

## Would you sacrifice one person to save five? - Eleanor Nelsen

Imagine you're watching a runaway **trolley** barreling down the tracks straight towards five workers who can't escape. You happen to be standing next to a **switch** that will **divert** the trolley onto a second track.

Here's the problem. That track has a worker on it, too, but just one. What do you do? Do you **sacrifice** one person to save five? This is the "trolley problem", a version of an **ethical dilemma** that **philosopher** Philippa Foot devised in 1967. It's popular because it forces us to think about how to choose when there are **no good choices**. Do we pick the action with the best **outcome** or stick to a **moral code** that **prohibits** causing someone's death? In one survey, about 90% of **respondents** said that it's okay to **flip** the switch, letting **one** worker die to save **five**.

And other studies, including a **virtual reality simulation** of the dilemma, have found similar results. These judgments are **consistent** with the **philosophical principle of utilitarianism**, which argues that the **morally correct** decision is the one that **maximizes** the **well-being** for the **greatest** number of people. The five lives **outweigh** one, even if achieving that outcome requires **condemning** someone to **death**. But people don't always take the **utilitarian** view, which we can see by changing the trolley problem a bit.

This time, you're standing on a bridge over the **track** as the runaway trolley **approaches**. Now, there's no second track, but there is a very **large** man on the bridge next to you. If you **push** him **over**, his body will stop the trolley, **saving** the five workers, but he'll **die**. To utilitarians, the decision is exactly the same: **lose one life to save five**. But in this case, only about 10% of people say that it's okay to **throw** the man **onto** the tracks. Our **instincts** tell us that **deliberately** causing someone's death is different than **allowing** them to die as **collateral damage**. It just feels wrong for reasons that are hard to explain.

This **intersection** between **ethics** and **psychology** is what's so interesting about the trolley problem. The dilemma in its many variations **reveals** that what we think is right or wrong **depends** on

factors other than a **logical weighing of the pros and cons**. For example, men are **more likely than women** to say it's okay to push the man over the bridge. So are people who watch a **comedy** clip before doing the thought experiment. And in one virtual reality study, people were **more willing to sacrifice men than women**.

Researchers have studied the brain activity of people thinking through the classic and bridge versions. Both **scenarios** activate areas of the brain involved in conscious decision-making and emotional responses. But in the bridge version, the **emotional response** is much **stronger**, and so is an activity in an area of the brain associated with **processing internal conflict**.

Why the difference? One explanation is that pushing someone to their death feels **more personal**, activating an emotional **aversion** to killing another person. But we feel **conflicted** because we know it's still the logical choice. Trolleyology has been criticized by some philosophers and psychologists. They argue that it doesn't reveal anything because its **premise** is so unrealistic that study participants don't take it seriously. But **new technology** is making this kind of ethical analysis more important than ever.

For example, driverless cars may have to handle choices like causing a small accident to prevent a larger one. Meanwhile, governments are researching **autonomous** military **drones** that could **wind up** making decisions on whether they'll risk civilian **casualties** to attack a **high-value** target. If we want these actions to be ethical, we have to decide, in advance, how to **value** human life and **judge** the greater good.

So, researchers who study **autonomous** systems are collaborating with philosophers to address the complex problem of **programming ethics** **into** machines, which goes to show that even **hypothetical** dilemmas can wind up on a **collision course** with the real world.

## The ethical dilemma of self-driving cars – Patrick Lin

This is a thought experiment. Let's say at some point in the **not-so-distant** future, you're **barreling** down the highway in your self-driving car, and you find yourself **boxed in** on all **sides** by other cars. Suddenly, a large, heavy **object** falls off the truck in front of you. Your car can't stop **in time** to avoid the **collision**, so it needs to make a decision: go straight and hit the object, **swerve** left into an **SUV**, or **swerve** right into a motorcycle.

Should it **prioritize** your safety by hitting the motorcycle, **minimize** danger to others by not **swerving**, even if it means hitting a large object and **sacrificing** your life, or take the **middle ground** by hitting the SUV, which has a high passenger safety **rating**? So what should the self-driving car do?

If we were driving that boxed-in car in **manual mode**, whichever way we'd react would be understood as just that, a **reaction**, not a **deliberate** decision. It would be an **instinctual** panicked move with no forethought or **malice**. But if a programmer were to **instruct** the car to make the same move, given **conditions** it may **sense** in the future, well, that looks more like **premeditated homicide**.

Now, to be fair, self-driving cars are predicted to **dramatically reduce** traffic accidents and **fatalities** by removing human **error** from the driving **equation**. Plus, there may be all sorts of other benefits: eased road **congestion**, decreased harmful **emissions**, and minimized **unproductive** and stressful driving time.

But accidents can and will still happen, and when they do, their outcomes may be determined months or years in advance by **programmers** or **policymakers**. And they'll have some difficult decisions to make. It's tempting to offer up **general decision-making principles**, like minimizing harm, but even that quickly leads to **morally murky** decisions. ‘

For example, let's say we have the same **initial** set-up, but now there's a motorcyclist wearing a helmet to your left and another one without a helmet to your right. Which one should your robot car **crash** into? If you say the biker with the helmet because she's more likely to **survive**, then aren't you **penalizing** the **responsible** motorist? If instead, you save the biker without the helmet because he's acting **irresponsibly**, then you've gone way beyond the initial design principle about minimizing harm, and the robot car is now meeting out street justice.

The ethical considerations get more complicated here. In both of our **scenarios**, the **underlying** design is **functioning** as a targeting **algorithm** of sorts. In other words, it's **systematically** favoring or discriminating **against** a certain type of object to crash into. The **owners** of the target vehicles will suffer the negative **consequences** of this algorithm through no fault of their own.

Our new technologies are opening up many other **novel** ethical dilemmas. For instance, if you had to choose between a car that would always **save as many lives as possible** in an accident or one that would **save you at any cost**, which would you buy? What happens if the cars start analyzing and factoring in the passengers of the cars and the **particulars** of their lives?

Could it be the case that a **random** decision is still better than a **predetermined** one designed to minimize harm? And who should be making all of these decisions anyhow? Programmers? Companies? Governments? Reality may not **play out exactly** like our thought experiments, but that's not the point. They're designed to **isolate** and **stress test** our **intuitions** on ethics, just like science experiments do for the physical world. Spotting these moral **hairpin turns** now will help us **manoeuvre** the unfamiliar road of technology ethics, and allow us to **cruise** confidently and **conscientiously** into our brave new future.