Automotive Ethics

Professor Dr. Wolf Schäfer
Department of Technology & Society

CFO Roundtable Presentation
Landesbank Baden-Wuerttemberg, New York City
May 9, 2019

Automotive Ethics = AI + E

I am a historian of science and technology who is heading a department of technology and society in a college of engineering and applied sciences at a university dominated by STEM – with other words: I can see that something is missing.

1. When I was born, nuclear physics ruled. But after Hiroshima, the head administrator of the Manhattan Project, James Bryant Conant, concluded, “Science is much too important to be left to the scientists.”

2. Now engineering rules and I have paraphrased Conant’s line for the motto of my dual-purpose department (technology & society), *Engineering has become much too important to be left to the engineers.*

3. Social scientific and humanistic inputs are too often missing from the innovations that push our technology-driven societies from one change to the next.

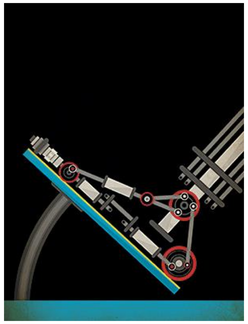
4. The algorithms of modern engineering have spawned disruptors such as Amazon, Airbnb, Facebook, Uber and Tesla – companies who are known for their bold nonconformity, but not their contributions to the Golden Rule (do unto others as you would have them do unto you).

Automotive Ethics is designed to reverse this trend. We already know: Electric vehicles (EVs) are here; autonomous vehicles (AVs) and autonomous electric vehicles (AEVs) are coming, and the latter are driven by AI. Let’s add the E of ethics to their artificial intelligence. **Hence, AI + E**.

AI + E

We have begun to tackle AI + E for autonomous transportation and electromobility at Stony Brook University (SBU).

We are starting this project in conjunction with an exciting new research and teaching initiative in the College of Engineering, called *Vertically Integrated Projects* (VIP). Automotive Ethics is part of this Program.[1]



Our VIP project has three goals:

1. Create a transdisciplinary laboratory for automotive ethics development and testing.
2. Establish an information repository about regional and country-based approaches regulating the behavior of autonomous vehicles.
3. Provide the car industry with evaluation and testing support for embedded autonomous ethics engines.

AI + E is clearly an issue for autonomous vehicles, but also a great challenge on our stage of human history. Allow me to illustrate this with two slides. The first will show the conundrum of humanity's progress; the second the particular achievement of the 4th Industrial Revolution.

By the way: I believe that academia, industry, regulatory agencies, politics and diplomacy must cooperate to achieve universal AI + E in the automotive realm.

[1] See Automotive Ethics: A New Vertically Integrated Projects Program.

Zooming Out: Humanity's Growth

4 global energy transitions (the domestication of: 1. fire, 2. plants & animals, 3. fossil fuels, and 4. nuclear power) have increased Earth's carrying capacity and triggered a series of population explosions:

- **1 million** (hunter-gatherers 10,000 BCE)
- **1 billion** (farmers 1800 CE)
- **2.5 billion** (semi-industrialized world ca. 1950)
- **7.7 billion now**
- **10 billion** (fully industrialized world ca. 2050)
- **100 billion** (nuclear fusion world in 3000?)

and Shrinking Time to Adapt

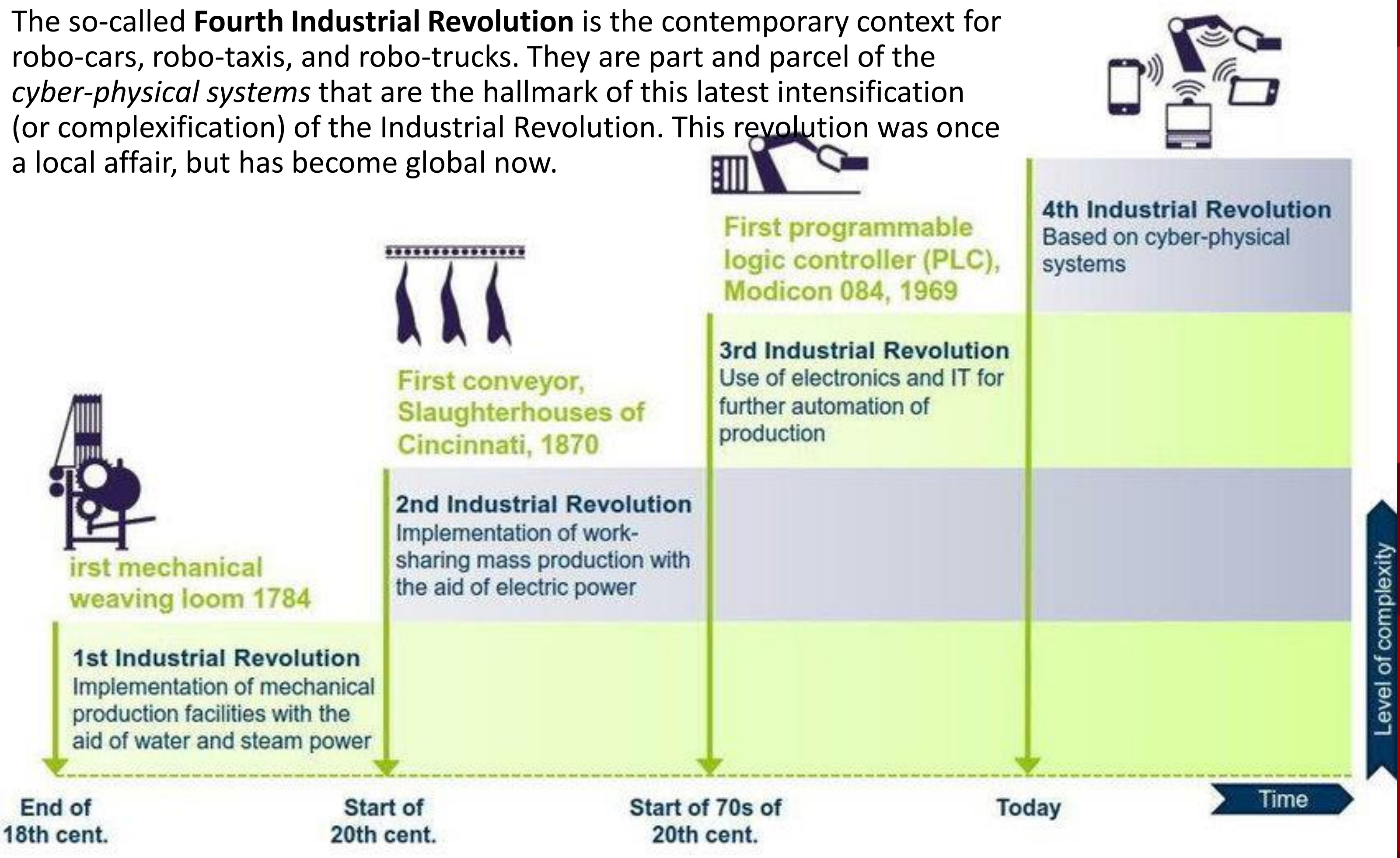
A major consequence of our inventive history: **Disruptive change has to be weathered in ever shorter time spans:**

- Hunter-gatherers foraged over **1 million years** or 90% of human history.
- Agriculturalists developed their economy in **12,000 years**.
- Industrialists had **200 years** so far.
- The nuclear age dawned in the **1940s**.

Bottomline: Humanity's adaptation window has shrunk by six orders of magnitude, from 10^6 to less than 10^1 years – is mankind prepared for fundamental change in ever shorter time spans? I don't think so.

Zooming In: Our Current Context

The so-called **Fourth Industrial Revolution** is the contemporary context for robo-cars, robo-taxis, and robo-trucks. They are part and parcel of the *cyber-physical systems* that are the hallmark of this latest intensification (or complexification) of the Industrial Revolution. This revolution was once a local affair, but has become global now.



irst mechanical weaving loom 1784

1st Industrial Revolution
Implementation of mechanical production facilities with the aid of water and steam power

First conveyor, Slaughterhouses of Cincinnati, 1870

2nd Industrial Revolution
Implementation of work-sharing mass production with the aid of electric power

First programmable logic controller (PLC), Modicon 084, 1969

3rd Industrial Revolution
Use of electronics and IT for further automation of production

4th Industrial Revolution
Based on cyber-physical systems

General Assumptions Tree for AI + E

1. Humanity is creating a global technoscientific civilization (Pangaea Two[1]).
2. Cyber-physical systems are the signature-feature of the Fourth Industrial Revolution.
3. Industry 4.0. aims at the smart design of everything, from genetically engineered fish to cars without steering wheels.
4. Artificial intelligence, biotechnology (CRISPR), nanotechnology, fifth-generation wireless technologies (5G), and quantum computing allow the design of cyber-physical systems across physical, biological, and digital borders.
5. Designing across natural and academic borders requires inter- and transdisciplinarity.
6. SBU, CEAS, DTS are employing/teaching/harnessing team science in our VIP program.
7. AI + E enabled vehicles are but one example of such cross-border convergence.
8. Creating the next disruptive innovation is the leading obsession of technology entrepreneurs.
9. No Holds Barred Disruption has become an aspirational norm and a badge of honor.
10. Ethical responsibility, social harmony, and large-scale sustainability are frequently absent from the initial design of disruption-promising tech projects.
11. AVs and AEVs warrant the inclusion of a moral machine because their AIs have to constantly weigh potentially harmful driving decisions.
12. **AI + E affords controlled disruption.**

[1] See Wolf Schäfer, 2014: “Pangaea II: The Project of the Global Age.” In *Global Challenges in Asia*, ed. by Hyun-Chin Lim, Wolf Schäfer, and Suk-Man Hwang. Seoul, Seoul National University Press, 97-122.

AI + E = Controlled Automotive Disruption

I am optimistic about the automotive industry's drive towards controlled disruption. Being one of the world's leading economic sectors by revenue, it will catch up with small, pioneering disruptors like Tesla and compete on a global scale.

If only for self-preservation and continued leadership, the established automotive sector will pay careful attention to controlling factors, such as

- national and international road safety regulations,
- demands for social and ecological sustainability, as well as
- the algorithmic ethics of AI + E.

My Forecast: The large-scale introduction of AEVs will disrupt established patterns of transportation and mobility worldwide. Yet controlled automotive disruption will also save countless lives. The introduction of highly effective, automatic safety measures will significantly reduce motor vehicle collisions[1] and curb the pollution caused by fossil fuel vehicles.[2] Additional beneficial consequences may be expected in the grey area of raw material procurement.[3]

[1] According to the WHO (*Global Status Report* on road safety 2018), road traffic injury is now the **8th leading cause of death** for all age groups and still increasing. Current total global burden of road traffic deaths: **1.35 million people**.

[2] The UK and France have set the end of gas and diesel vehicles by **2040**; Norway has decreed its respective deadline for **2025**.

[3] Presently, electromobility based on lithium-ion batteries incurs brutal socio-natural costs from the unregulated mining of raw materials (child labor, steep environmental degradation, huge health and safety hazards).

Zwischenruf (heckle) : How does automotive ethics *work*? What *is* automotive ethics?

I am playing the heckler here for a reason. In April 2019, I googled "automotive ethics" and received 62.7 million answers in less than 1 second. And the top answer was: **"to perform high-quality repair service at a fair and just price."** This result was linked to the venerable Code of Ethics of the National Automotive Service Association.

However, **"Automotive Ethics – Stony Brook University"** was also listed on the first results page, and that was pretty good news considering that I had advertised our VIP initiative on a Stony Brook website only two weeks earlier.

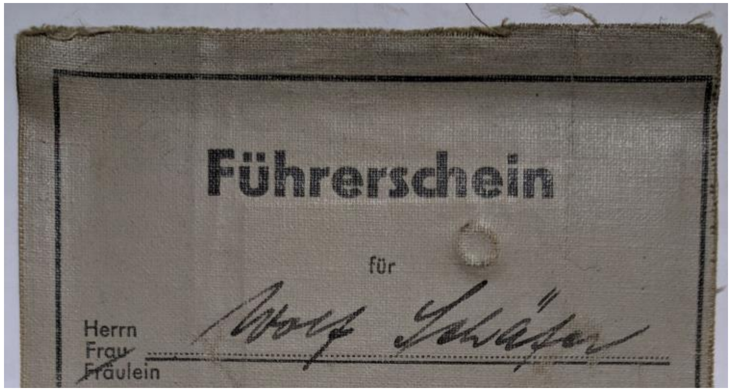
This mixed Google search result is indicative of two things in my view:

> *First*, the strong automotive experience of the last century and its associated concerns are still with us.
>
> *Second*, the understanding of automotive ethics as AI + E is brand new and not yet self-evident. Henceforth, we must answer the heckler's question. And we must admit that automotive ethics is far from ready-made and still a work in progress.

How To AI + E?

The moment to discuss AI + E and make practical progress toward solving it, is Now. Partially automated vehicles are already driving on our streets and highways, yet large parts of the public around the world are still invested in the old paradigm of a human behind the steering wheel.

I myself remember the proud moment of receiving my *Führerschein* (driver's license) in the early 1960s and I am holding on to both – the memory and the floppy license that made me a *Führer* (leader) of a motor vehicle with a combustion engine of classes 1, 3, and 4.



Führerschein

für

Herrn ~~Frau~~ ~~Fräulein~~ Wolf Schäfer

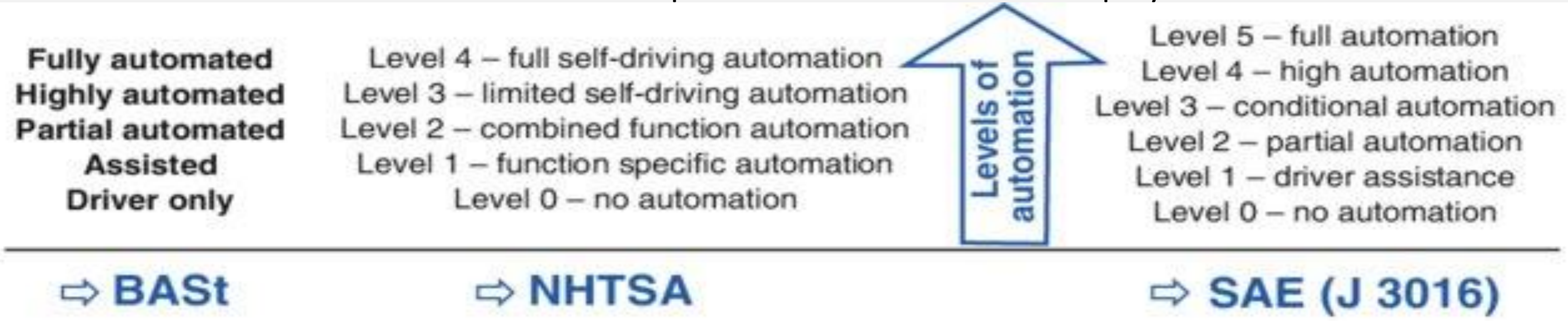
By the way: The opposition against driverless cars is getting organized in the US. In early 2018, a "Human Driving Association" rallied and declared, "The War on Driving Is Here."[1] A few months earlier, the Pew Research Center had reported that 56% of Americans would not want to ride in a driverless car because they do not want to give up control.[2]

[1] See M. R. O'Connor. "The Fight for the Right to Drive." The New Yorker, April 30, 2019. See also "HDA: Human Driving Association." *Human Driving Association*. Accessed May 5, 2019.

[2] See Aaron Smith and Monica Anderson. "Americans' Views on Driverless Vehicles." *Pew Research Center* (blog), Oct. 4, 2017.

From Partially Automated to Fully Automated

One can rightly say, *The race to increase the levels of vehicle automation is on*, because that is true. One can also say, *The AEV revolution is underway*, since that's also true. But claiming the paradigm shift to software-driven vehicles is all but accomplished, is wrong. However, that is what Elon Musk is saying. He pronounced a few weeks ago that a "Tesla fleet" of autonomous robo-taxis will be deployed in 2020.[1] *Consumer Reports* and the recently formed industry-coalition Partners for Automated Vehicle Education (PAVE) immediately intervened and declared that such claims are premature and "not backed up by the data."[2]



⇨ BASt	⇨ NHTSA	⇨ SAE (J 3016)
		Level 5 – full automation
Fully automated	Level 4 – full self-driving automation	Level 4 – high automation
Highly automated	Level 3 – limited self-driving automation	Level 3 – conditional automation
Partial automated	Level 2 – combined function automation	Level 2 – partial automation
Assisted	Level 1 – function specific automation	Level 1 – driver assistance
Driver only	Level 0 – no automation	Level 0 – no automation

Levels of automation defined by the *Bundesanstalt für Straßenwesen* (BASt), the National Highway Traffic Safety Administration (NHTSA), and the international Society of Automotive Engineering (SAE).

[1] See Russ Mitchell. "Elon Musk Claims a Million Teslas Will Drive Themselves in a Year. Safety Advocates Have Concerns." latimes.com, 22 April 2019.

[2] "Tesla Must Prove Safety Before Claiming 'Self-Driving' Ability." Consumer Reports, April 22, 2019. See also Bill Visnic. "PAVE Coalition Formed," January 8, 2019, sae.org/news/2019/01/pave-coalition-announcement.

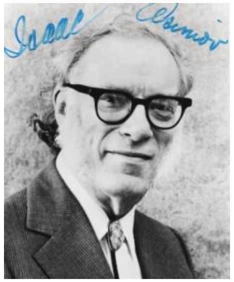
AI + E: Problems

Human thinking about ethics has no problem entering into a discourse about the previously mentioned values of responsibility, social harmony, and large-scale sustainability. Philosophers address these things all the time. However, when an engineer sets out to program an automotive robot for *responsible* driving, all sorts of problems emerge.

Thought experiment: A SAE Level 5 vehicle (fully automated but still a distant dream car) cannot avoid crashing into one of two cyclists, yet it can select which one: the one, who is wearing a helmet, or the other one, who is not. What is the "responsible" choice? Hitting the protected or the unprotected cyclist? The answer seems to be: *Target the person that is more likely to survive.* Hence, the carefully programmed car hits the guy with the helmet and not the other who might die. Now, the autonomous car has saved a life, yet also penalized lawful and prudent road behavior. Furthermore, this "responsible" crash-optimization program may have now set the stage for more cyclists to forego helmets, because doing so has become safer with respect to the optimal targeting choice of the autonomous car.[1]

Writing our moral values into the algorithms of autonomous cars is necessary and unavoidable, but also fraught with uncertainty and unexpected consequences.

[1] For an excellent philosophical discussion of these kinds of moral problems, see Patrick Lin, 2016: "Why Ethics Matters for Autonomous Cars" at https://link.springer.com/chapter/10.1007/978-3-662-45854-9_4.



AI + E: An Early Answer

Normative distinctions, such as good or bad, right or wrong, are preferences that can be *determined* by democratic consent or authoritarian diktat, whereas **technical values**, such as correct or incorrect, can be *discovered* by science and engineering. One of the first attempts to resolve this notorious tension for the sake of intelligent robots was formulated in 1942 by science fiction writer **Isaac Asimov** in his **Three Laws of Robotics**:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*[1]

[1] See Andrew Liptak. "Isaac Asimov and the Three Laws of Robotics." Kirkus Reviews, March 14, 2013.

Asimov's laws extend humanity's religious and philosophical rules-tradition – which includes **Do No Harm** – from fellow human beings to robots, the holders of artificial intelligence. That works well as a general guideline. Yet problems with Asimov's AI + E solution emerge, when we ask, What exactly counts as harm? Does lying, trespassing, or intrusion of a person's privacy constitute "harm"? What about if "harm" is only *likely*? Can't any action or inaction result in some kind of "harm"? And what happens when all available choices are harmful? Should the robot stop dead? There is only one conclusion: **We have to specify exact thresholds to save robots from debilitating paralysis.**[2]

[2] See Derek Leben. *Ethics for Robots: How to Design a Moral Algorithm*. Abingdon, Oxon; New York, NY: Routledge, 2018, p. 2.

AI + E: A Current Answer

Recent work on automotive ethics has applied Asimov's laws to robo-vehicles:

1. *An automated vehicle should not collide with a pedestrian or cyclist.*
2. *An automated vehicle should not collide with another vehicle, except where avoiding such a collision would conflict with the First Law.*
3. *An automated vehicle should not collide with any other object in the environment, except where avoiding such a collision would conflict with the First or Second Law.*[1]

[1] See J. Christian Gerdes and Sarah M. Thornton. "Implementable Ethics for Autonomous Vehicles." In *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, edited by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, 87–102. Springer: Berlin Heidelberg, 2015, doi.org/10.1007/978-3-662-45854-9_5.

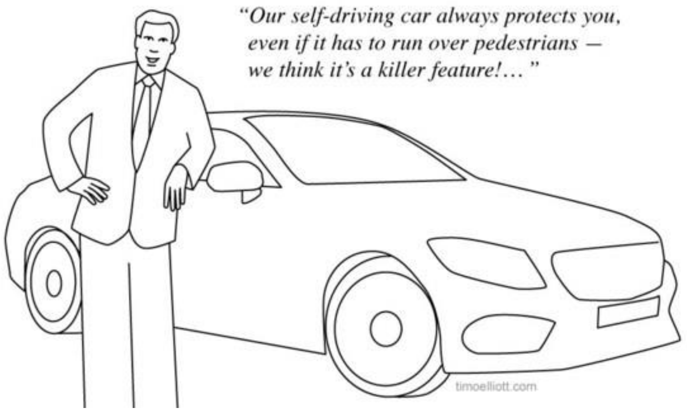
The term of art philosophers use for strict obligations like those above is *deontic*. (Kantian ethics would be an example of a deontic moral theory.) The problem with deontology is its moral rigor. **Deontological rules leave no wiggle room for negotiation or situational adaptation.**

Consequentialism stands opposite deontology and argues that **any action or inaction that produces a good outcome is morally right**. (American pragmatist philosopher John Dewey could be called a consequentialist).

Gerdes and Thornton use both frameworks – deontology plus consequentialism – to impart optimal ethical behavior into automotive AI. To achieve this goal, they map both theories onto mathematical programs that control the vehicle's decision-making.

AI + E = Mathematics + Philosophy

Optimal control theory is a mathematical method “directly analogous to consequentialist approaches in philosophy” (Gerdes & Thornton). It translates the ethical implications of all actions and inactions into **cost functions**, which in turn allow the software controller to reward desired and penalize undesired actions. This method can guide a rocket to its destination and give an autonomous vehicle smooth steering and safe breaking. Calculating the cost of property damage versus personal injury, or the difference between occupant versus pedestrian protection, is in its reach. The hairy problem is neither mathematical nor philosophical, but rather the development of acceptable cost functions.

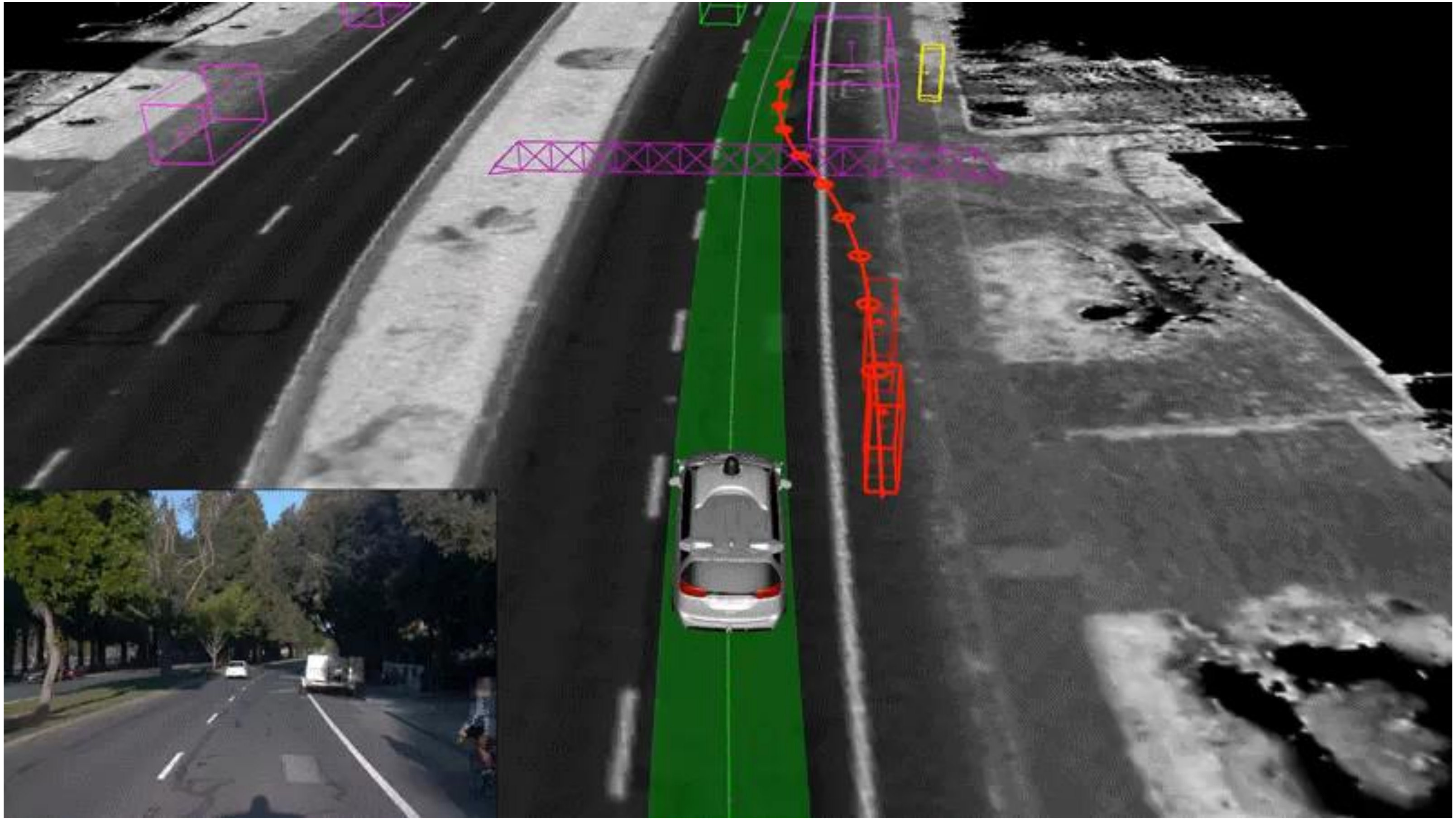

“Our self-driving car always protects you, even if it has to run over pedestrians — we think it’s a killer feature!… ”
timoelliott.com

In 2016, a Daimler manager was quoted saying, “If you know you can save at least one person, at least save that one. Save the one in the car ... If all you know for sure is that one death can be prevented, then that’s your first priority.” The media exploded. *Car and Driver* headlined, “Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians.”[1] There’s a lesson here. “Morality by math” (Patrick Lin) is not natural; it has to be researched, taught, and explained. We must make the mathematical morality of control algorithms understandable and transparent.

[1] See Car and Driver, 7 Oct. 2016.

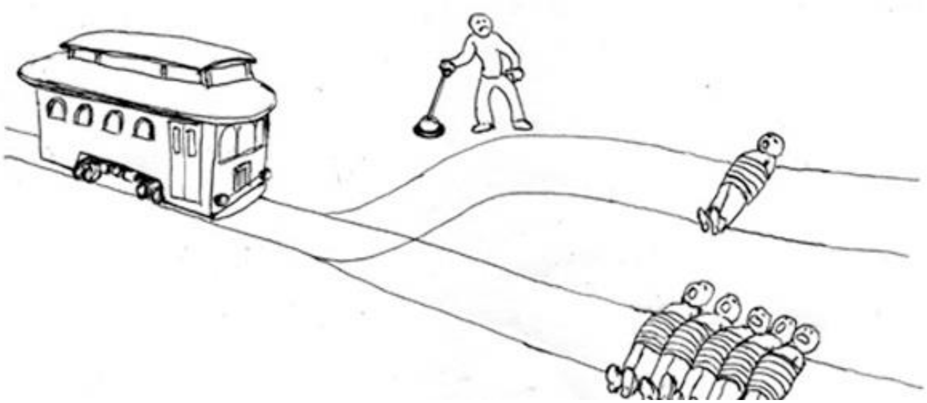
Math Morality Made Transparent

See at youtube.com/watch?v=NG_O4RyQqGE



The video shows how the collision with two cyclists is prevented by a programming switch from consequentialist to deontic logic. Instead of making the potential collision a million times more costly than a lane change, the programmer has put a firm constraint on the optimization process as illustrated by the purple crossbar blocking both lanes. This bar is an encoded hard constraint, i.e. a mathematical imperative instructing the AI: *Do Not Pass These Cyclists*.

Lethal Moral Math



Yes, this is the famous/infamous[1] **Runaway Trolley Problem** depicting a dilemma situation that asks you to make a life and death decision: 5 will live and 1 will die, or 1 will live and 5 will die. Take the human at the switch away and make the trolley autonomous, then the vehicle's AI is forced to make the lethal decision. Guess what!

Nobody dies, the programmer may say, *because the trolley's AI will face the deontic crossbar in its code and stop before the split in the track*. – Well, that smart answer is our engagement problem.

Original Equipment Manufacturers (OEM) and AI engineers are reluctant to engage with thought experiments about lethal dilemmas. They are chasing the science fiction goal of total safety in which accidents have become a thing of the past. But dilemma-style situations may occur, robots will be hacked, and overrides can fail. Hence: **We must investigate and openly discuss the moral math of programming autonomous vehicles.**

[1] Heather Roff has strongly argued that the Trolley Problem is misleading and that one has to take "the technology on its own terms," especially the math of Partially Observed Markov Decision Processes (POMDP). See Heather M. Roff. "The Folly of Trolleys: Ethical Challenges and Autonomous Vehicles." *Brookings* (blog), Dec. 17, 2018.

Collaborative AI + E

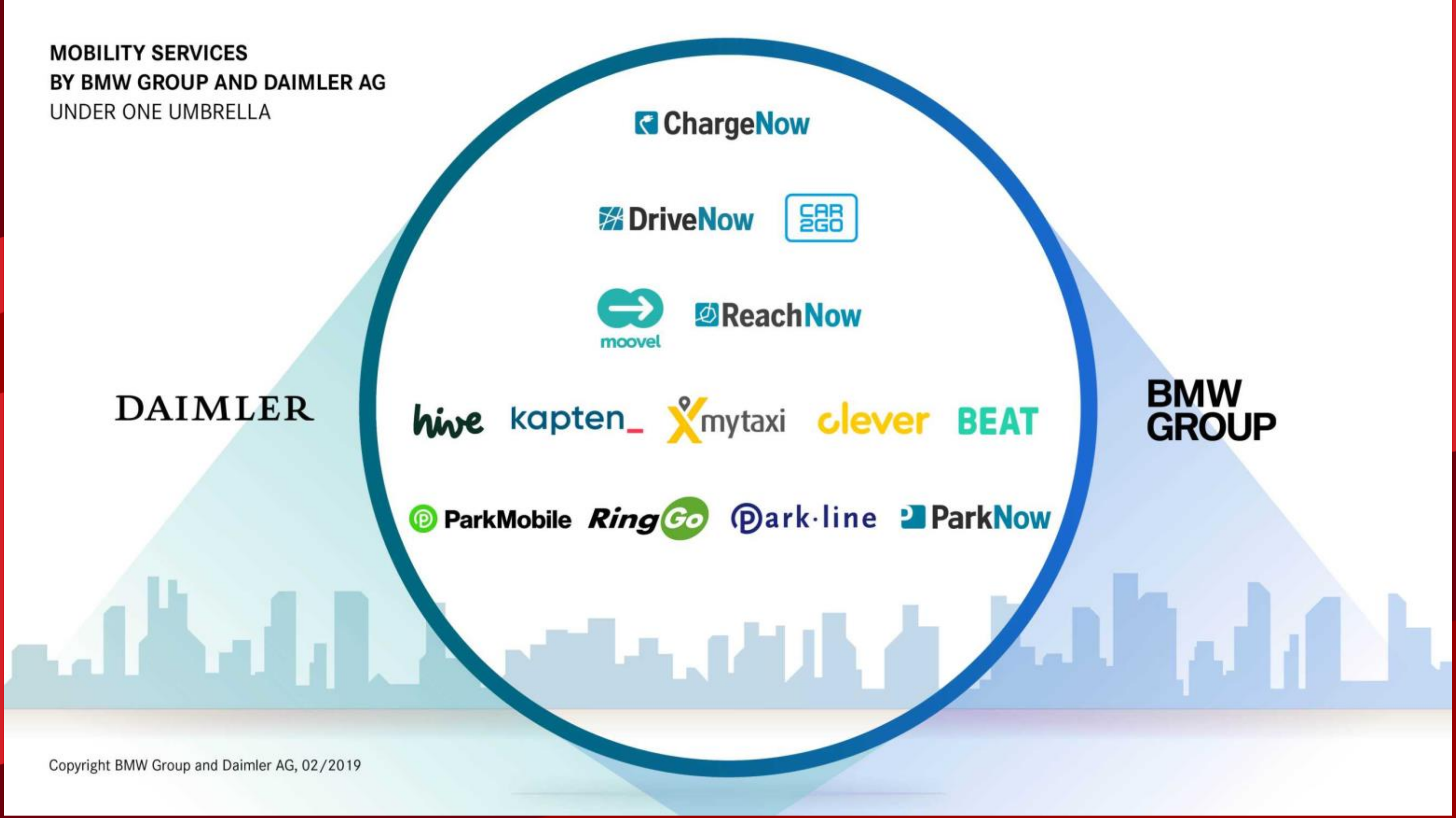
As I said in the beginning, academia, industry, politics, and diplomacy must cooperate to achieve universal AI + E for autonomous mobility.

- Similar to our VIP team, the Automotive Ethics Lab at Stony Brook will cut across academic disciplines.
- I am seeking support from the university and the automotive industry for my lab and graduate students, but will reach out to politics and diplomacy as well. The borders we must transcend are not only disciplinary, but also national and cultural.[1]
- Our ethics orientation will be universal and not particular. I don't think we want to give autonomous car buyers a menu of different ethics choices. Or build American cars with libertarian and German cars with Kantian preferences.[2]
- I firmly believe that the best approach to automotive ethics is **not proprietary but collaborative**. As a mobility customer of car2go, I was excited by the decision to bundle Daimler's car2go and BMW's DriveNow operations. In fact, I take this win-win move as a model for the support structure of universal automotive ethics.

[1] See Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563 (24 October 2018): 54–64, https://www.nature.com/articles/s41586-018-0637-6.

[2] See Patrick Lin. "Here's a Terrible Idea: Robot Cars With Adjustable Ethics Settings." *Wired*, 18 August 2014. For a defense of this possibility, see Tom Fournier. "Will My next Car Be a Libertarian or a Utilitarian? Who Will Decide?" *IEEE Technology and Society Magazine*, vol. 35, issue 2, June 2016.

Exemplary Mobility Partnering



MOBILITY SERVICES
BY BMW GROUP AND DAIMLER AG
UNDER ONE UMBRELLA

Copyright BMW Group and Daimler AG, 02/2019

AI + E Partnering For Success

Imagine we could say:

Wir bündeln unsere Kräfte und erschaffen gemeinsam eine weltweite Lösung der ethischen Herausforderungen autonomer Fahrzeuge. We bundle our strengths and come together to create a worldwide solution to the ethical challenges of autonomous vehicles.

Joint industry funding of AI + E will create a unique win-win situation for both industry and academia.

Imagine the photo opportunities...

Berlin, 22 February 2019: Harald Krüger, BMW, and Dieter Zetsche, Daimler, shake hands over their joint car-sharing venture.

Thank You!