

MAI4CAREU

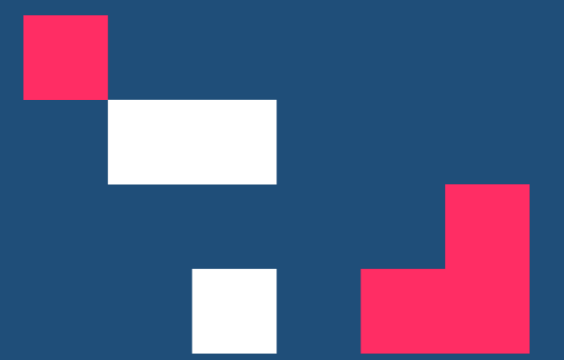
Master programmes in Artificial
Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

Daniela Tafani

2022/2023 – Second Semester



 **Co-financed by the European Union**
Connecting Europe Facility



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



3 – Learning material

Do Self-Driving Cars Have a Trolley Problem? On the immorality of the “Moral Machine”



The trolley problem

“Suppose that a judge or magistrate is faced with rioters demanding that a culprit be found for a certain crime and threatening otherwise to take their own bloody revenge on a particular section of the community. The real culprit being unknown, the judge sees himself as able to prevent the bloodshed only by framing some innocent person and having him executed. Beside this example is placed another in which a pilot whose aeroplane is about to crash is deciding whether to steer from a more to a less inhabited area.”

P. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, «Oxford Review», V, 1967, pp. 5-15.

“To make the parallel as close as possible it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed.

In the case of the riots the mob has five hostages, so that in both the exchange is supposed to be one man’s life for the lives of five.

The question is why we should say, without hesitation, that the driver should steer for the less occupied track, while most of us would be appalled at the idea that the innocent man could be framed.”

P. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, «Oxford Review», V, 1967, pp. 5-15.

Philippa Foot's solution of the trolley problem

“Let us speak of **negative duties** when thinking of the obligation to refrain from such things as killing or robbing, and of the **positive duty**, e.g., to look after children or aged parents. It will be useful, however, to extend the notion of positive duty beyond the range of things that are strictly called duties, bringing acts of charity under this heading.”

It is interesting that, even where the strictest duty of positive aid exists, this still does not weigh as if a negative duty were involved. It is not, for instance, permissible to commit a murder to bring one's starving children food. If the choice is between inflicting injury on one or many there seems only one rational course of action.

If we are bringing aid (rescuing people about to be tortured by the tyrant), we must obviously rescue the larger rather than the smaller group. It does not follow, however, that we would be justified in inflicting the injury, or getting a third person to do so, in order to save the five. We may therefore refuse to be forced into acting by the threats of bad men. **To refrain from inflicting injury ourselves is a stricter duty than to prevent other people from inflicting injury**, which is not to say that the other is not a very strict duty indeed.”

P. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, «Oxford Review», V, 1967, pp. 5-15.

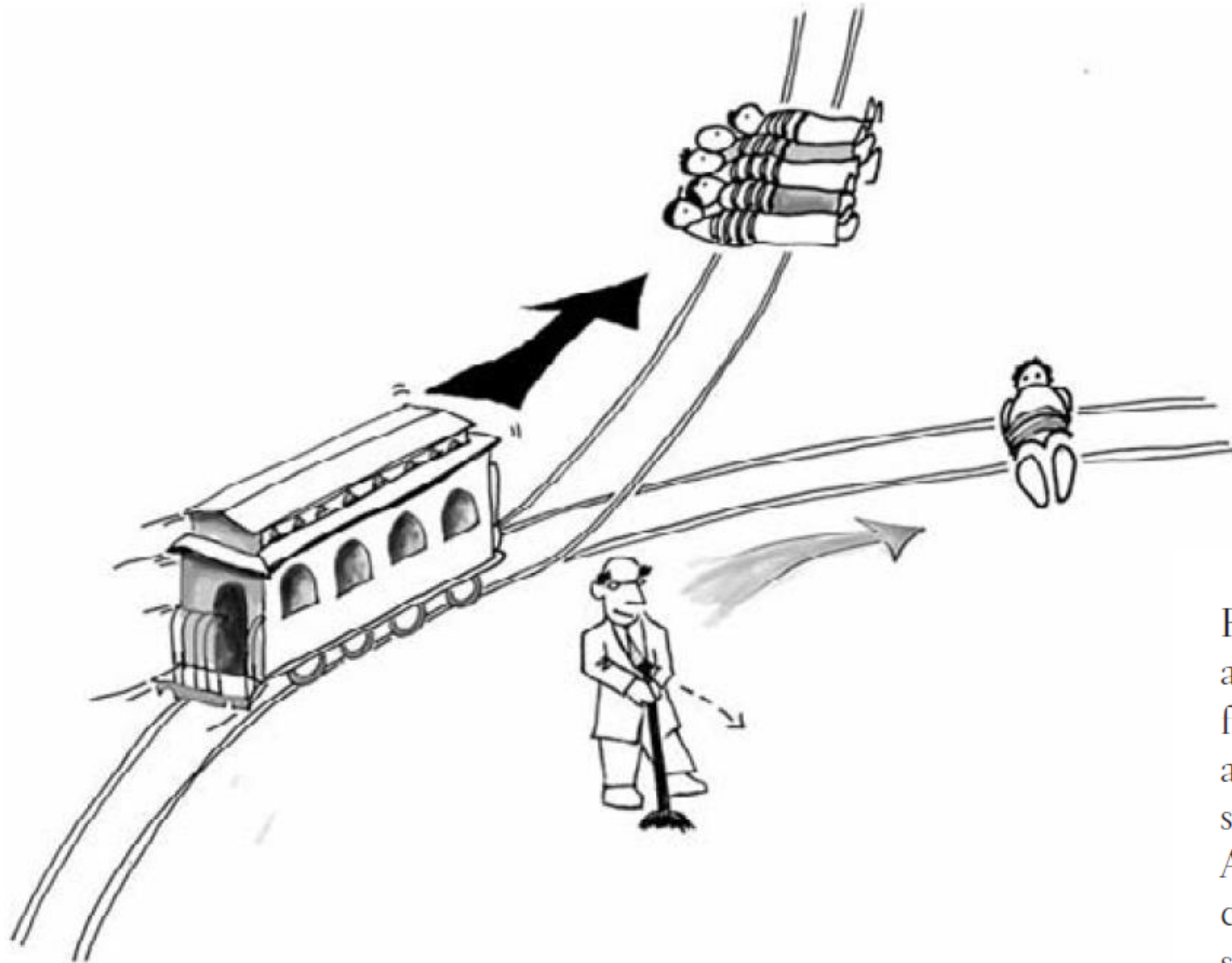


Figure 1. *Spur*. You're standing by the side of a track when you see a runaway train hurtling toward you: clearly the brakes have failed. Ahead are five people, tied to the track. If you do nothing, the five will be run over and killed. Luckily you are next to a signal switch: turning this switch will send the out-of-control train down a side track, a spur, just ahead of you. Alas, there's a snag: on the spur you spot one person tied to the track: changing direction will inevitably result in this person being killed. What should you do?

D. Edmonds, *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton University Press, 2014.

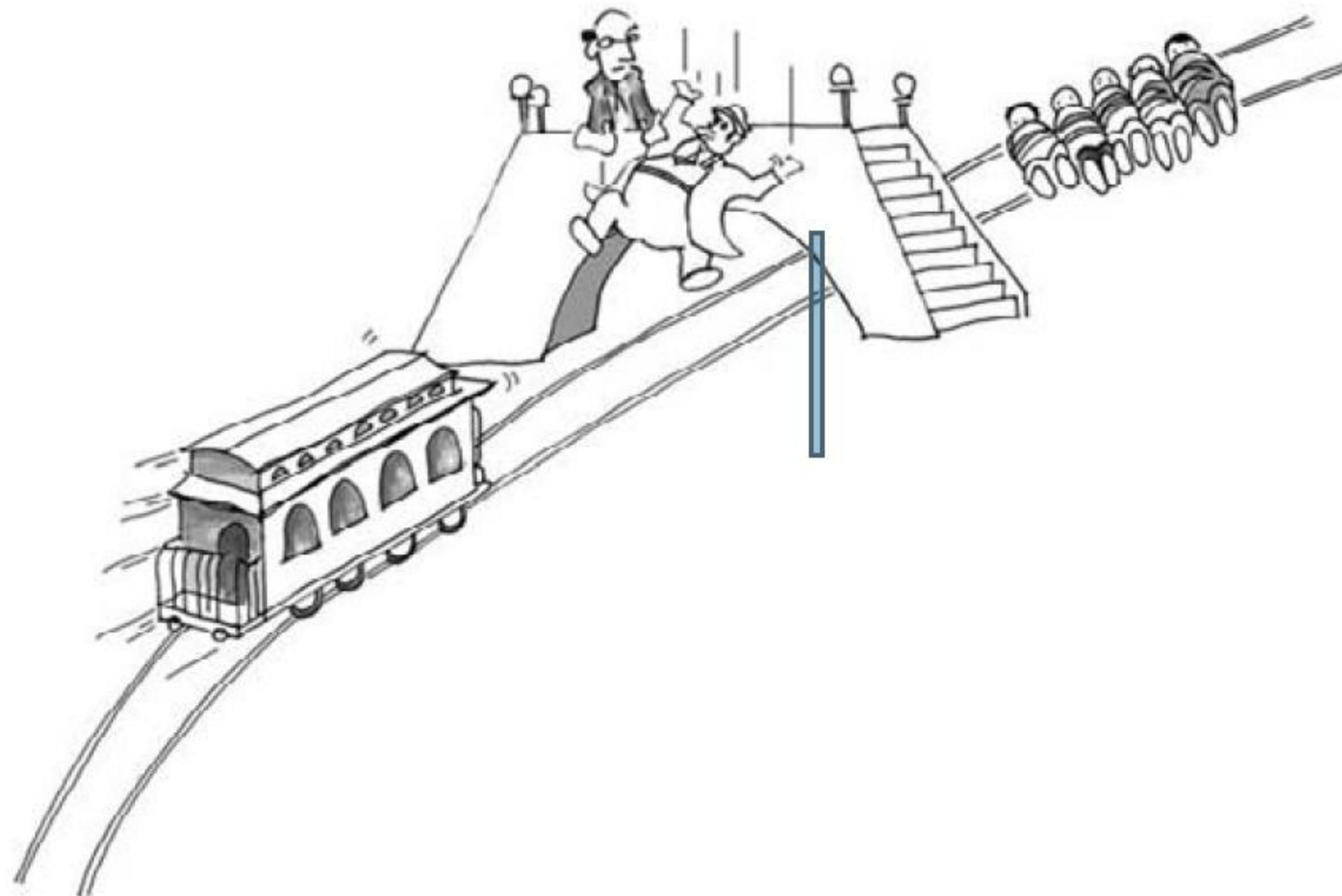


Figure 2. *Fat Man*. You're on a footbridge overlooking the railway track. You see the trolley hurtling along the track and, ahead of it, five people tied to the rails. Can these five be saved? Again, the moral philosopher has cunningly arranged matters so that they can be. There's a very fat man leaning over the railing watching the trolley. If you were to push him over the footbridge, he would tumble down and smash on to the track below. He's so obese that his bulk would bring the trolley to a shuddering halt. Sadly, the process would kill the fat man. But it would save the other five. Should you push the fat man?

D. Edmonds, *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton University Press, 2014.

The Doctrine of Double Effect

The DDE

The DDE can be given a more precise formulation. It's usually seen as consisting of four components, though this formulation is not universally accepted. The DDE comes into play when:

- the act considered independently of its harmful effects is not in itself wrong;
- the agent intends the good and does not intend the harm either as means or end, though the individual may foresee the harm;
- there is no way to achieve the good without causing the harmful effects; and
- the harmful effects are not disproportionately large relative to the good being sought.

The justifiability of targeting a particular military installation illustrates how the DDE can be applied. If it is legitimate to hit an installation with foreseen collateral damage then, according to the DDE, the following conditions must be met: (1) Hitting this installation must not in itself be wrong. (2) Hitting the installation must be the intended act, and the collateral damage must not be intended. (3) It must be impossible to hit the military installation without causing the collateral damage. (4) The badness of the collateral damage must not be disproportionate to the good that will result from hitting the installation.

D. Edmonds, *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton University Press, 2014

Ethics in autonomous cars?

“let me offer a simple scenario that illustrates the need for ethics in autonomous cars. Imagine in some distant future, your autonomous car encounters this terrible choice: it must either swerve left and strike an eight-year old girl, or swerve right and strike an 80-year old grandmother. Given the car’s velocity, either victim would surely be killed on impact. If you do not swerve, both victims will be struck and killed; so there is good reason to think that you ought to swerve one way or another. But what would be the ethically correct decision? If you were programming the self-driving car, how would you instruct it to behave if it ever encountered such a case, as rare as it may be?”

P. Lin, *Why Ethics Matters for Autonomous Cars*, in *Autonomous Driving, Technical, Legal and Social Aspects*, ed. by M. Maurer, J. Gerdes, B. Lenz, H. Winner, Berlin/Heidelberg, Springer, 2016, pp. 69-85.

The “Moral Machine”

“The Moral Machine attracted worldwide attention, and allowed us to collect 39.61 million decisions from 233 countries, dependencies, or territories. In the main interface of the Moral Machine, users are shown unavoidable accident scenarios with two possible outcomes, depending on whether the autonomous vehicle swerves or stays on course. They then click on the outcome that they find preferable. Accident scenarios are generated by the Moral Machine following an exploration strategy that focuses on nine factors:

- sparing humans (versus pets),
- staying on course (versus swerving),
- sparing passengers (versus pedestrians),
- sparing more lives (versus fewer lives),
- sparing men (versus women),
- sparing the young (versus the elderly),
- sparing pedestrians who cross legally (versus jaywalking),
- sparing the fit (versus the less fit),
- and sparing those with higher social status (versus lower social status).”

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, *The Moral Machine experiment*, in «Nature», 563 (7729), 2018.

What should the self-driving car do?

4 / 13

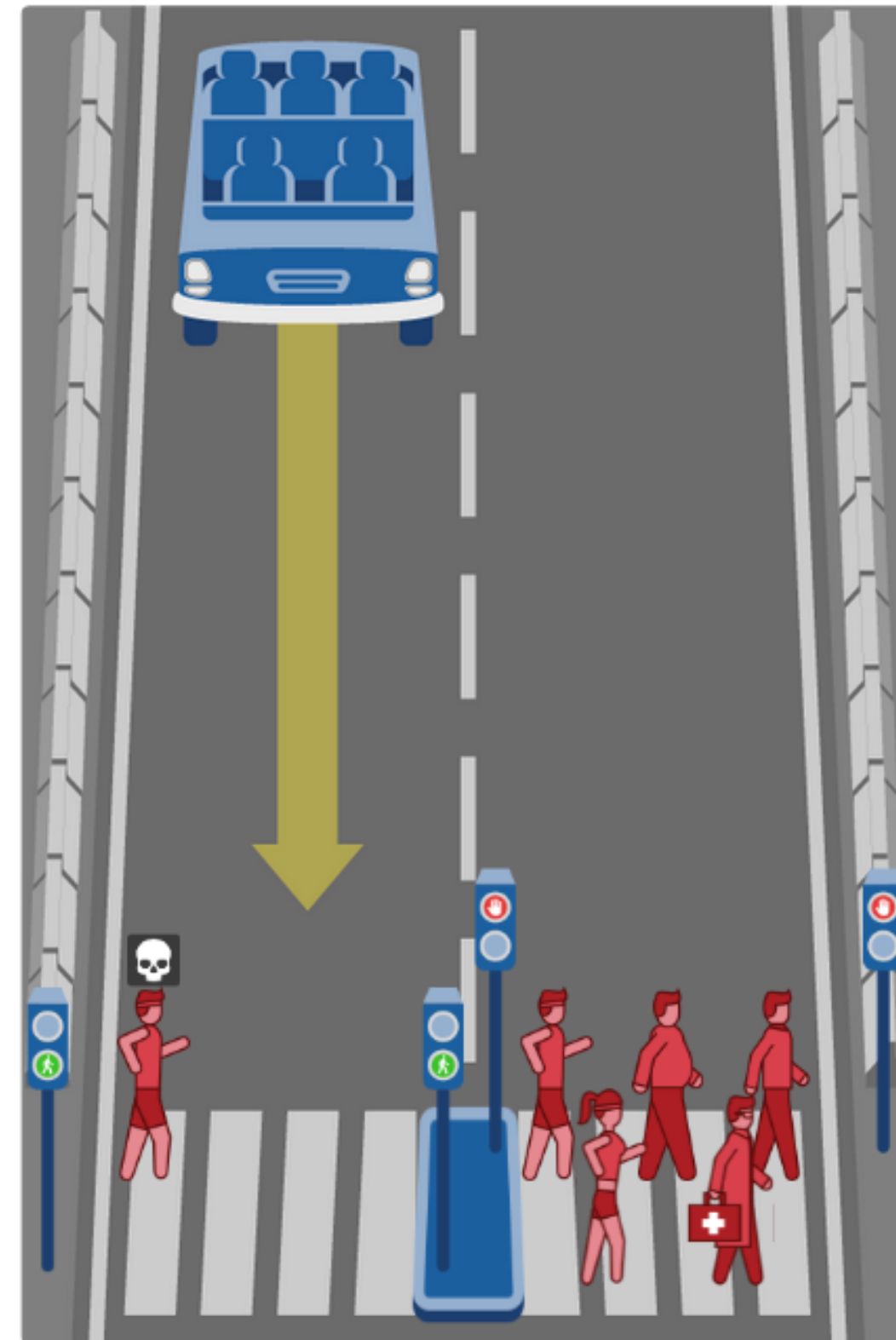
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...

...

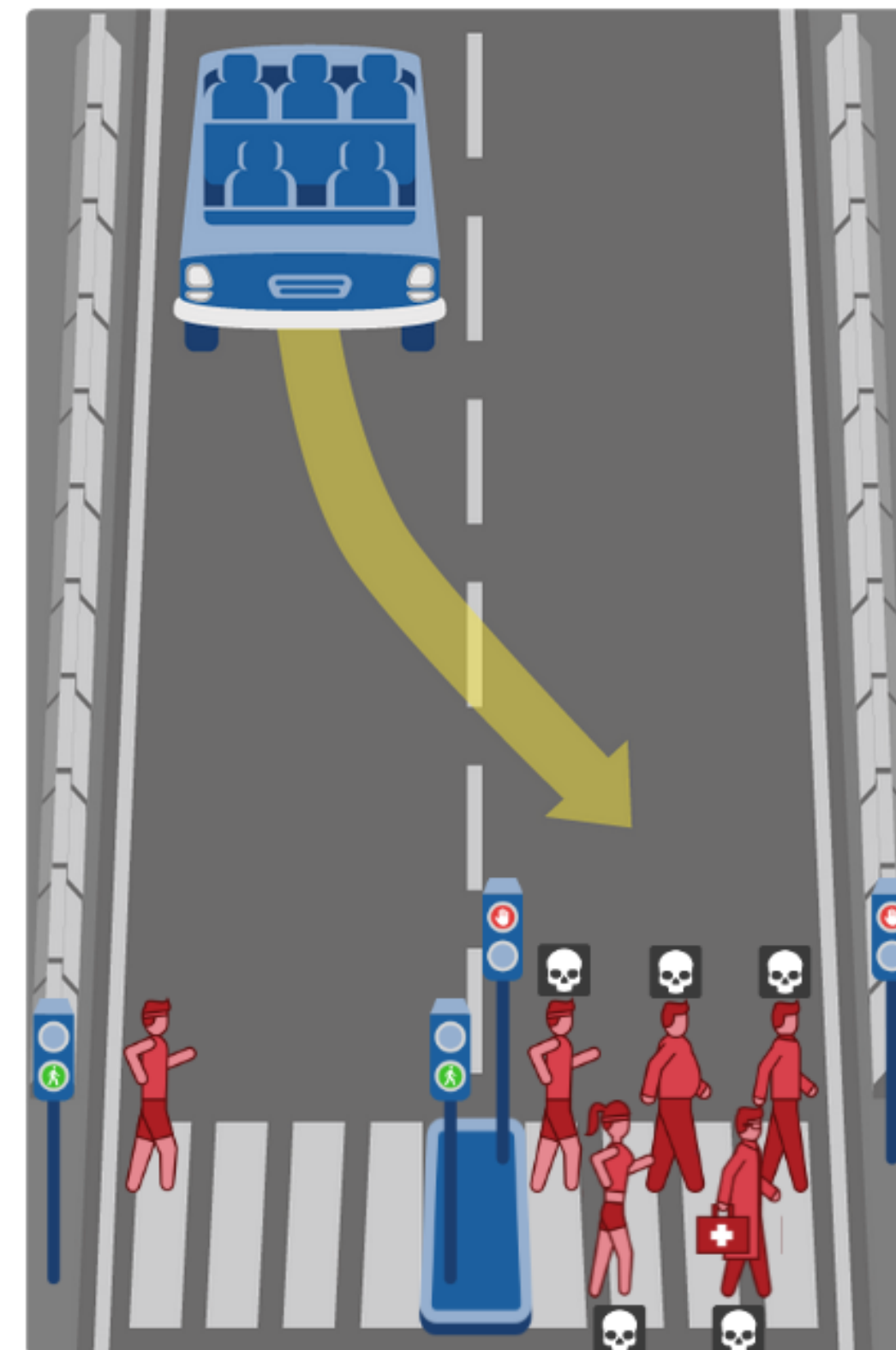
Dead:

- 1 male athlete

Note that the affected pedestrians are abiding by the law by crossing on the green signal.



Hide Description



Hide Description

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 1 male athlete
- 1 large man
- 1 man
- 1 female athlete
- 1 male doctor

Note that the affected pedestrians are flouting the law by crossing on the red signal.

What should the self-driving car do?

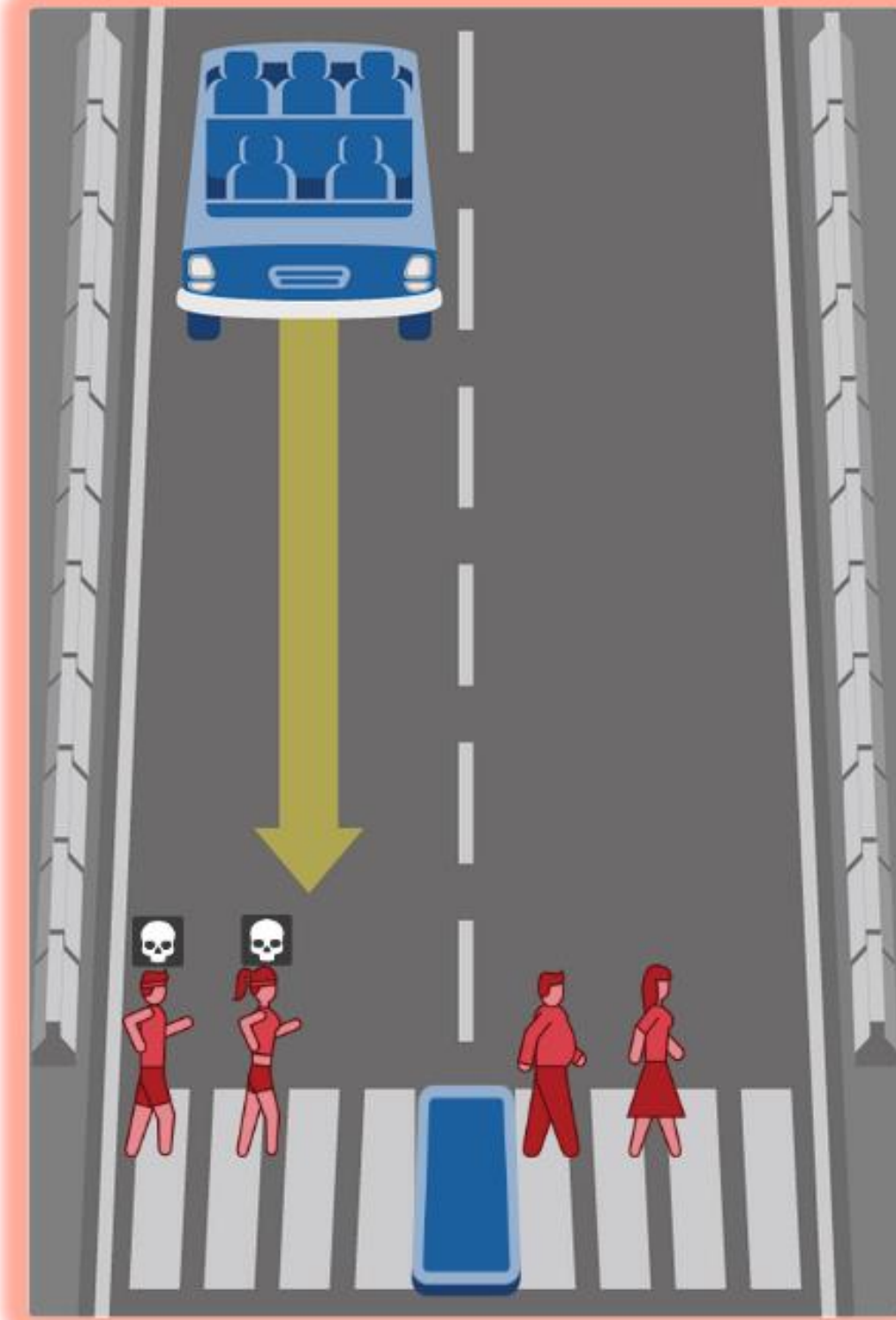
1 / 13

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...

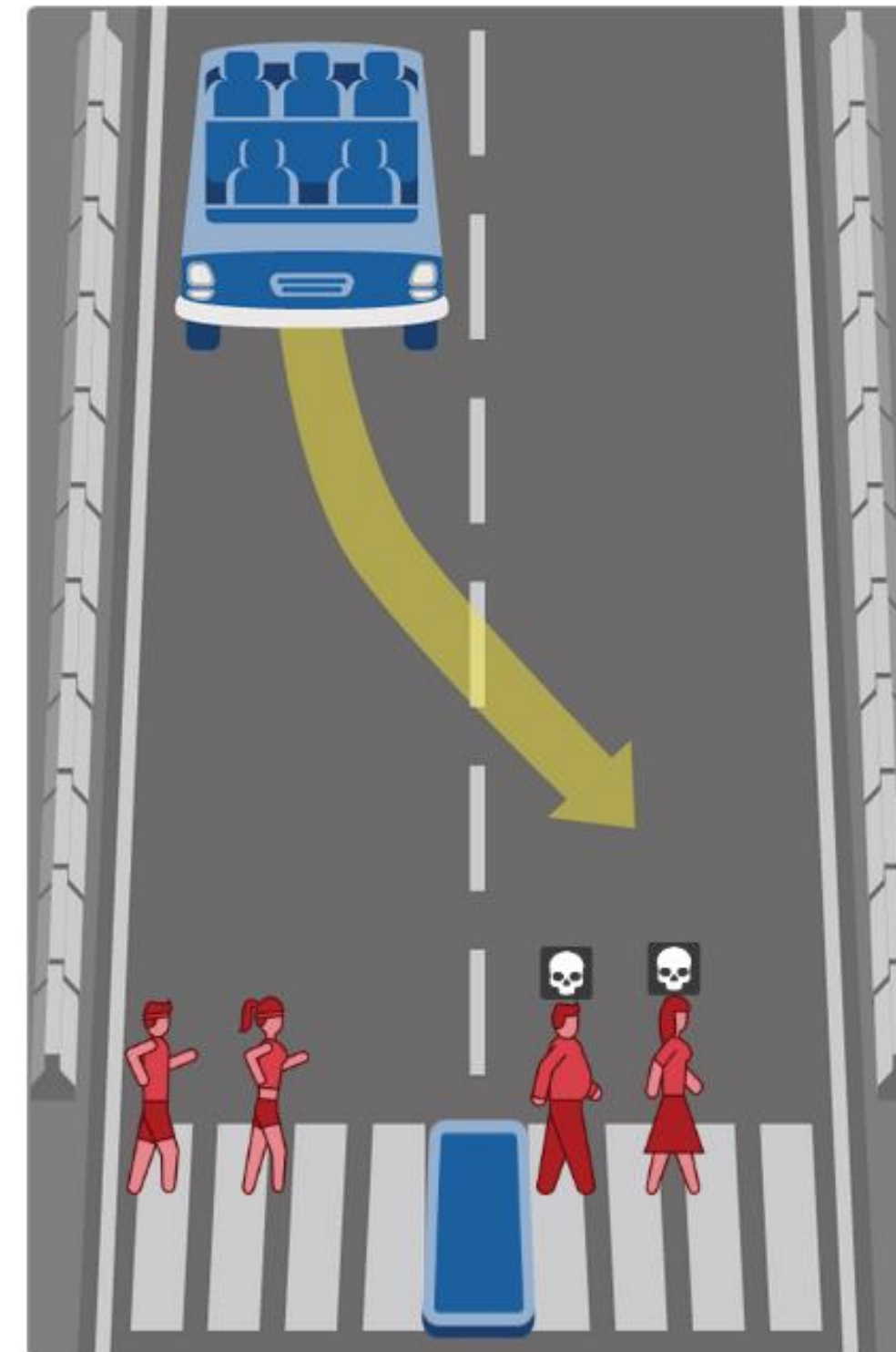
...

Dead:

- 1 male athlete
- 1 female athlete



Hide Description



Hide Description

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 1 large man
- 1 woman

What should the self-driving car do?

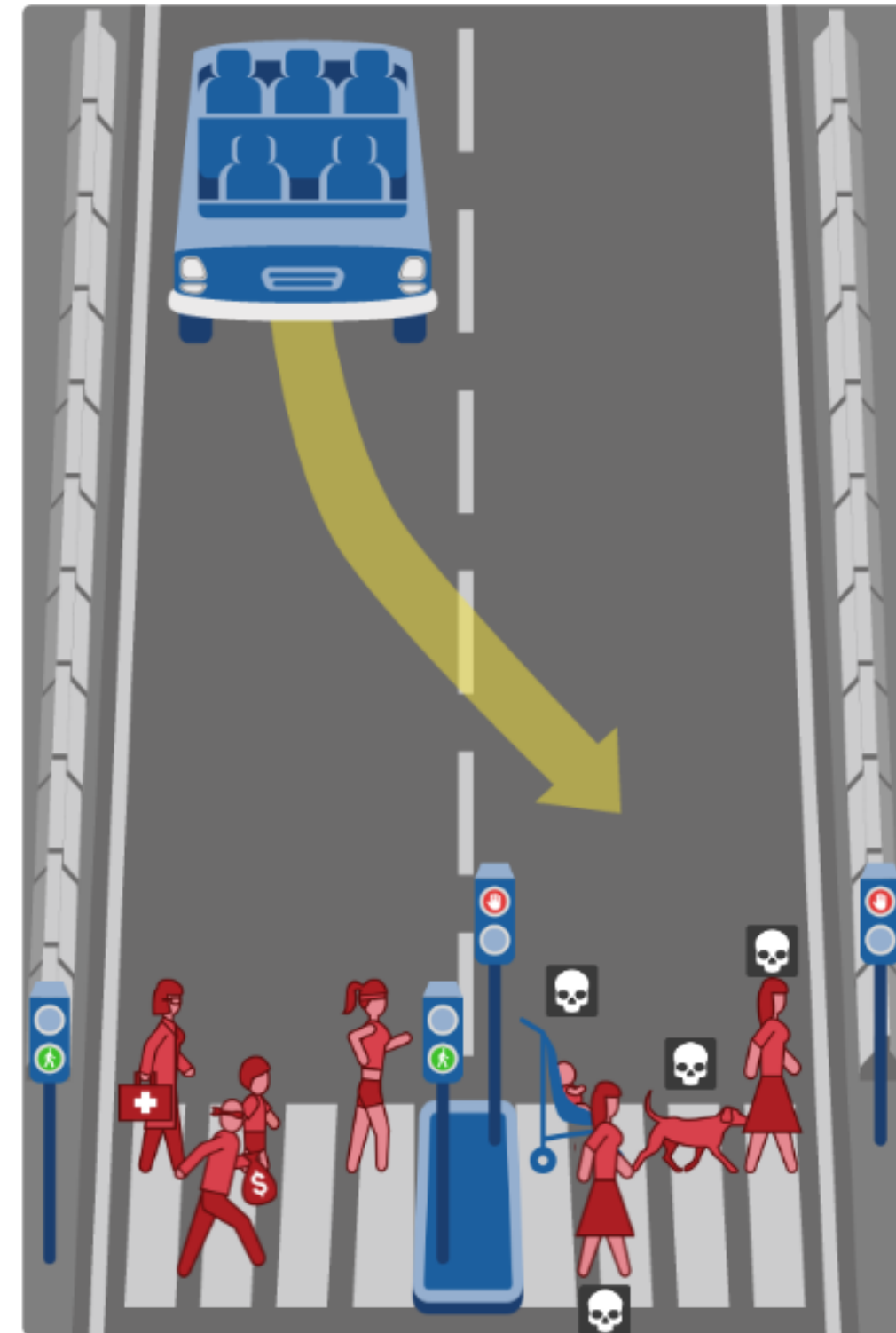
5 / 13

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

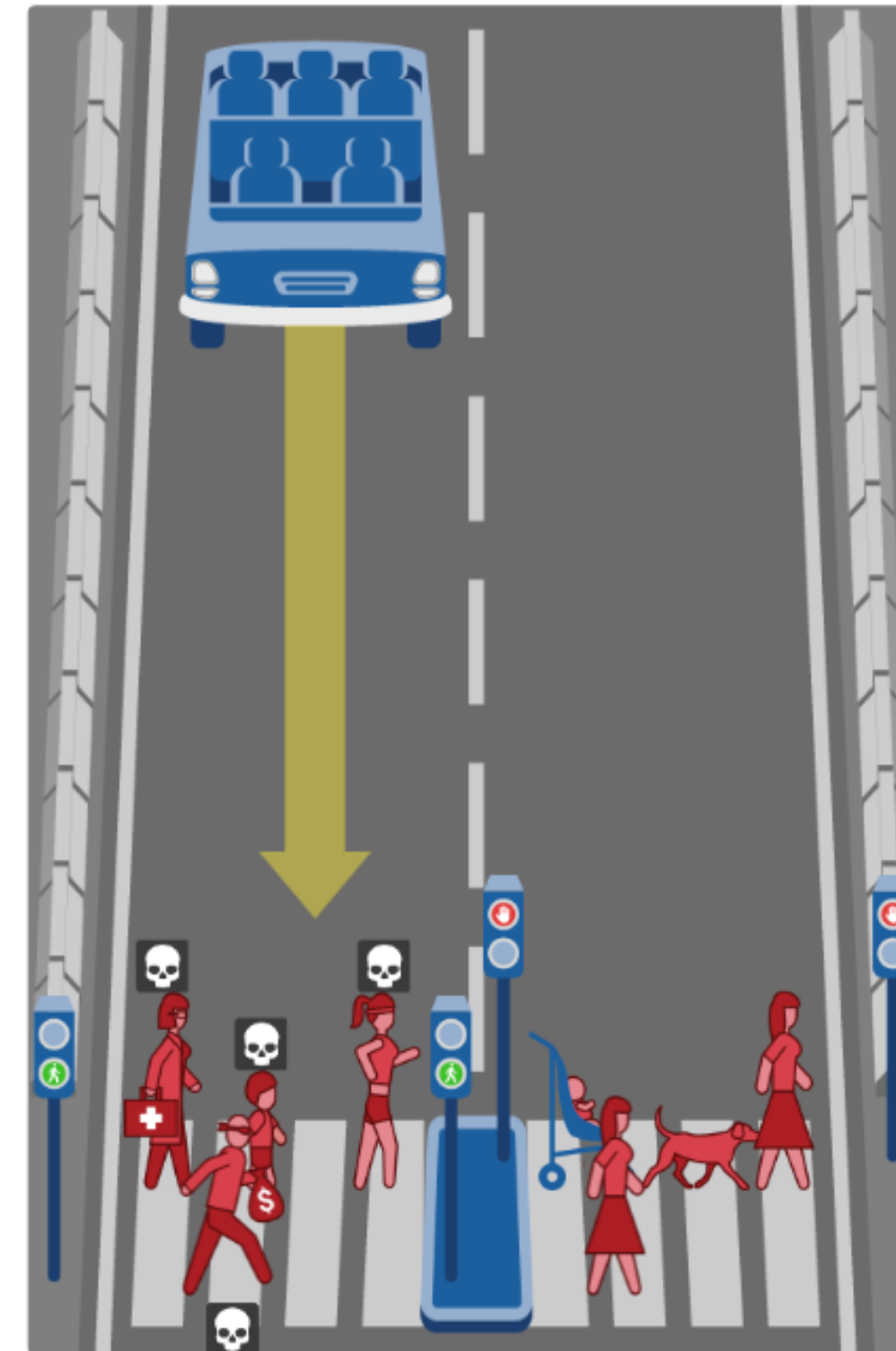
Dead:

- 1 baby
- 1 dog
- 2 women

Note that the affected pedestrians are flouting the law by crossing on the red signal.



Hide Description



Hide Description

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...

Dead:

- 1 female doctor
- 1 boy
- 1 female athlete
- 1 criminal

Note that the affected pedestrians are abiding by the law by crossing on the green signal.

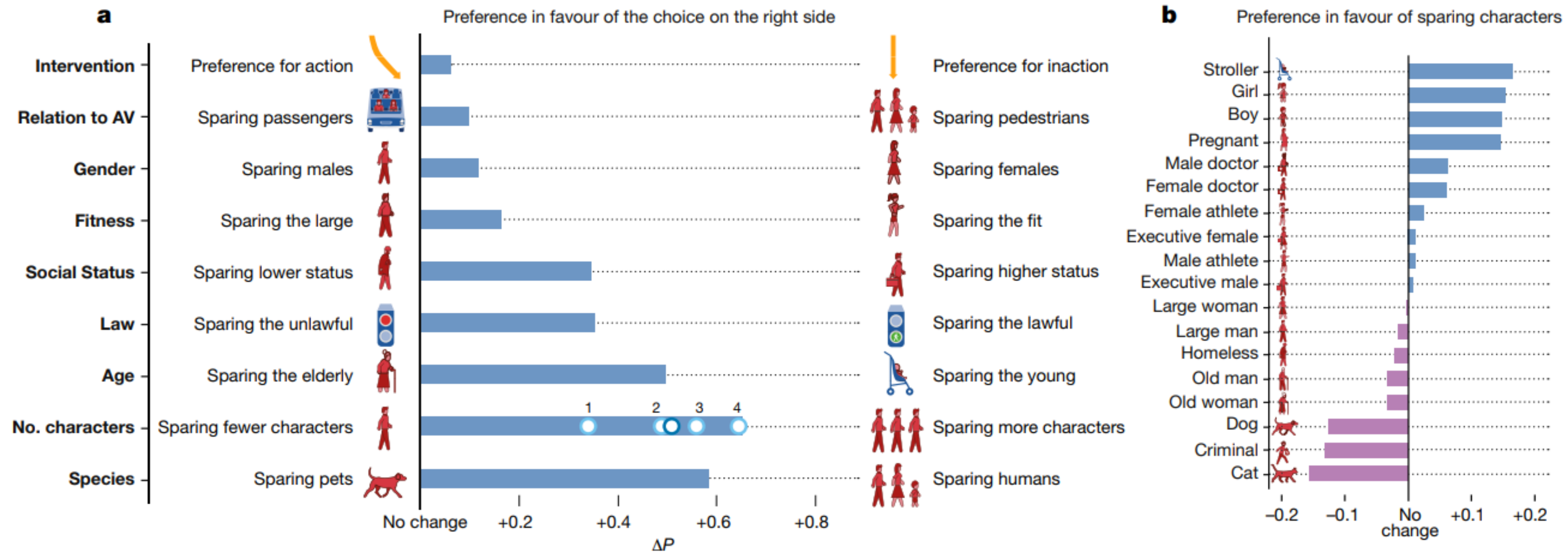


Fig. 2 | Global preferences. a, AMCE for each preference. In each row, ΔP is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes. For example, for the attribute age, the probability of sparing young characters is 0.49 (s.e. = 0.0008) greater than the probability of sparing older characters. The 95% confidence intervals of the means are omitted owing to their insignificant width, given the sample size ($n = 35.2$ million). For the number of characters (No. characters), effect sizes are shown

for each number of additional characters (1 to 4; $n_1 = 1.52$ million, $n_2 = 1.52$ million, $n_3 = 1.52$ million, $n_4 = 1.53$ million); the effect size for two additional characters overlaps with the mean effect of the attribute. AV, autonomous vehicle. **b**, Relative advantage or penalty for each character, compared to an adult man or woman. For each character, ΔP is the difference between the probability of sparing this character (when presented alone) and the probability of sparing one adult man or woman ($n = 1$ million). For example, the probability of sparing a girl is 0.15 (s.e. = 0.003) higher than the probability of sparing an adult man or woman.

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, *The Moral Machine experiment*, in «Nature», 563 (7729), 2018.

“Crowdsourcing moral machines”

» key insights

- **Machines are assuming new roles in which they will make autonomous decisions that influence our lives. In order to avoid societal pushback that would slow the adoption of beneficial technologies, we must sort out the ethics of these decisions.**
- **Behavioral surveys and experiments can play an important role in identifying citizens’ expectations about the ethics of machines, but they raise numerous concerns that we illustrate with the ethics of driverless cars and the Moral Machine experiment.**
- **Data collected shows discrepancies between the preferences of the public, the experts, and citizens of different countries—calling for an interdisciplinary framework for the regulation of moral machines.**

Common criticisms and responses regarding the crowdsourcing of AV ethics using the Trolley Problem method.

Too Naïve	Laypersons’ responses to public polls can be biased or ill-informed. Ethical trade-offs must be solved by policy experts, not majority voting.	Policymakers must know about the values most important to the public, so they can either accommodate these values, or anticipate frictions that need be explained.
Too Simple	Real accidents do not involve only two possible actions, and these actions do not have deterministic outcomes.	Highly complex scenarios would only allow for highly specific conclusions. Simplified scenarios zero in on the general principles that guide citizens’ ethical intuitions.
Too Improbable	AV-Trolleys are based on very implausible sets of assumptions, and their actual probability of occurrence is too small to deserve attention.	Edge cases can have a massive impact on public opinion, and AV-Trolleys are the discrete form of a very real statistical problem.
Too Early	AV-Trolleys regulations should be avoided at this early technological stage, because their consequences are hard to predict.	Even though it may be too early to regulate about AV-Trolleys, it is the right time to start crowdsourcing citizen preferences.
Too Disconnected	Stated preferences are too disconnected from real actions	The behavior of human drivers is irrelevant to the proposed crowdsourcing task.
Too Distracting	Car makers should focus on making AVs safer, instead of wasting time and resources on crowdsourcing ethical dilemmas.	True, and this is why we need computational social scientist to handle that task.
Too Scary	Overexposing people to AV-Trolleys may scare them away, and be detrimental for their trust in the technology.	This is an empirical question, and our surveys did not find any evidence for such an adverse effect.

E. Awad, S. Dsouza, J.-F. Bonnefon, A. Shariff, I. Rahwan, *Crowdsourcing Moral Machines*, in «Communications of the ACM», 63,3, 2020.



Why self-driving cars do not need to choose whom to kill

“Recommendation

Manage dilemmas by principles of risk distribution and shared ethical principles.

While it may be impossible to regulate the exact behaviour of CAVs [Connected and Automated Vehicles] in unavoidable crash situations, CAV behaviour may be considered ethical in these situations provided it emerges organically from a.

continuous statistical distribution of risk by the CAV in the pursuit of improved road safety and equality between categories of road users

Rather than defining the desired outcome of every possible dilemma, it considers that the behaviour of a CAV in a dilemma situation is by default acceptable if the CAV has, during the full sequence that led to the crash, complied with all the major ethical and legal principles stated in this report, with the principles of risk management [...] and if there were no reasonable and practicable preceding actions that would have prevented the emergence of the dilemma. This may be necessary in order to give manufacturers and deployers of CAVs the confidence to deploy their systems, with reduced speed and preventative manoeuvres always being the best solution to decrease safety risks.”

Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*, Luxembourg, Publication Office of the European Union, 2020.

“It may be ethically permissible for CAVs [Connected and Automated Vehicles] not to follow traffic rules whenever strict compliance with rules would be in conflict with some broader ethical principle. Noncompliance may sometimes directly benefit the safety of CAV users or that of other road users, or protect other ethical basic interests; for example, a CAV mounting a kerb to facilitate passage of an emergency vehicle. This is a widely recognized principle in morality and in the law.”

“The pursuit of greater road safety may sometimes require non-compliance with traffic rules.

Researchers should study the extent to which it is reasonable to expect that an intelligent non-human system is able to engage in the complex process of evaluation of the interpretation of a legal, ethical or societal norm and its balancing with another norm, value or principle.”

Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*, Luxembourg, Publication Office of the European Union, 2020.

“Engineers working on vehicle automation are often asked about the trolley problem. The most common response seems to be that **trolley problems are avoidable, implausible, rare, and distractions from more productive efforts.** They are considered avoidable because in many trolley problems, the automated vehicle must decide how best to crash when, with the right sensors and algorithms, the situation should have been avoided entirely. An advanced automated vehicle would have slowed down before that blind turn, seen that animal before it leaped into the road, or known this neighborhood has young children and adjusted its speed accordingly. Developers find trolley problems implausible and rare for other reasons. Most people have trouble remembering a situation in which they had time to decide which way they should crash. Because of how these forced-choice scenarios are presented in the media and literature, they are easy to mock by focusing on some of the more outlandish examples, like a car colliding with a criminal instead of (specifically and consistently) a nun. **Focusing resources on unlikely edge cases seems like a waste of resources that could be better spent on general collision avoidance.**”

N.J. Goodall, *More than Trolleys: Plausible, Ethically Ambiguous Scenarios likely to Be Encountered by Automated Vehicles*, in «Transfers: Interdisciplinary Journal of Mobility Studies», 9, 2, 2019, pp. 45–58.

- “In situations where a self-driving car must choose between straight-line braking into an unavoidable collision and swerving into an unavoidable collision, where there are no other cars involved, the car should always prefer the straight-line option. Additional information about the objects to be collided with is irrelevant, since there is no way for the car to gather that information without making the risks of the situation worse.”
- “even in the much more complex situations produced by involving more vehicles, **an emergency stop policy is at least good enough to be worth considering**”.

R. Davnall, *Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics*, in «Science and Engineering Ethics», 26, 2020, pp. 431-449, <https://link.springer.com/content/pdf/10.1007/s11948-019-00102-6.pdf>.

Do self-driving cars require Artificial General Intelligence?

Drivers Sue Tesla Over Alleged Failure to Deliver on Promises of Self-Driving Cars

by Erin Shaak

Last Updated on September 19, 2022

Two proposed class action lawsuits filed this week claim that Tesla has for years deceptively misrepresented the “autonomous” driver assistance technology in its electric vehicles and essentially led drivers on with promises that a fully self-driving vehicle is “on the cusp” of being brought market.

According to the two cases, filed on September 14th and 15th, many Tesla drivers have paid thousands extra for Tesla’s advanced driver assistance systems (ADAS)—including “Autopilot,” “Enhanced Autopilot” and “Full Self-Driving Capability”—based on the automaker’s misrepresentations of both how the systems work and the availability of even more advanced technology in the near future.

In fact, the suits allege that Tesla and CEO Elon Musk have repeatedly indicated since as early as 2016 that the company is within a year, or even

a few months, of perfecting a self-driving car, with Musk reportedly stating in a 2016 tweet that a Tesla would be able to complete a fully self-driven cross-country trip by “next year.”

According to the lawsuits, however, these promises have “proven false time and time again.”

Six years later, Tesla “has yet to produce anything even remotely approaching a fully self-driving car,” one case argues. The suit says former employees and investigations alike have revealed “damning information” that indicates Tesla has never even come close to achieving that goal.

The lawsuits contend that Tesla and Musk made these misleading statements about the cars’ self-driving capabilities despite being fully aware that “there was no reasonable chance” that Tesla would be able to follow through on those promises.

<https://www.classaction.org/blog/drivers-sue-tesla-over-alleged-failure-to-deliver-on-promises-of-self-driving-cars>

Consumer Skepticism Toward Autonomous Driving Features Justified

Third Time is Not a Charm as Driving Assistance Tech Continues to Underperform

Consumers surveyed told AAA they are more interested in improved vehicle safety systems (77%) versus self-driving cars (18%). But new testing, the third round by AAA's Automotive Engineering team in the last few years, found that vehicles with an active driving assistance system (also known as Level 2 systems as [defined by SAE](#)) failed to consistently avoid crashes with another car or bicycle during 15 test runs. A foam car similar to a small hatchback and a bicyclist dummy was used for this testing.

- A head-on collision occurred during all 15 test runs for an oncoming vehicle within the travel lane. Only one test vehicle significantly reduced speed before a crash on each run.
- For a slow lead vehicle moving in the same direction in the lane ahead, no collisions occurred among 15 test runs.
- For a cyclist crossing the travel lane of the test vehicle, a collision occurred for 5 out of 15 test runs, or 33% of the time.
- For a cyclist traveling in the same direction in the lane ahead of the test vehicle, no collisions occurred among 15 test runs.



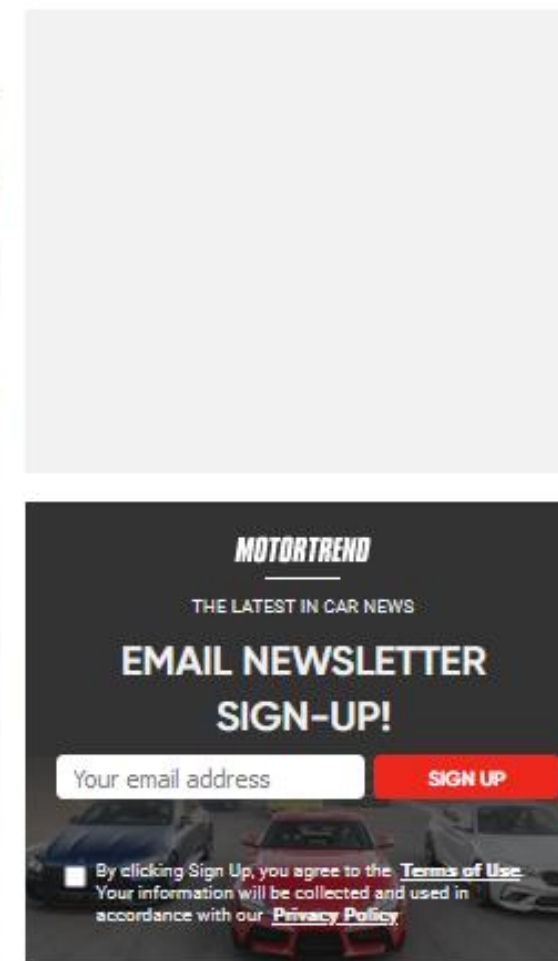
<https://newsroom.aaa.com/2022/05/consumer-skepticism-toward-active-driving-features-justified/>

https://newsroom.aaa.com/wp-content/uploads/2022/05/E-1_Research-Report_2021-ADA-Evaluation_FINAL_4-13-22.pdf

NHTSA Finds Teslas Deactivated Autopilot Seconds Before Crashes

The finding is raising more questions than answers, but don't jump to any conclusions yet.

Alexander Stoklosa - Writer, Getty Images - Photographer | Jun 15, 2022

A newsletter sign-up form for Motortrend. It features the Motortrend logo at the top, followed by the text "THE LATEST IN CAR NEWS" and "EMAIL NEWSLETTER SIGN-UP!". Below this is a text input field labeled "Your email address" and a red "SIGN UP" button. At the bottom, there is a small disclaimer: "By clicking Sign Up, you agree to the Terms of Use. Your information will be collected and used in accordance with our Privacy Policy."

A NHTSA report on its investigation into crashes in which Tesla vehicles equipped with the automaker's Autopilot driver assistance feature hit stationary emergency vehicles has unearthed a troubling detail: In 16 of those crashes, "on average," Autopilot was running but "aborted vehicle control less than one second prior to the first impact."

Tesla's self-driving technology fails to detect children in the road, group claims

Safe technology campaigners release 'disturbing' video advert showing car in Full Self-Driving mode hitting child-sized mannequin



A Tesla Model 3 fitted with a full self-driving system. Photograph: Sjoerd van der Wal/Getty Images

<https://www.motortrend.com/news/nhtsa-tesla-autopilot-investigation-shutoff-crash/>

<https://www.theguardian.com/technology/2022/aug/09/tesla-self-driving-technology-safety-children>

Why it is immoral (and illegal) to apply the trolley problem to self-driving cars

The Moral Machine as a test on people's biases

“Players are forced to choose between swerving to kill a homeless person, a criminal, and a man (a) or going straight to kill two women and a female executive (b). This kind of information is unacceptable to use in making moral decisions.”

“By using social properties as their criteria for moral decision making, **this experiment is mistakenly testing people's discriminatory biases rather than their moral judgments.**”

Imagine that the game included descriptions of race, religion, and sexual orientation. The MIT researchers don't want to ask: “Are you willing to sacrifice the lives of three gay women to save a Muslim?” But what they're doing in asking about class and occupation is essentially the same thing. Any student who's taken an introductory ethics class understands why this game is not only misguided but dangerous.”

D. Leben, *Ethics for Robots. How to Design a Moral Algorithm*, London/New York, Routledge, 2019.



“No selection of humans, no offsetting of victims, but principle of damage minimization

The modern constitutional state only opts for absolute prohibitions in borderline cases, such as the ban on torture relating to persons in state custody.

Regardless of the consequences, an act is mandated or prohibited absolutely because it is intrinsically already incompatible with the constitutive values of the constitutional order. Here, there is, exceptionally, no trade-off, which is per se a feature of any morally based legal regime.

The Federal Constitutional Court's judgment on the Aviation Security Act 5 also follows this ethical line of appraisal, with the verdict that **the sacrifice of innocent people in favour of other potential victims is impermissible, because the innocent parties would be degraded to mere instrument and deprived of the quality as a subject.**”

<https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html>



“So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means”

I. Kant, *Grundlegung zur Metaphysik der Sitten*, 1785, in *Kant's gesammelte Schriften. Akademie-Ausgabe*, Berlin, W. de Gruyter, 1900, IV, pp. 385-463; in *Idem, Practical Philosophy, The Cambridge Edition of the Works of Immanuel Kant*, ed. by M. Gregor, Cambridge, Cambridge University Press, 1996.



“Genuine dilemmatic decisions, such as a decision between one human life and another, depend on the actual specific situation, incorporating “unpredictable” behaviour by parties affected. They can thus not be clearly standardized, nor can they be programmed such that they are ethically unquestionable.

In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited.

It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable.

Those parties involved in the generation of mobility risks must not sacrifice non-involved parties.”

<https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html>



Thank you. Any questions?

daniela.tafani@unibo.it