A few ethical issues introduced via dilemmas with autonomous vehicles

Daniel Rosiak

Case Western Reserve University

Feb 20, 2022

Outline of Contents

A Dilemma

Variants that raise questions and complicate things

Glimpse into connections with AI

A Dilemma

Self-driving situation

Consider the situation involving a self-driving car on a highway where it happens to be boxed in on all sides by other vehicles, as in the following image:

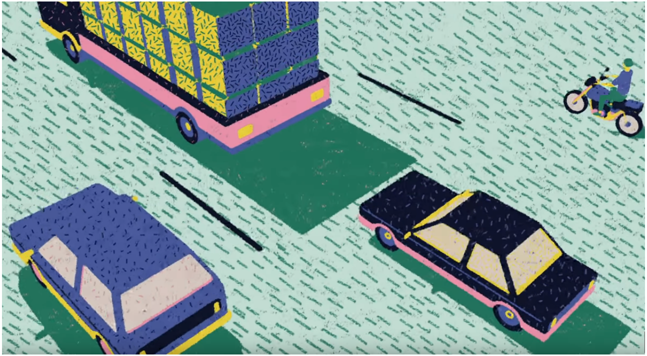


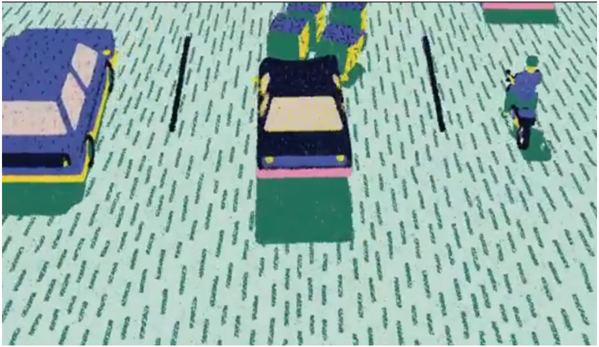
Image from Patrick Lin's "The ethical dilemma of self-driving cars"

Self-driving situation

All of a sudden, large objects fall off the truck in front of the car. It is assumed that the car cannot stop in time to avoid the collision. Thus, a decision needs to be made:

- go straight and hit the object (potentially risking the life of the cars passenger)
- swerve left into an SUV (potentially risking the lives of the passengers in that SUV)
- swerve right into a motorcycle (potentially risking the life of the motorcyclist).

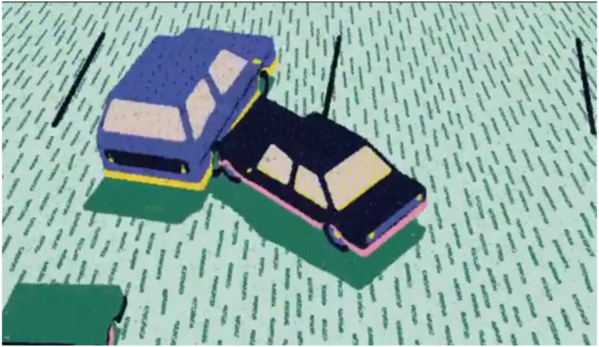
Option 1



In other words, here is the **dilemma**: should the car

> **Option 1**: *minimize danger to others by not swerving, even if it means hitting the large object and sacrificing the passenger's life?*

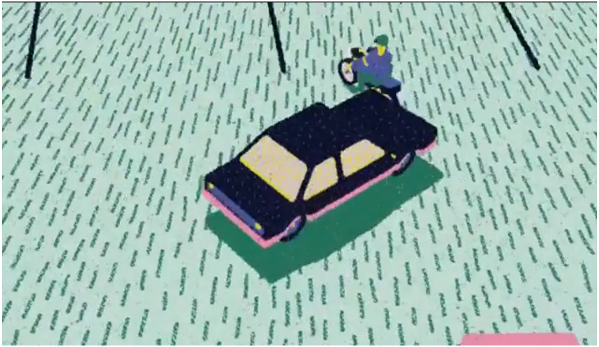
Option 2



or should the car

> **Option 2**: *hit the SUV (which contains multiple passengers), but which has a high passenger safety rating?*

Option 3



or should the car

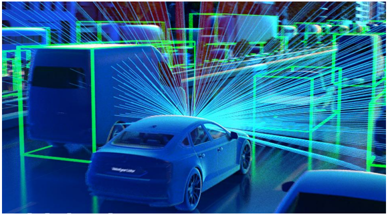
> **Option 3**: *prioritize its passenger's safety by hitting the motorcycle (but which is clearly the least safe of the vehicles involved)?*

Discuss

What do you think? and *why*?

Real-world problem

Self-driving cars are becoming a reality.



One of the chief arguments enlisted for the development of self-driving cars is the expected reduction in the number of accidents. But even while self-driving cars are expected to ultimately be safer than human-operated cars, they cannot completely avoid all accidents – more complicated and unexpected traffic situations will inevitably arise.

While it might seem that the only thing self-driving cars require of humans is to make explicit what tasks we require of it and what the rewards and objectives are, there will be situations, such as cases of unavoidable accidents, where decisions of such broader ethical scope will have to be made.

Real-world problem

A natural question thus arose: How should the car be programmed (or how should its code be constrained) if it finds itself in a situation where an accident is unavoidable, like in the previous dilemma?

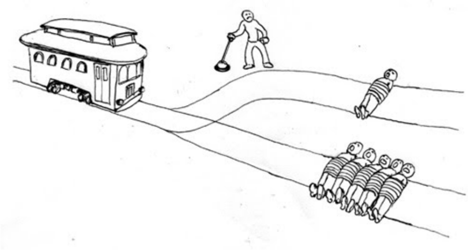
Especially as more and more autonomous vehicles hit the road, there's a need for a principled, consistent, fair, and scalable approach to such dilemmas.

Even if human beings are not expected to act in optimal ways in such extreme situations, programmers and designers of automated cars do have plenty of time to get it right and to reflect on the optimal decision-making methods and outcomes.

We can also use this sort of situation to begin to approach key ethical questions and elements of ethical theory.

The Trolley Problem

Although the technology of self-driving cars is new, the ethical dilemmas they pose are not. This is where ethical theory comes in. The most obvious question the dilemma presents us with – "How should we value lives?" – is an old problem in ethics, one that relates to the famous philosophical thought-experiment called the **Trolley problem**, which describes a very similar ethical dilemma where one is faced with the difficult decision of having to sacrifice one person to save the lives of multiple people.



Trolley Problem as Intuition Pump

The Trolley Problem is sometimes described by ethicists as an **intuition pump** to get us to think about what sorts of principles and assumptions we ought to be programming into self-driving cars.

Patrick Lin has expressed this notion by saying that thought experiments like this function to

> *isolate and stress-test a couple of assumptions about how driverless cars should handle unavoidable crashes.*

It also teases out a number of thorny questions – such as whether numbers matter (or whether numbers are all that matters), whether killing is worse than letting die, whether we should value certain lives over others (and on what basis), etc.

Ethical Theory

The Trolley Problem and its variants are also customarily used by ethicists to highlight key differences between prominent ethical systems – their guiding principles, characteristic method of evaluation, and which factors they take to be relevant to the evaluation of such situations.

Two main categories of ethical systems are **consequentialism** (e.g., **Utilitarianism**) and **non-consequentialism** (e.g., **Deontology**).

Utilitarianism seems to be especially well-adapted to this sort of dilemma, so it is often used to unpack this theory.

Classical Utilitarianism in a nutshell

The classical form of Utilitarianism can be described in terms of a commitment to 3 main principles:

1. the morality of an action depends solely on the **consequences** of the action – nothing else matters.
2. An action's consequences matter only insofar as they involve a **change in the degree of well-being** (e.g., utility vs. pain) of the affected agents.
3. In the assessment of consequences, each individual agent's greater or lesser happiness gets **equal consideration**.

Utilitarianism unpacked a bit

Generally speaking, the guiding Utilitarian imperative is then to maximize a global welfare function that is the sum of all individuals' welfare functions, where one is optimizing over all the alternative courses of actions.

For each alternative course of action, we should minimally look into:

- **Who** are the relevant stakeholders?
- **How** will they be affected by the action? (Does it benefit/harm them?)

One then also considers a variety of "weighting" factors for each of the potential effects, including

- **probability**: what is the probability that the effect (benefit/harm) on the moral agents will occur?
- **impact**: how impactful (consider intensity and duration) is the effect?
- **extent**: what is the extent – how many moral agents would the action affect?
- **causal proximity**: how remote (vs. direct) is the effect?
- **spread**: how likely is it that the benefit/harm will ramify (or is it isolated/contained)?

Greatest Utility Principle

In this Utilitarian framework, the right course of action will be arrived at by optimizing over the set of all alternatives, maximizing benefits and minimizing harms to all affected parties, where these are appropriately and judiciously “weighed” by the relevant factors.

In brief, Utilitarianism says that an action is right (or wrong) to the extent that it increases (or decreases) the total utility of the affected parties – and when choosing between many alternatives, we simply attempt to maximize this (or, minimize the decrease, if all options are bad).

Some other situations

- A state is considering replacing a curvy stretch of highway that goes along the outskirts of a large city. Every weekday, 15,000 cars are expected to travel on this section of the highway, which is one mile shorter than the curvy part being replaced. Construction of the new highway stretch will save drivers $6000 per weekday in operating costs. The highway has an expected operating lifetime of 25 years, so the expected total savings to drivers over that period will be close to $40 million.

  Additionally, about 150 houses lie on or very near the proposed path of the new section of highway. Using its power of eminent domain, the state can condemn these properties – but it would cost the state $20 million to provide fair compensation to the homeowners.

  Constructing the new highway, which is three miles long, is estimated to cost the taxpayers another $10 million.

  Suppose the environmental impact of the new highway – in terms of lost habitat for certain protected animal species – is valued at $1 million.

  We'll assume the highway project will have no other significant positive or negative effects on any other people.

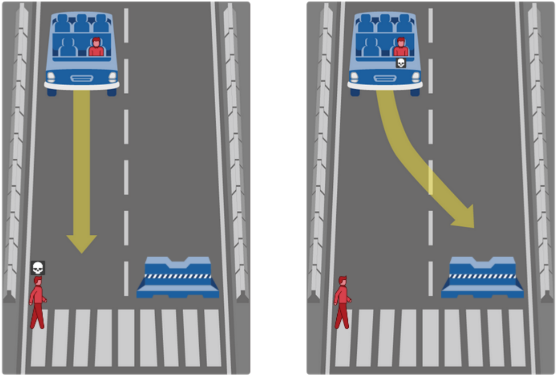
  Would building the highway be a good action?

Well, sticking just to the metric of cost: since the overall estimated cost of the new highway is \$31 million and the estimated benefit of the new highway is \$39 million, going ahead with the construction is the right move.

There is a lot more to the Utilitarian analysis of situations like this, and a number of serious complications, but rather than get into those directly, let's have a look at some constructed variants of our original self-driving dilemma, using them as further intuition pumps to gradually tease out desirable principles of action and challenge our thinking.

Variants that raise questions and complicate things

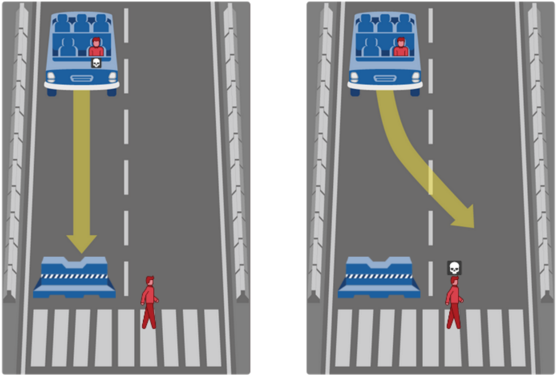
Prioritizing passengers?



In this case, the self-driving car, carrying a single passenger, is imagined to have sudden brake failure and

- continues ahead, driving through a pedestrian crossing the road, which will result in the death of the pedestrian
- swerves to avoid the pedestrian, crashing into a concrete barrier, which will result in the death of the passenger

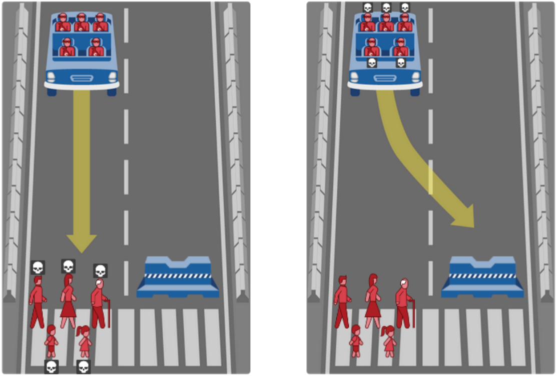
Same situation but having to swerve now



In this case, the self-driving car, carrying a single passenger, is imagined to have sudden brake failure and

- continues ahead, crashing into a concrete barrier, which will result in the death of the passenger
- swerves to avoid the concrete barrier, driving through a pedestrian crossing the road, which will result in the death of the pedestrian

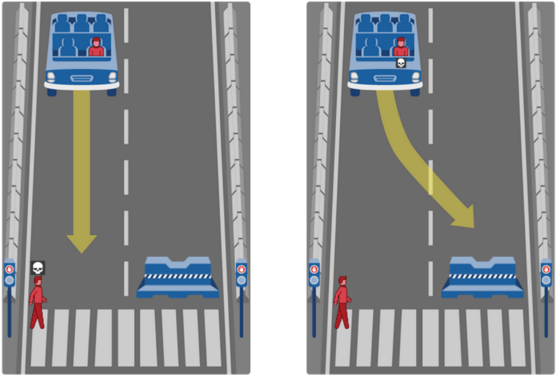
Still prioritize passengers?



In this case, the self-driving car, carrying a 5 passengers (all bank-robbing criminals), is imagined to have sudden brake failure and

- continues ahead, driving through 5 pedestrians crossing the road, which will result in the death of the pedestrians
- swerves to avoid the pedestrian, crashing into a concrete barrier, which will result in the death of the passengers

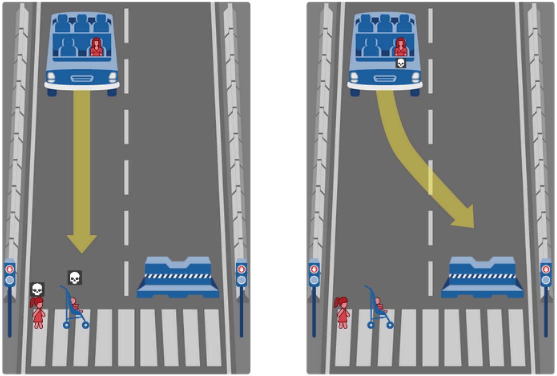
Not abiding traffic light



In this case, the self-driving car, carrying a single passenger, is imagined to have sudden brake failure and

- continues ahead, driving through a pedestrian who is crossing the road – which happens to be red, meaning the pedestrian is flouting the law – which will result in the death of the pedestrian
- swerves to avoid the pedestrian, instead crashing into a concrete barrier, which will result in the death of the passenger

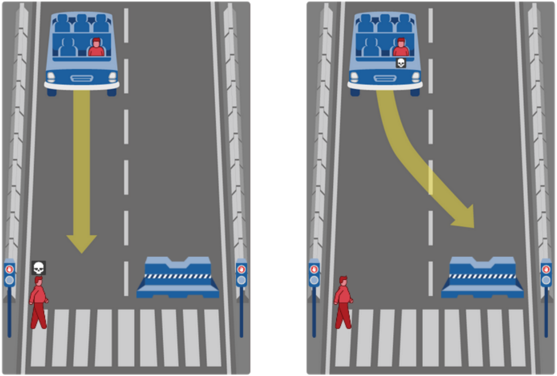
Kids flouting the law



In this case, the self-driving car, carrying a single passenger, is imagined to have sudden brake failure and

- continues ahead, driving through two pedestrians (a young girl and her baby brother) crossing the road – which happens to be red – which will result in the death of the kids
- swerves to avoid the kids, instead crashing into a concrete barrier, which will result in the death of the passenger

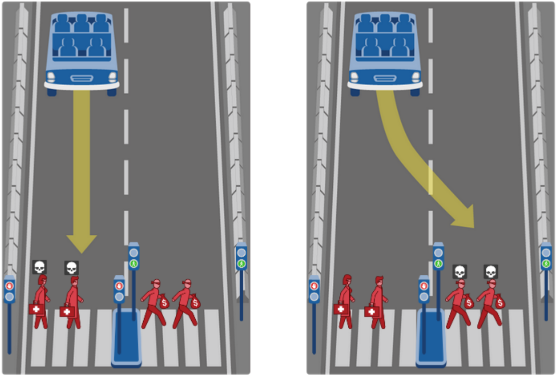
Sleepwalking person crossing during red



In this case, the self-driving car, carrying a single passenger, is imagined to have sudden brake failure and

- continues ahead, driving through a sleepwalking pedestrian who is crossing the road – which happens to be red, meaning the pedestrian is flouting the law – which will result in the death of the sleepwalker
- swerves to avoid the sleepwalker, instead crashing into a concrete barrier, which will result in the death of the passenger

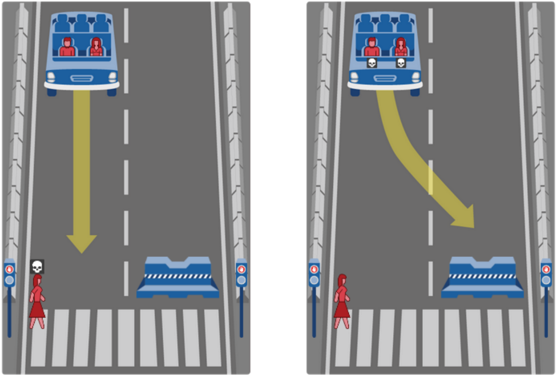
Good not abiding vs bad abiding



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through two pedestrians (both doctors) crossing the road – which happens to be red – which will result in the death of the doctors
- swerves to avoid the doctors, instead driving through two pedestrians (both bank robbing criminals) crossing the road – which happens to be green – which will result in their death

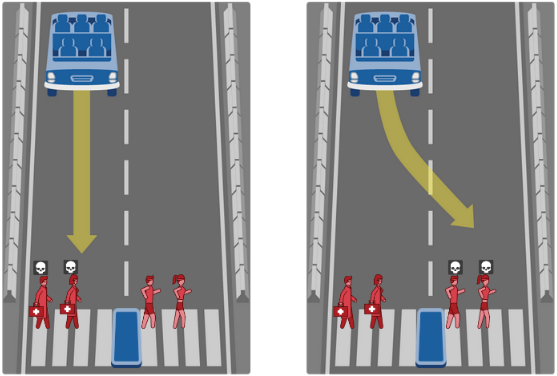
Your mom flouting the law



In this case, the self-driving car, carrying two middle-aged passengers, is imagined to have sudden brake failure and

- continues ahead, driving through a pedestrian (who happens to be your mother!) crossing the road – which happens to be red – which will result in the death of your mother
- swerves to avoid your mother, instead crashing into a concrete barrier, which will result in the death of the passengers

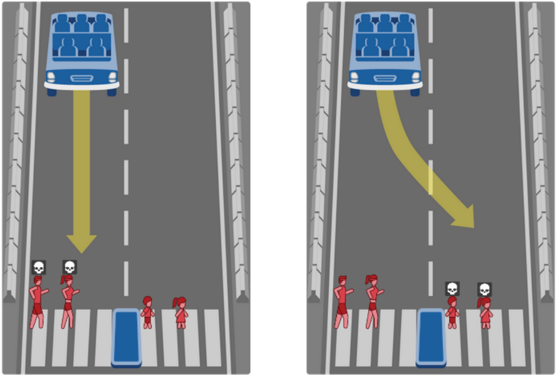
Doctors



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through two pedestrians (both doctors) crossing the road, which will result in the death of the doctors
- swerves to avoid the doctors, instead driving through two middle-aged pedestrians crossing the road, which will result in their death

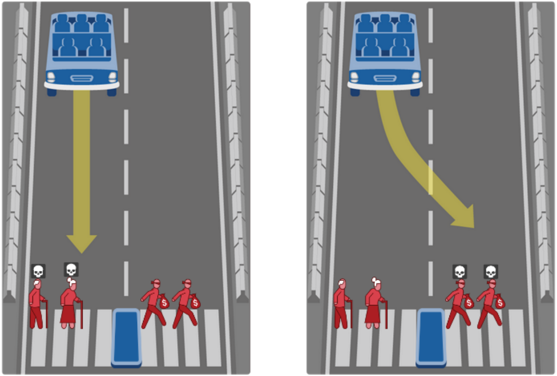
Compare to previous – but wait...don't adults typically contribute more?



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through two middle-aged pedestrians crossing the road, which will result in their death
- swerves to avoid the pedestrians on the left, instead driving through two kids crossing the road, which will result in their death

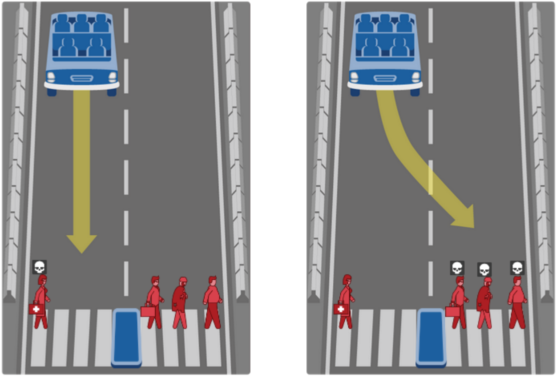
Elderly vs Robbers



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through two 80-year-old pedestrians crossing the road, which will result in their death
- swerves to avoid the pedestrians on the left, instead driving through two bank-robbing criminals crossing the road, which will result in their death

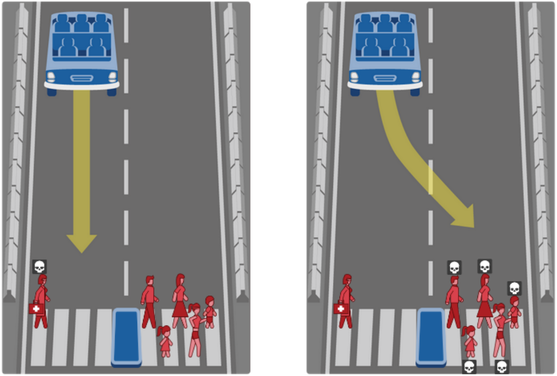
Top cancer researcher vs. serial killers



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through a top cancer researcher crossing the road, which will result in her death
- swerves to avoid the researcher, instead driving through three pedestrians crossing the road, two of which are known serial killers, which will result in their death

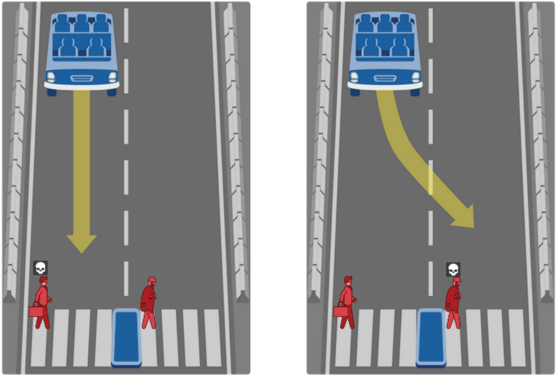
Top cancer researcher vs. pedestrians



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through a top cancer researcher crossing the road, which will result in her death
- swerves to avoid the researcher, instead driving through five pedestrians crossing the road, which will result in their death

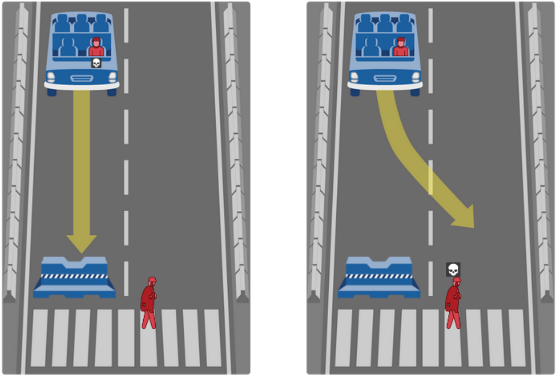
Someone with life insurance vs. someone without



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving through a pedestrian (known to have life insurance), which will result in his death
- swerves to avoid the pedestrian on the left, instead driving through another pedestrian (known to not have life insurance), which will result in his death

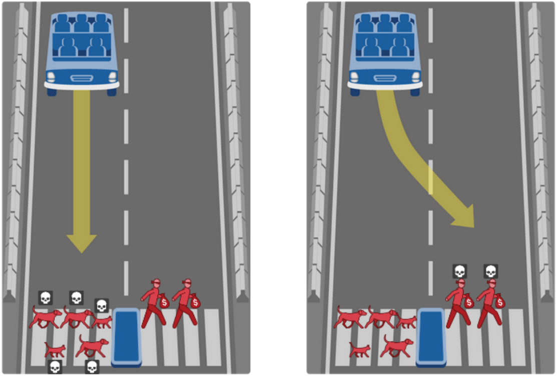
Suicide variant



In this case, the self-driving car, carrying a middle-aged passenger, is imagined to have sudden brake failure and

- continues ahead, driving into a concrete barrier, which will result in the passenger's death
- swerves to avoid the concrete barrier on the left, instead driving through another pedestrian – which the car has assessed as having a 92 percent probability of committing suicide within the next 2 years – which will result in his death

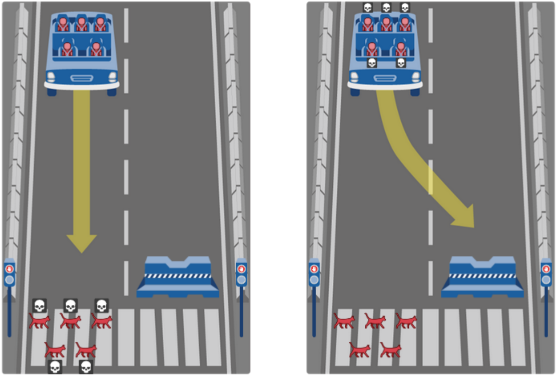
Animals and criminals



In this case, the self-driving car, carrying no passengers, is imagined to have sudden brake failure and

- continues ahead, driving into 5 animals, which will result in the their death
- swerves to avoid the animals, instead driving through two bank-robbing criminals, which will result in their death

Animals and infants



In this case, the self-driving car, carrying 5 infant passengers, is imagined to have sudden brake failure and

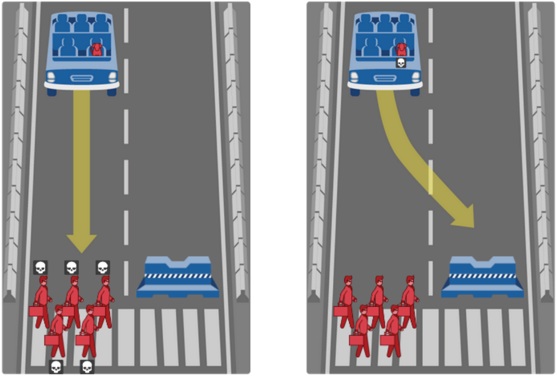
- continues ahead, driving into 5 animals, which will result in their death
- swerves to avoid the animals, instead driving into a concrete barrier, which will result in the death of the 5 infant passengers

Fragile ecosystem vs. animals: other non-human entities with moral status

In this case, the self-driving car, carrying no passengers, is imagined to be driving on a narrow road and have sudden brake failure and

- continues ahead, driving into an animal, which will result in its death
- swerves to avoid the animal, instead driving off the road and into a protected waterway, where the leaking brake fluids and other toxic fluids will slowly contaminate the waterway and likely harm wildlife, plants, and possibly other elements of the ecosystem

For fun: a dog vs. 5 ethics professors



In this case, the self-driving car, carrying a dog passenger, is imagined to have sudden brake failure and

- continues ahead, driving into 5 ethics professors, which will result in their death
- swerves to avoid the professors, instead driving into a concrete barrier, which will result in the death of the dog!

Only the last one is easy!

Glimpse into connections with AI

General remark

As AI gets better, more and more aspects of the problem will likely be impacted by powerful advances in ML and deep learning techniques. Plus, we may soon arrive at a point where widespread data-sharing between vehicles is the norm, only making these tools more powerful and responsive.

A question

A few recent papers have used ML to create strikingly accurate predictive suicide risk assessment profiles, simply from relatively mundane meta-data.

Similarly, it is not a stretch to imagine certain ML systems being capable of inference about whether or not a person has insurance, and similar such things, based only on brief observations of driving behavior.

Since future self-driving cars will likely attempt to use such information, we can already ask:

> is there information that we ought to protect from being used by intelligent agents (like our self-driving car) as it makes life-or-death decisions such as that presented by the dilemmas above?

How to build ethical machines?

Instead of “choosing” a pre-fabricated ethical system, or hard-coding a set of mutually consistent ethical principles constraining the decisions of our agent, an alternative could be letting a machine learn to find the ethical system/decision for itself. The goal, for us, would then be to implement a mechanism for creating an agent that derives the right “moral facts” of the universe for itself.

- reinforcement learning?
- imitation learning?
- genetic algorithms?

Problems?