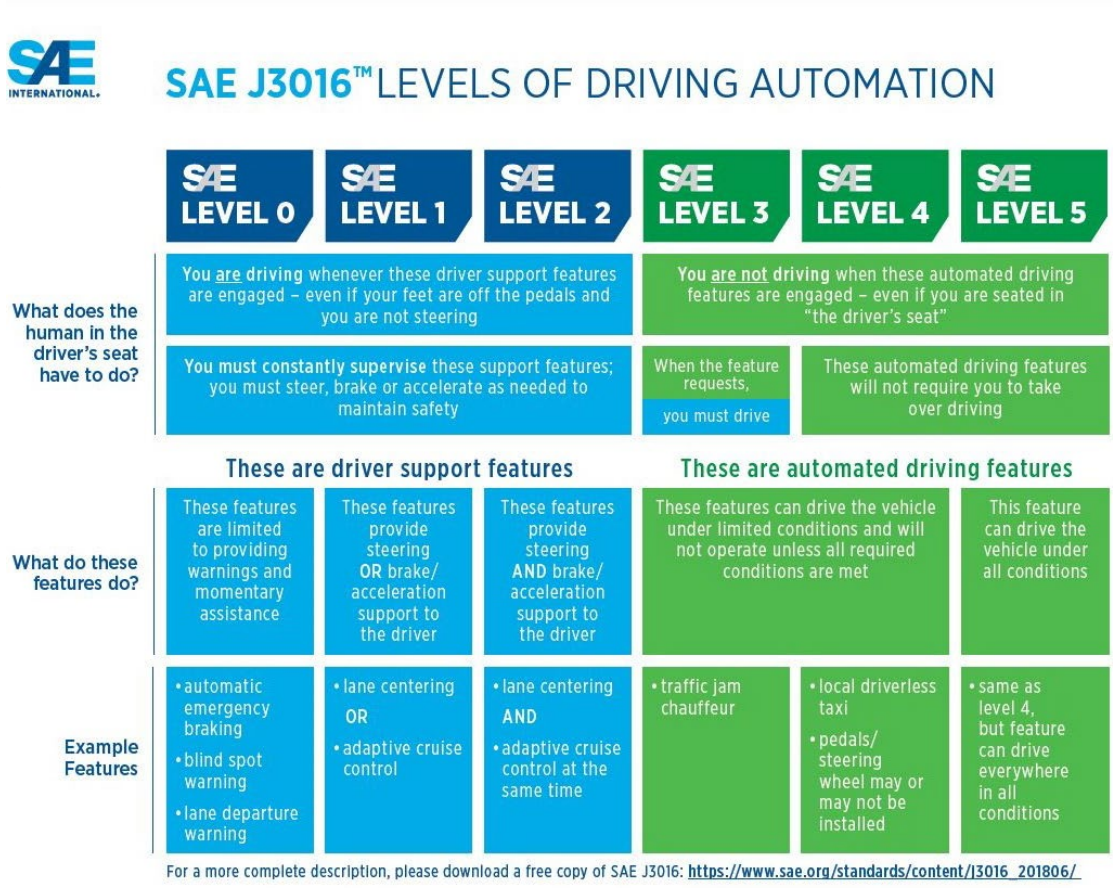
Engineering Fiction Turning Real

A proliferation of smart cars could **halve** the global number of traffic deaths and injuries by 2030 – which is the target of the WHO.

*

The International Society of Automobile Engineers (SAE International) is a global association with over 128,000 members. Its taxonomy of driving automation has been adopted by the NHTSA.

Automation increases over five steps and ranges from no to full automation (Levels 0 to 5). Please note that Level 4 and 5 AVs can come without pedals and steering wheels.



FAR BEYOND

Stony Brook University

Reality Check

"Full Self-Driving" (FSD) is Musk's hyperbole for the driver support features of his Electric Vehicles (EVs). All Tesla cars are **SAE Level 2** since 2020 and the FSD software is in beta mode since then.

Mercedes-Benz is the first automaker that has been awarded **Level 3 certification** in two US states (Nevada and California) for select S-Class and EQS Sedan models.*

Drive Pilot – an Automated Driver Assist System (ADAS) – will be available for these vehicles via subscription starting with model year 2024.

FAR BEYOND

* Kyle Edward, "Mercedes-Benz First To Gain U.S. Approval For Level 3 Automated Driving System." Forbes, 28 Sept. 2023 https://www.forbes.com/sites/kyleedward/2023/09/28/mercedes-benz-first-to-gain-us-approval-for-level-3-automated-driving-system

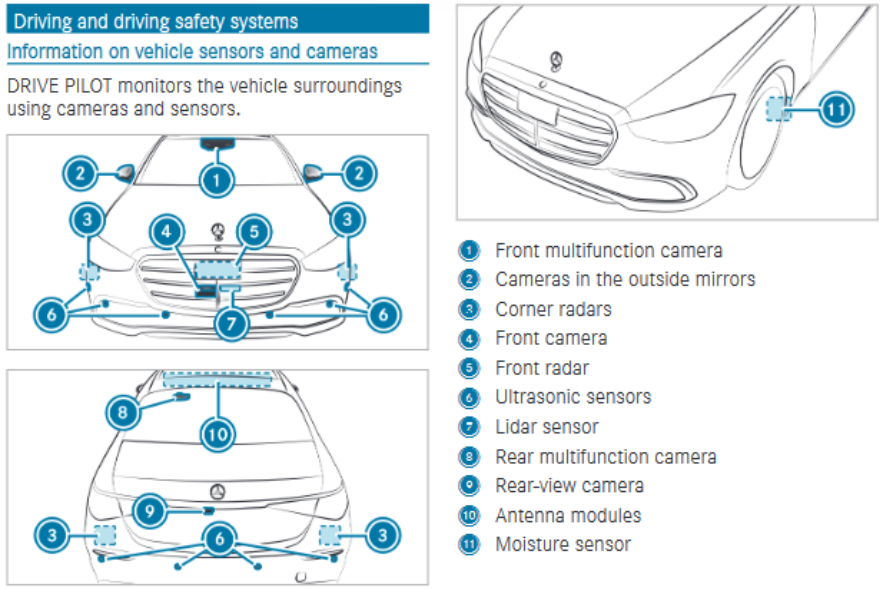
Stony Brook University

Reality Check

"Full Self-Driving" (FSD) is Musk's hyperbole for the driver support features of his Electric Vehicles (EVs). All Tesla cars are **SAE Level 2** since 2020 and the FSD software is in beta mode since then.

Mercedes-Benz is the first automaker that has been awarded **Level 3 certification** in two US states (Nevada and California) for select S-Class and EQS Sedan models.*

Drive Pilot – an Automated Driver Assist System (ADAS) – will be available for these vehicles via subscription starting with model year 2024.



FAR BEYOND

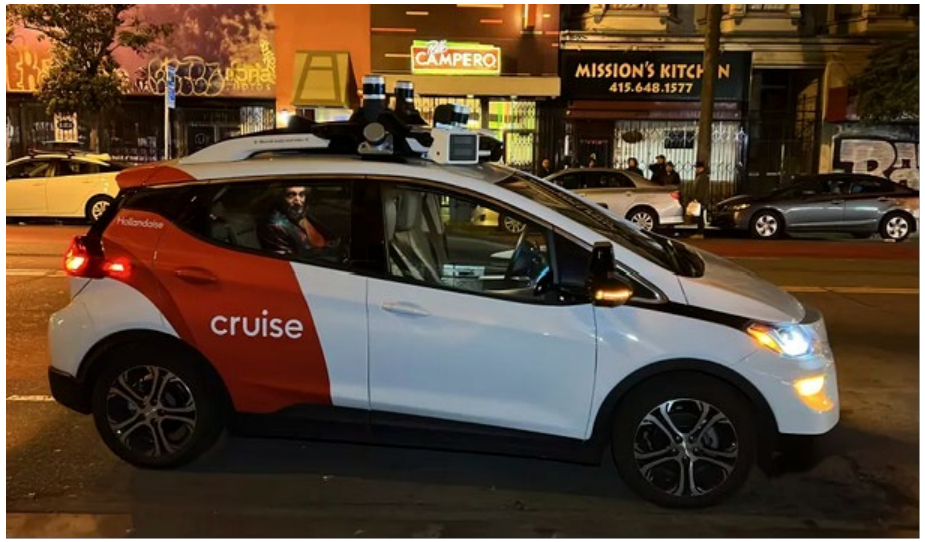
* Kyle Edward, "Mercedes-Benz First To Gain U.S. Approval For Level 3 Automated Driving System." Forbes, 28 Sept. 2023 https://www.forbes.com/sites/kyleedward/2023/09/28/mercedes-benz-first-to-gain-us-approval-for-level-3-automated-driving-system

Stony Brook University

Look Mam, a Waymo. No pedals, no steering wheel

Level 4 and 5 cars are not yet commercially available and won't be anytime soon. Mercedes' CTO Markus Schäfer (no relation) thinks Level 4 may be "doable" by 2030.

Driverless robotaxis, such as Google's **Waymo** and GM's **Cruise**, are Level 4 AVs. They were permitted in San Francisco in August 2023. Two months later, the Cruise permit was pulled after one of its robotaxis had driven itself over a pedestrian and dragged her 20 feet.*



The operating environment of Level 4 is demarcated by an operational design domain (ODD). ODD specifies under which conditions the AV functions safely. It indicates certain traffic, road, environmental, geographical, or time-of-day limitations.

FAR BEYOND

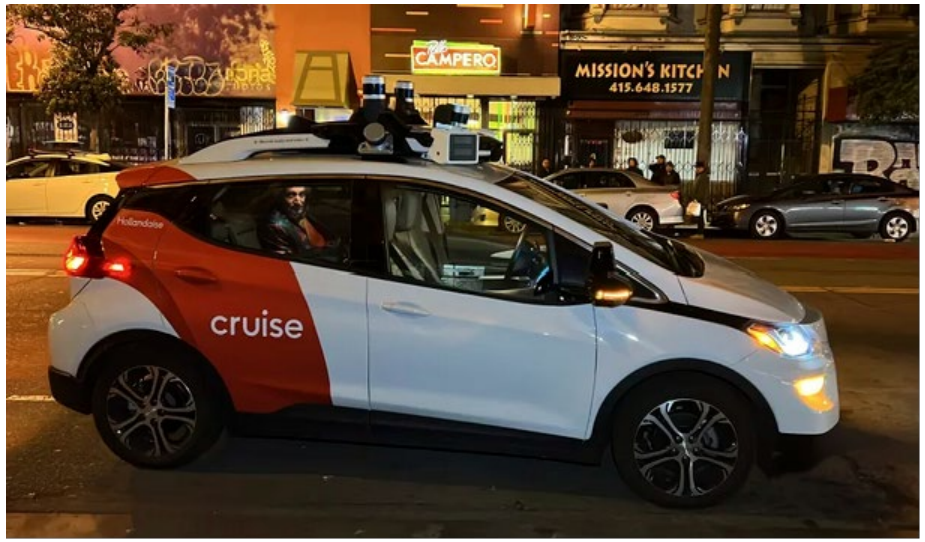
* See Thadani, Trisha: "How a Robotaxi Crash Got Cruise's Self-Driving Cars Pulled from Californian Roads." *Washington Post*, October 30, 2023 https://www.washingtonpost.com/technology/2023/10/28/robotaxi-cruise-crash-driverless-car-san-francisco

Stony Brook University

Look Mam, a Waymo. No pedals, no steering wheel

Level 4 and 5 cars are not yet commercially available and won't be anytime soon. Mercedes' CTO Markus Schäfer (no relation) thinks Level 4 may be "doable" by 2030.

Driverless robotaxis, such as Google's **Waymo** and GM's **Cruise**, are Level 4 AVs. They were permitted in San Francisco in August 2023. Two months later, the Cruise permit was pulled after one of its robotaxis had driven itself over a pedestrian and dragged her 20 feet.*



The operating environment of Level 4 is demarcated by an operational design domain (ODD). ODD specifies under which conditions the AV functions safely. It indicates certain traffic, road, environmental, geographical, or time-of-day limitations.

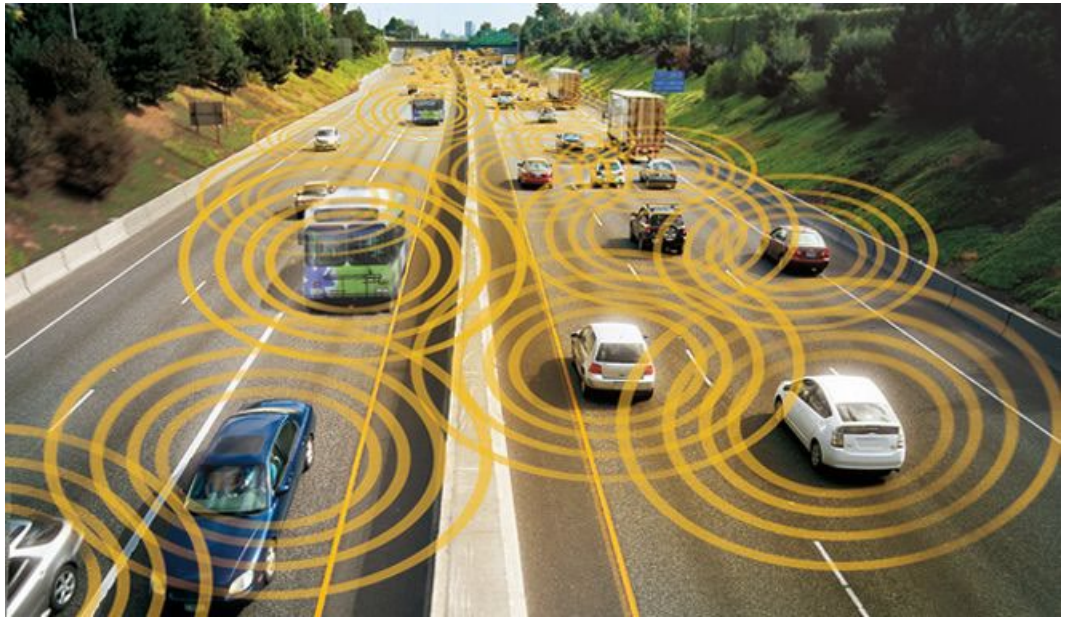
FAR BEYOND

* See Thadani, Trisha: "How a Robotaxi Crash Got Cruise's Self-Driving Cars Pulled from Californian Roads." *Washington Post*, October 30, 2023 https://www.washingtonpost.com/technology/2023/10/28/robotaxi-cruise-crash-driverless-car-san-francisco

Stony Brook University

Connected Autonomous Vehicles (CAVs)

We must assume that Level 4 and 5 vehicles are CAVs with panoptic views of their environment. They will be informed when a vehicle in their vicinity experiences malfunction. This knowledge has ethical and, possibly, legal consequences.



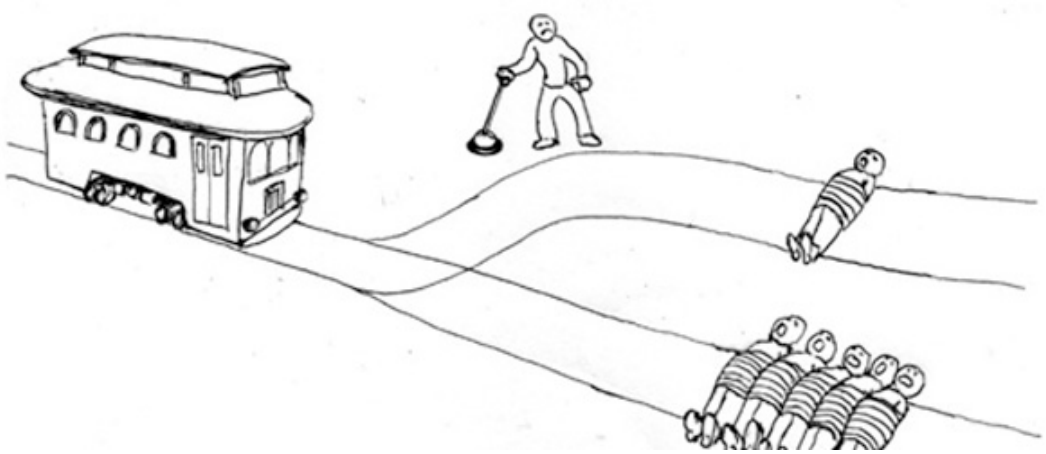
Even lower-level AV accidents already command high media and legal attention. Yet crashes on Levels 4 and 5 will most likely be examined like airplane crashes. Therefore, my students and I are considering this small but super important class of accidents.

Crashes that cannot be avoided, receive top scrutiny, and are the result of an automotive AI choosing between alternative consequences, let's say harming a deer or rear-ending an occupied AV.

Stony Brook University

The Trolley Catch

A trolley has lost its braking function and is barreling downhill. It will kill 5 people tied onto the tracks – unless a bystander switches the trolley over to a sidetrack with 1 person tied up. What should the bystander do? Let 5 people die or pull the lever to divert the trolley and kill 1 person?



The dilemma of the Trolley Problem has become a **moral archetype**, much-discussed in philosophy, medical ethics, and, lately, also automotive ethics.

Alternative accident impacts (ACI) caused by advanced AVs will be too rare to be useful for machine learning. But such events will happen. And they will trigger machine-based decision-making. And that means, moral deliberations must now be written into algorithms.

Consequently, my students and I investigate ACI as a programming challenge.

Stony Brook University

Moral Machine Explorers

The VIP* students in my Automotive Ethics Lab explore moral decision-making in edge cases. They build and program model cars that simulate AV behavior when unavoidable accidents are guided by the automotive AI to unfold differently.

They go from asking, “Should the guy at the lever divert the trolley?” to pondering, “How can we incorporate Kantian ethics and/or Utilitarian ethics into Python?”

They are students of philosophy but not philosophy students. They are engineering students who know their programming languages but realize that engineering has become too important to be left to the engineers.

They tackle the Moral Machine Problem without asking their smartphones to let a random generator app decide what to do in edge cases.

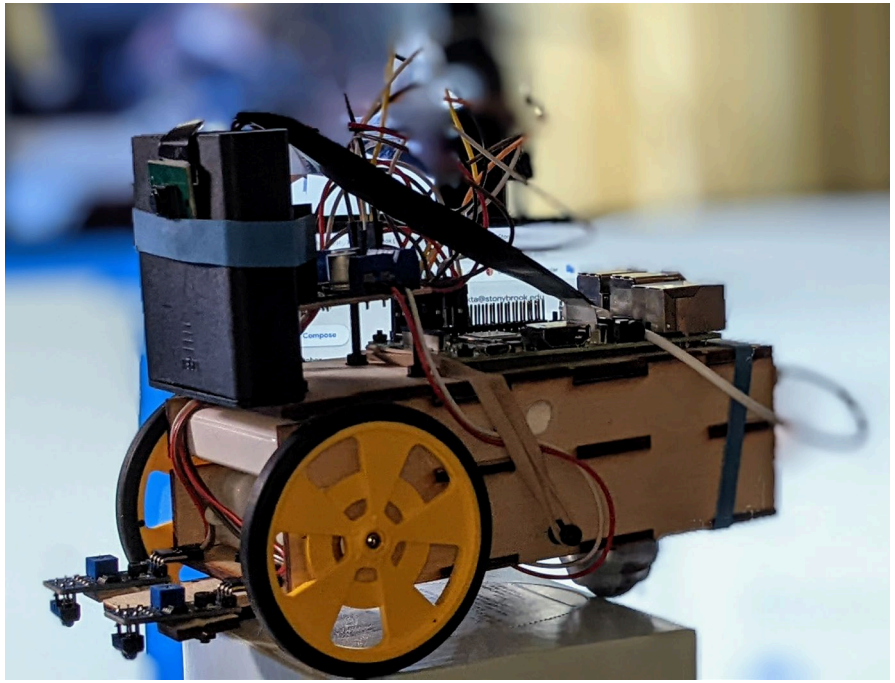
FAR BEYOND

* VIP stands for Vertically Integrated Project and is a new CEAS instruction format. VIP classes include freshman, sophomore, junior, and senior students, plus Master and PhD students.

Stony Brook University

Moral Machine Explorers

The VIP* students in my Automotive Ethics Lab explore moral decision-making in edge cases. They build and program model cars that simulate AV behavior when unavoidable accidents are guided by the automotive AI to unfold differently.



They go from asking, “Should the guy at the lever divert the trolley?” to pondering, “How can we incorporate Kantian ethics and/or Utilitarian ethics into Python?”

They are students of philosophy but not philosophy students. They are engineering students who know their programming languages but realize that engineering has become too important to be left to the engineers.

They tackle the Moral Machine Problem without asking their smartphones to let a random generator app decide what to do in edge cases.

FAR BEYOND

* VIP stands for Vertically Integrated Project and is a new CEAS instruction format. VIP classes include freshman, sophomore, junior, and senior students, plus Master and PhD students.

Stony Brook University

Moral Machine Explorers

The VIP* students in my Automotive Ethics Lab explore moral decision-making in edge cases. They build and program model cars that simulate AV behavior when unavoidable accidents are guided by the automotive AI to unfold differently.

They go from asking, “Should the guy at the lever divert the trolley?” to pondering, “How can we incorporate Kantian ethics and/or Utilitarian ethics into Python?”

They are students of philosophy but not philosophy students. They are engineering students who know their programming languages but realize that engineering has become too important to be left to the engineers.

They tackle the Moral Machine Problem without asking their smartphones to let a random generator app decide what to do in edge cases.

FAR BEYOND

* VIP stands for Vertically Integrated Project and is a new CEAS instruction format. VIP classes include freshman, sophomore, junior, and senior students, plus Master and PhD students.

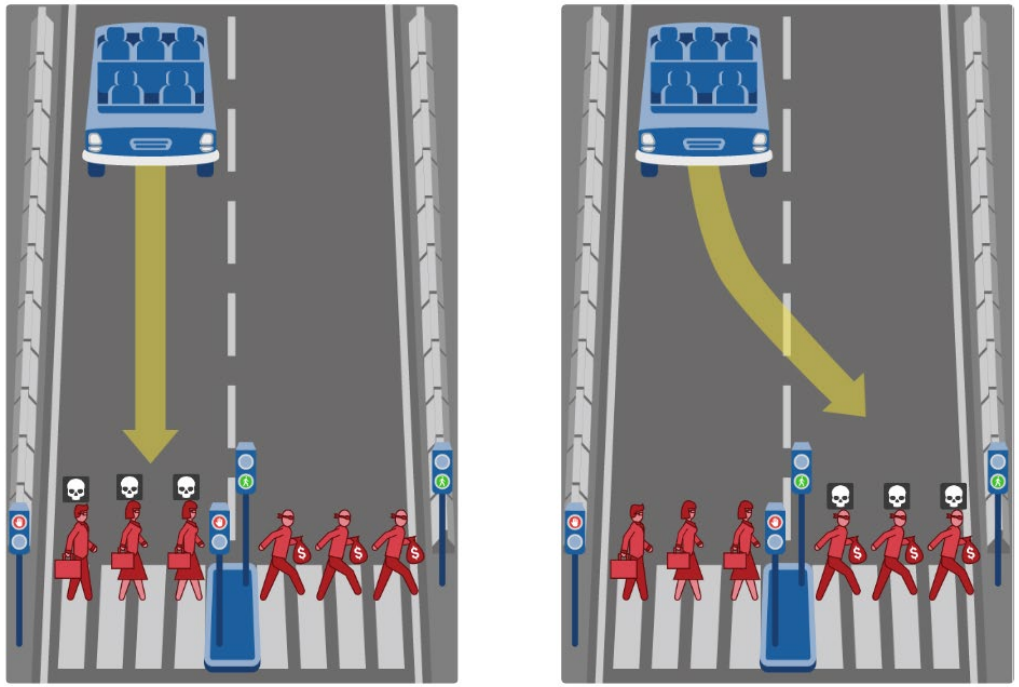
Stony Brook University

MIT's Moral Machine Experiment https://www.moralmachine.net/

The Moral Machine experiment was an online research project that harnessed mass collaboration. People from over 233 countries and territories provided over 40 million ethical decisions in 10 languages.*

The graphic shows one of the **many scenarios** of the experiment. The brakes of an AV failed, but the vehicle has two steering options: Kill 1 male and 2 female executives (left) or kill 3 criminals (right). Which group should the car's AI kill?

Nota bene: *First*, the executives are disregarding a red light, whereas the criminals are crossing legally at green. *Second*, my students would do the right thing and program the empty AV to come to a halt by crashing into one of the concrete barriers lining the road.



The MIT experiment tested the moral programming of people and not AVs.

FAR BEYOND

* Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan, 2018. The Moral Machine Experiment. *Nature*, 563, 59-64. Retrieved from nature.com/articles/s41586-018-0637-6.

Moral Machines should not consider the social circumstances of potential victims, such as their occupation or criminal record.

Abstain from doing harm is our guiding principle.

If harm, however, is unavoidable, then rich or poor, slim or overweight, religion, gender or sexual orientation etc. should not determine AV behavior.

We are not in favor of enabling owners and users of AVs to set their moral preferences. An AI that "brakes for animals" may not brake for humans.

We are equally skeptical of "ethics bots" that learn "specific ethical preferences from a user and subsequently apply these preferences to the user's machine."*

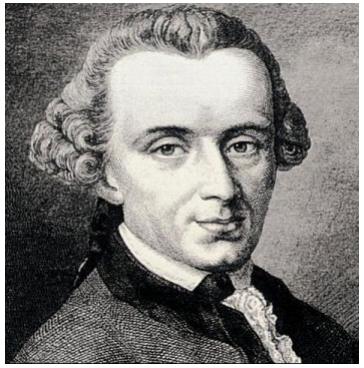
Lastly, we question the application of ethical diversity along national lines, say German AVs with Kantian, American with Libertarian, and Chinese with Utilitarian ethics.

* A. Etzioni & O. Etzioni, Incorporating Ethics into Artificial Intelligence. *Journal of Ethics* (2017) 21: 403-418.

Stony Brook University

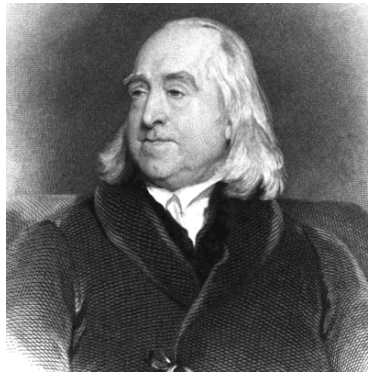
Too Many Theories

Kant's *Categorical Imperative* requires that no person is ever used as a mere means to an end (even if that end is laudable). His ethics also decree that all moral decisions must be capable of universalization to be just.



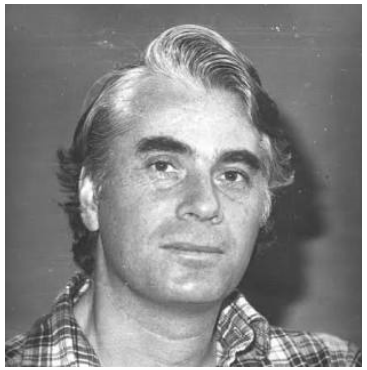
Immanuel Kant
1724-1804

Utilitarianism – **Bentham's moral theory** – seeks the "greatest good for the greatest number." This Greatest Happiness Principle is computational-friendly and allows to sacrifice some lives for a greater good.



Jeremy Bentham
1748-1832

Libertarianism's highest value is the *right to be free*. The concept has French anarchistic origins dating back to the 1850s and was adopted in the US in the 1950s. It defends the right of individuals to their material holdings and asserts that the sole legitimate purpose of government is to protect these rights.*



Robert Nozick
1938-2002

FAR BEYOND

* Libertarianism has a strong following among US billionaires and in the Republican Party, but the Automotive Ethics Lab is not modeling it for AVs. We regard Kantianism and Utilitarianism – or a combination of both – as the most valid theories for industry adoption.

Stony Brook University

Our Goal: A Universal Moral Machine
A rather unlikely prospect at present, perhaps even unattainable

The result of our research so far: The plurality of ethical theories with different moral obligations creates different moral machines with starkly different behaviors. The model cars in our lab demonstrate that: Different moral machines act dissimilar in edge cases.

A Utilitarian AV would harm one person to protect a larger group of people, whereas a Kantian AV might not do that. For instance, an unoccupied Utilitarian car would interfere with an out-of-control vehicle that carries 1 passenger and is about to crash into 5 pedestrians, whereas an unoccupied Kantian vehicle would honor the prohibition against using a person as a *mere means*. It would stop, not hit the malfunctioning car and harm its passenger, but let the out-of-control vehicle proceed towards harming 5 people. That's the problem with multiple moral theories.

Our expectation: AVs with Level 4 and 5 intelligence spur a global effort toward a universal moral regime "beyond" Kantian, Utilitarian, Libertarian or any other parochial ethics.