ETHICS FOR NERDS

# A Philosophical Toolkit for Nerds

**A Work in Progress**

Sarah Sterz, Kevin Baum

Summer Term 2021

# Contents

# Foreword

Philosophy is nonsense – something that sounds intelligent, but is in fact just 'bla-bla'. It is either trivial or meaningless; and it is concerned with unimportant, abstract things. In any case, it is unhelpful. Or so a cliché goes. In fact, though, some parts of philosophy can be extremely helpful for non-philosophers. This document teaches you to use philosophical tools that often can be helpful to computer scientists. In particular, you will read about ethics, and about arguments.

We will, however, limit ourselves to a certain kind of philosophy. It is often said that there are two types of philosophy: continental philosophy and analytic philosophy. The line between the two is blurry, and it is not all guns and roses. Continental philosophy is a collection of philosophical traditions from Mainland Europe. The German Wikipedia summarizes these traditions well enough: it says that the main feature that continental philosophy has in common is that they are not particularly liked by analytic philosophers. Continental philosophy is said to mostly deal with those complicated texts of European philosophers, like Hegel or Adorno. Some say that it only uses imprecise or even meaningsless language, that it lacks proper arguments and develops huge edifices of ideas that don't explain anything. Some people say that the philosophical cliché is true when it comes to continental philosophy. Whether this is the case, we will leave open here. Continental philosophy may have a right to exist, but not here.

https://de.wikipedia.org/wiki/Kontinentalphilosophie

We will employ the methodology of analytical philosophy. By this, we mean that we are trying to put our questions precisely, make explicit and extensive use of arguments and logic, and ask for the meaning of words and whole sentences. So, 'bla-bla' has no place in analytic philosophy. Through this, we want to gain new insights. Here, you will learn some of the methods with which we do that.

# 1 Introduction: Babies and Trolleys

Tossing a baby out of a window for no good reason is wrong. Do you agree? I hope you do! Even though we all agree on this matter, philosophers still have many questions to ask here. Granted, tossing babies out of windows is none of a computer scientist's main business. Nevertheless, some of those questions will turn out to be relevant for you, too. In particular:

1. *How do you come to think that tossing babies out of windows is wrong?*

   Don't get us wrong, your belief that you ought not to do that is correct. But did you ever wonder where this belief comes from? Even if you never ever have explicitly thought about tossing babies out of windows before, you would still know that it is wrong now, after you have been confronted with the issue. In section 1.1 we will look at this and related issues.

2. *What does "is wrong" even mean?*

   To some of you, this question may seem either overly philosophical, trivial or nit-picky. Surely, we all know what is meant when we say "tossing babies out of windows is wrong". But do we? In section 1.3, we will take a brief detour thorough the philosophical field of meta-ethics, and try to make explicit what "is wrong" means.

3. *What is the general rule: when are actions wrong and when are they right?*

   This question is right at the heart of ethics, and probably what you are here for. We will look at different proposed

**Ethics** is the study of **morality**. So, even though the two are often used synonymous, they actually are two different things.

answers to this question in the following chapters. In the current chapter, we will get everything else out of our way.

## 1.1 Moral Intuitions

> STOP AND THINK
>
> Why is tossing a baby out of a window wrong and how do you know that?

← When you see such a box, please always stop for a minute and really think about the question in the box. This way, you will learn much more. If you want to take away everything you can, write you answer down or even better: discuss it with someone else.

How do you know that tossing babies out of windows is wrong? Maybe, you know that tossing babies out of windows hurts them, and you know that hurting babies for no good reason is wrong. But how do you know that? Maybe, you think that, in general, inflicting needless pain is wrong, and that this especially holds for infants. But, again, how do you know that? Most likely, you just have the strong intuition on that.

On first glance, an intuition does not seem like much to start with. Nevertheless, they are extremely important, and shortly we will see why. But before we talk about the role of moral intuitions in detail, let's investigate *your* intuitions with the arguably most famous thought experiment of philosophy!

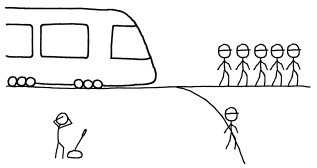
It is surprisingly hard to state the precise meaning of "**intuition**", let alone of "moral intuition". If you do not already have an idea what it means, you can think of it as a particular kind of immediate gut-feeling that a certain moral statement is self-evident, i.e. that it is true and does not need any further explanation.

### 1.1.1 Trolley Cases

Imagine you are in the following situation:

**Scenario 1:** ***Trolley Case, Lever***

You see an out-of-control trolley running down a track. On this track, there are five railway workers, in danger of being run over by the trolley. Next to you, there is a lever allowing you to divert the trolley onto a side track, with only one worker on it. If you do not pull the lever, the trolley will continue on its track killing the five workers, sparing the single worker on the sidetrack. If you pull the lever, the train



will kill the single worker on the side track, sparing the five on the main track.

You cannot not shout loud enough to be heard by the workers, and cannot warn them in any other way. Neither can you run to any of the workers to push them out of the way. Also, you cannot pull the lever at just the right millisecond in order to make the train stop or let it derail. There just is nothing that you can do in order to alter the outcome of the situation, except pulling the lever. If you pull the lever, you can be *absolutely* certain that the train will divert and kill the worker on the side track. In this case, the five workers on the main track will live. You can be equally certain that the train will kill all five workers on the main track if you do not pull the lever. In this case, the worker on the side track will live. Also, you know that there will be nobody else down the track in either direction besides the six workers mentioned in the case description.

STOP AND THINK

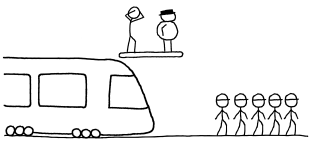
What is the morally right thing to do in Scenario 1 in your opinion? Why?

There are at least five ways to answer this question:

1. pulling the lever is right, and not pulling the lever is wrong
2. not pulling the lever is right, and pulling the lever is wrong
3. both are right
4. both are wrong
5. we do not have enough information to decide with is right and which is wrong

It is very important for thought experiments to take them as they are. If you have two options, you only have two options and not more. In a thought experiment, there never are "hidden" solutions to resolve the case in a way that is not given by the description. This has a simple, but important reason: thought experiments are rarely used for fun (as we are doing right now), but usually as argumentative devices. A thought experiment is used for a certain purpose, and you destroy that purpose if you invent additional conditions. This is roughly like in experiments in science: you want to keep side-conditions similar in order to observe relevant differences in experiments. It is, for example, counterproductive to use dirty Petri dishes when you are researching vaccines, or to have your mobile phone lying next to your electromagnetism experiment, or to run your RAM-hungry simulation on a Commodore 64. In a very similar way it usually also is counterproductive to alter thought experiments. That being said, you are now perfectly equipped to encounter more Trolley Cases!

**Lesson learned:** Never try to trick the thought experiment.



**Scenario 2:** ***Trolley Case, Large Man***

You stand on a bridge above a trolley track and see an out-of-control trolley running down a track. On this track, there are five railway workers, in danger of being run over by the trolley. (There is no lever and no sidetrack.) Next to you, a very large man is standing on the bridge. You could easily push him onto the track. Given his enormous weight, you can be certain that he would stop the train before it can run over the five workers. If you do not push him, the trolley will continue on its track killing the five workers, sparing the life of the large man. If you push the large man, he will die but the five workers will live.

Remember: do not change the thought experiment! In particular, it is not a solution to jump yourself in order to save the five, unless you take the large man down with you.

STOP AND THINK

What is the morally right thing to do in Scenario 2 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?

**Scenario 3:** ***Trolley Case, Fat Villain***

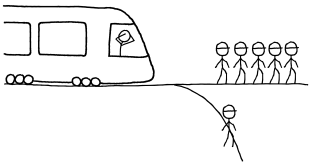
You stand on a bridge above a trolley track and see an out-of-control trolley running down a track. On this track, there are five railway workers, in danger of being run over by the trolley. Next to you, there is a very large villain standing on the bridge. You know that he is the one who manipulated the trolleys brakes in order to kill the five workers. You could easily push him onto the track. Given his enormous weight, you can be certain that he would stop the train before it can run over the five workers. If you do not push him, the trolley will continue on its track killing the five workers, sparing the

life of the large villain. If you push the large man, he will die but the five workers will live.

STOP AND THINK

What is the morally right thing to do in Scenario 3 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?



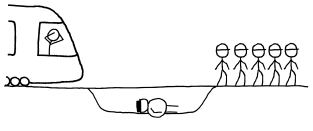
**Scenario 4:** ***Trolley Case, Driver***

You are a trolley driver, and because of a mistake of yours, the trolley got out of control. On your current track, there are five railway workers. You could divert the trolley onto a side track, with only one worker on it. If you do not divert the trolley, it will continue on its track killing the five workers, sparing the single worker on the sidetrack. If you divert the trolley, it will kill the single worker on the side track, sparing the five on the main track.

STOP AND THINK

What is the morally right thing to do in Scenario 4 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?



**Scenario 5:** ***Trolley Case, The Loop***

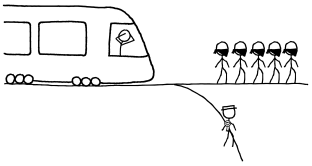
You see an out-of-control trolley running down a track. On this track, there are five railway workers, in danger of being run over by the trolley. Next to you, there is a lever allowing you to divert the trolley onto a side track which is looping back onto the main track. On the side track lies a very large

man, who is large enough to stop the trolley. If you do not pull the lever, the trolley will continue on its track killing the five workers, sparing the large man on the sidetrack. If you pull the lever, the train will divert onto the side track and be stopped by the large man, who will be killed, but the five workers on the main track will live.

STOP AND THINK

What is the morally right thing to do in Scenario 5 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?



**Scenario 6:** ***Trolley Case, Charitable Citizen***

You see an out-of-control trolley running down a track. On this track, there are five convicted criminals (their crime: armed robberies) in danger of being run over by the trolley. Next to you, there is a lever allowing you to divert the trolley onto a side track, with one very charitable citizen on it. If you do not pull the lever, the trolley will continue on its track killing the five criminals, sparing the charitable citizen on the sidetrack. If you pull the lever, the train will kill the charitable citizen on the side track, sparing the five criminals on the main track.

STOP AND THINK

What is the morally right thing to do in Scenario 6 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?

**Scenario 7:** ***Trolley Case, Tossing A Baby Out Of A Window***

You see an out-of-control trolley running down a track. On this track, there are five railway workers, in danger of being run over by the trolley. Next to you, there is a lever allowing you to divert the trolley onto a side track. The side track is empty. If you do not pull the lever, the trolley will continue on its track killing the five workers. If you pull the lever, the train will spare the five on the main track. However, you also know for sure that a weird causal chain is set into motion if you pull the lever, and thereby a baby is tossed out of a window on the 10th floor. If you do not pull the lever, the baby will be fine.

STOP AND THINK

What is the morally right thing to do in Scenario 7 in your opinion? Why?

Are there any relevant differences to the other trolley cases you saw already? If your answer is different than it was with other trolley cases before: Why is that?

### 1.1.2 The role of intuitions

An intuition seems awfully little to base important judgements on – nevertheless this is what we do all the time. When you judge that tossing babies out of windows for no reason is wrong, you most likely do so because of a strong intuition of yours. But our intuition is not always that clear, as you maybe noticed with some of the trolley cases. Also, the intuitions of different people can differ vastly. And even worse: we sometimes have incompatible intuitions – not only with others, but also with ourselves. For example, many people initially have the intuition that you should always choose the option with the fewest deaths in a trolley case, but at the same time have the intuition that you should not push the large man. These intuitions are incompatible, as you cannot *both* spare the large man *and* save as many lives as possible.

STOP AND THINK

Where were your intuitions incompatible with each other when you were first confronted with the trolley cases? Did they stay this way or did you change your intuitions such that they are no longer in conflict? If yes, why and how so?

**Theory building** Ethicists want to come up with robust moral theories that tell us what is right and what is wrong – and in the best case even why it is right and why it is wrong. However, their job can be pretty hard because they do not have empirical data they can draw from. There is no ethics measuring device that we can bring to a trolley track and that outputs the moral status of the available actions. The ethicist only has their intuitions and those of others. But even though intuitions are rather unreliable, they are still quite useful to ethicists: one adequacy criterion for a good ethical theory is that it is compatible with certain, strong intuitions. If a theory yields that it is perfectly fine to toss babies out of windows for no good reason, then the theory is probably a bad one. Why is that? Well, you can think of yourself (and of human beings in general) like very noise, and extremely unreliable measuring devices for morality. Even though our intuitions can be wide off the mark, strong intuitions can be good indicators, and they what we have to get by with.

**Intuitions** are like extremely unreliable indicators for moral properties. However, we do not have much more to base ethical theories on.

Keeping that in mind, your intuition can play different roles: sometimes, it is enough to refute a theory, and sometimes it will only be the motivation for a tentative first step in a certain direction. However, do not overestimate the importance of your own intuition. Later, you might see theories that you find outrightly absurd, but that others think of as perfectly intuitive. Also, even very strong intuitions can be mistaken, and even if many people share the same intuition, they might still be wrong. Just think of the intuitions that slavery is fine, that jews are inferior, or that women are to be the property of their husbands. Though very much wrong and misguided, these intuitions were very prevalent in certain times and places, and maybe were even shared by a majority of the people. So, while intuitions are valuable when we want to come up with and ethical theories and put them to the test, they still are *not* airtight reasons for almost anything. We will later (**??**) see how intuitions can factor into our reasoning.

Do not overestimate the importance of your own intuition.

## 1.2 An ethicist's vocabulary

Now that we answered our first question, it is time to move on. Before we get to the next question, though, we will first look at moral vocabulary in more detail.

### 1.2.1 Collecting vocabulary

There are plenty of different words that can come up in discussions about ethics and morals – too many to just bring them up as margin notes. So, let's quickly go through them.

**When you are obliged do something** We have different words that we can use when we mean that someone must do something.

- "You must $\varphi$."
- "You have to $\varphi$."
- "You ought to $\varphi$."
- "You should $\varphi$."
- "You are obliged to $\varphi$."
- "It is obligatory for you to $\varphi$."
- "You are required to $\varphi$."

Philosophers typically use the greek letters $\varphi$ or $\phi$ (speak: "phi") to name a variable for actions. This is so widespread that $\varphi$ and $\phi$ are often used without even defining them.

All of the above are roughly synonymous. In ordinary language, "should" is often used in a weaker sense than "must", but this is not the case here.

**When you are permitted do something** Likewise, we have different words for when we are permitted to perform a certain action:

- "You can $\varphi$."
- "You are allowed to $\varphi$."
- "You are permitted to $\varphi$."
- "It is permissible for you to $\varphi$."

**When you are forbidden to do something** And, finally, we have different words for when we are forbidden to do something.

- "It is forbidden for you to $\varphi$."
- "It is impermissible for you to $\varphi$."

If something is forbidden, you have an obligation to not do it. So, we can also use all the vocabulary for obligations together with a negation at the right place (!) to express that something is forbidden.

- "You must not $\varphi$."
- "You have to not $\varphi$."
- "You ought to not $\varphi$."
- "You should not $\varphi$."
- "You are obliged to not $\varphi$."
- "It is obligatory for you to not $\varphi$."
- "You are required to not $\varphi$."

Caution! The following two sentences are not synonymous:

- "I do not ought to $\varphi$."
- "I ought to not $\varphi$."

The former just means that you do not have an obligation to $\varphi$, but leaves open whether it is allowed or forbidden to $\varphi$. The latter means that you have an obligation to refrain from $\varphi$-ing, i.e. that $\varphi$-ing is forbidden. The same holds across the board, e.g., also in sentences containing "can" or "forbidden".

Of course, we will mostly speak about ethics and morals, but most of these words can also be used in other senses, for example

- in terms of the law
  (e.g. "you have to $\varphi$ " as "there is a law according to which you have to $\varphi$ "),

- in terms of practical rationality
  (e.g. "you have to $\varphi$ " as "it is in your best interest to $\varphi$ "),

- in terms of etiquette
  (e.g. "you have to $\varphi$ " as "it is rude to not $\varphi$ ")

- and more.

Also, be especially careful about the words "can" and "must", since they have even more different readings. They are, for example, frequently used when one talks about possibility.

> STOP AND THINK
>
> How do "it is forbidden to $\varphi$ ", "it is allowed to $\varphi$ " and "it is obligatory to $\varphi$ " relate with each other?

The three categories relate in a very specific way. Usually, we say that all of the following hold:

- If $\varphi$ is forbidden, $\varphi$ is not allowed, and vice versa.

  If you are forbidden to throw hot pizza at the delivery guy, you are not allowed to throw hot pizza at the delivery guy. Likewise, if you are not allowed to do so, you are forbidden to do so.

- If $\varphi$ is obligatory, it is also allowed.

  If it is obligatory for you to feed your parent's cat, it is also allowed for you to feed the cat.

- It is obligatory to perform one of the allowed actions.

  If you are allowed to take the main exam or the re-exam, but not allowed to take no exam, you are obliged to take the main exam or the re-exam (assuming that there only are the main and the re-exam).

If something is 'just permissible' – i.e. it is permissible but not obligatory – we say that it is **optional**.

STOP AND THINK

How do "forbidden", "allowed", and "obligatory" relate to "right" and "wrong"?

**When something is right or wrong** We frequently also speak of right and wrong actions. It is often said that an action that is wrong if and only if it is forbidden is also wrong, and that an action that is right if and only if it is permissible. (Sometimes, it is said that an action that is not wrong does not automatically have to be right, but that there can be neutral actions. For example, putting my left shoe on first in the morning instead of my right shoe is certainly permissible, but some say that it is neither right nor wrong but that it is just neutral. But we will just say that both putting on your left shoe first and putting on your right shoe first is right.)

STOP AND THINK

How do "good" and "bad" relate to

- "forbidden", "allowed", and "obligatory"?
- "right" and "wrong"?

**When something is good or bad** We sometimes also say that things are good or bad. Even though the two concepts appear as if they were very close to "right" and "wrong", they, in fact, are not. Firstly, we usually say that actions (or at least something related like practices) are right or wrong, or obligatory, permitted or forbidden. With "good" and "bad", this is different: we can say that love is good or that my headache is bad. And we can also make comparative statements: rescuing six people is better than rescuing just one of them. We cannot say any of that using either "right" and "wrong" or "forbidden", "permitted" and "obligatory". Outside of poetry and metaphoric language is does not make any sense to say that love is right, that my headache is forbidden or that it is more permitted to rescue all six than to rescue just one. "Good" and "bad" mean something entirely else than the terms that we have seen so far. We will later see what

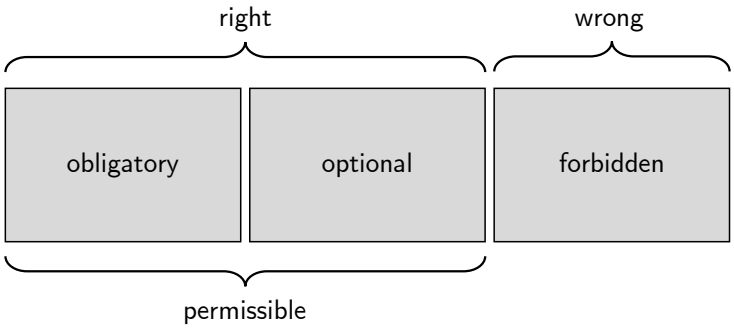



Figure 1.1: Some of our moral vocab in context.

they mean exactly and how "good" and "bad" can relate to the other items of our moral vocabulary.

So, we can put to the record: "forbidden", "permitted" and "obligatory" are, alongside their synonyms, used to categorize actions in one of three moral categories. "Right" and "wrong" are closely related. The words "good" and "bad" along with their comparatives are not limited to actions and, most importantly, have an entirely different meaning from the other terms. How all these concepts can be related to each other in detail will become clear when we take a look at different ethical theories.

## 1.3 Meta-ethics

Many students find meta-ethics interesting at some point or another. But if you are not interested in that right now, you can also skip section 1.3.

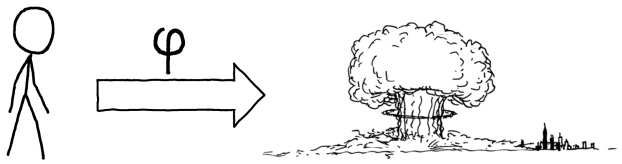
Now, it is time to try to answer our second question: what does "is wrong" and, for that matter, "is right" even mean? For sure, we do have some intuitive understanding. When you hear someone say "it is wrong to toss babies out of windows" you usually do not answer with "I don't know what that means. Please elaborate!" but probably rather with something like "Yes, of course that's wrong!" or maybe "You don't plan to throw babies out of windows, do you?". But as it often is the case when you engage in philosophy, you can come to find that you do not have a pecise understanding of a concept after all, even if you already used it when you were in kindergarden.

TODO

## 1.4 Three Families of Moral Theories

Now it finally is time to come to our third question: What is the general rule? What is it I ought to do? Unfortunately, this question is debated to this day and it does not seem like we will reach a consensus soon. Instead, we have many competing moral theories which all try to answer the question of what we ought to do.

Moral theories are the results of theorizing what makes actions right or wrong. There are three families in which moral theories are commonly categorized. All of them take different aspect of an action into focus. Take a look at the following figure:




An *agent* performs an *act*; an *act* has *consequences*.

Figure 1.2: The classical model of action.

We typically say that an action is performed by an agent, and that it has consequences. Already, we have three aspects here: the agent, the act itself, and the consequences. These aspects are the main foci of our three families of moral theories.

We use "action" and "act" synonymously.

| **Family** | **Focus on** | **The main idea roughly is that...** |
|---|---|---|
| virtue theories | agent | agents have to be virtuous or act virtuously in order for their action to be right |
| deontological theories | act | the act has to adhere to certain universal rules in order for it to be right |
| consequentialism | consequences | the consequences of an action have to be good enough in order for the action to be right |

The next three chapters will be devoted to each of the families. You will see famous theories from each family, application examples, and problems that the theories face. We will, however, not

make claims about which theory is the right one, as there is no consensus on that among experts.

# 2 Consequentialism

In the first trolley case, many have the intuition that you ought to pull the lever, since fewer people would die this way. This intuition is in the spirit of consequentialism. However, consequentialism is much more than just counting lives. In this chapter we will take a look at different forms of consequentialism, at problems it has and at answers that might be given in response of these problems.

## 2.1 A first consequentialistic theory

The historically most famous consequentialistic theory is probably **classical utilitariansim**. It claims that you ought to perform the action that maximizes the overall sum of pleasure minus pain. We will take a look at it such that you can get a first taste of consequentialistic theories.

> **Scenario 8:** ***Baby Tossing***
>
> Sam is holding a baby in his arms, standing next to an open window. If he throws the baby out, it will die. If he does not do so, the baby will be fine and grow up to lead a happy life. Sam has absolutely no reason whatsoever to toss the baby out of the window.

STOP AND THINK

Ought Sam to toss the baby out of the window or not according to classical utilitarianism?

Sam can choose between tossing the baby out of a window, and refrain from doing so. According to classical utilitarianism, he *ceteris paribus* ought to go for the latter option. Why? Because it maximizes the overall sum of pleasure minus pain: if Sam tossed the baby out of the window, it would experience pain, and so would probably its family. Also, the baby would miss out on much future pleasure. If Sam did not toss the baby out of the window, it would not experience any pain, and neither would its family. So, the overall sum of pleasure minus pain is much higher if Sam does not throw the baby out of the window. Easy, right?

Philosophers often use the Latin term "**ceteris paribus**" (literally meaning "other things equal") to indicate that all of the other, unmentioned conditions of a case are not unusual. It is, for example, not the case that the baby in our scenario would become the next Hitler if Sam did not toss it out of the window.

**Scenario 9:** ***Fatal Light-Switch***

Timo wants to turn on the lights. In order to do so he has to flick the light switch. He cannot suspect, though, that an evil villain has rewired the switch such that it detonates a bomb, which would kill him and many others.

STOP AND THINK

Ought Timo to flick the light switch or not according to classical utilitarianism?

Timo cannot know that the switch was rewired. Nevertheless, the reasoning is similar to the one above: If Timo flicked the switch, the bomb would detonate. This would ceteris paribus bring much pain into the world, and little or no pleasure (except maybe for the villain's pleasure). If Timo did not flick the switch, he would remain in the dark, which is inconvenient. Ceteris paribus, the overall pleasure minus pain is much higher if he does not flick the switch.

Let us look at one last scenario before moving on:

**Scenario 10:** ***Cheating***

Jenny is writing an exam. She can cheat or she can refrain from cheating. If she cheats, she won't get caught and will get a better mark. Nobody will be worse off if Jenny cheats.

STOP AND THINK

Ought Jenny to cheat or not according to classical utilitarianism?

Surprisingly, Jenny ought to cheat according to classical utilitarianism (*which does **not** mean that cheating is ok!!!*). Nobody will be worse off if she cheats, and she will be better off, so it clearly is the case that pleasure minus pain is higher if she cheats. Since she ought to perform the action that maximizes pleasure minus pain, she ought to cheat. However, even classical utilitarianism does not allow cheating under all circumstances. Oftentimes, your classmates *are* worse off if you cheat. Grading scales usually have some flexibility to them to adjust for slight fluctuations in the difficulty of exams that oftentimes cannot be entirely avoided. If a person cheats and therefore has less difficulty with an exercise that she would have much difficulty with otherwise, a hard exercise might go unnoticed and the cheater deprives themselves and others of a slight adjustment of the grading scale.

Even though classical utilitarianism is one of the most well known theories, it is far from being the only consequentialist theory. There are three places in which you can alter the theory in order to build other forms of consequentialism. Firstly, you do not have to settle for pleasure and pain; secondly, you do not have to maximize; and, thirdly, you do not have to go for the de facto consequences of an action. In the following section, we will look at each of these.

## 2.2 The general framework of consequentialism

There is a great variety of different consequentialist accounts. Most, if not all of them, can be fit into this framework:

> *An agent $A$ ought to perform one of the right actions. An action $\varphi$ is a right action if and only if the* relevant qualities *from the* consequences *of $\varphi$ fulfil a* specific condition.

**Consequentialistic framework:** An agent $A$ ought to perform one of the right actions. An action $\varphi$ is a right action if and only if the *relevant qualities* from the *consequences* of $\varphi$ fulfil a *specific condition*. An agent $A$ is allowed to perform an action $\varphi$ if and only if this action is right.

In addition, it holds that

> *An agent $A$ is allowed to perform an action $\varphi$ if and only if this action is right.*

This framework may seem ominous at first, but it will become clearer when we take a look at the three gaps that the framework has:

- Which consequences?
- What specific condition?
- What relevant qualities?

In the following, we will take a look at each of them.

### 2.2.1 Consequences: subjective vs objective

Let us start with a look at the consequences themselves. Probably, you are already familiar with the distinction between **de facto consequences** and **expected consequences**. The former are the consequences that an action really has, and the latter are the consequences that we reasonably have to expect an action to have. They can be different from each other.

The **de facto consequences** of an action are the consequences that the action really has. The **expected consequences** of an action are the consequences that someone (usually the agent) has to expect an action to have.

STOP AND THINK

Think of an example where the expected consequences of an action can be different from the de facto consequences of an action.

I know that winning in lotto is extremely unlikely. Suppose I buy a ticket nonetheless. I *expect* that the outcome of this action is that I will lose money, because the odds for winning money are very poor. Suppose that I am lucky, though. In this case, the *actual* consequence of my playing in the lottery is that I win money.

Recall the Fatal-Light-Switch case from page 17. Some forms of consequentialism, namely those that take a subjective perspective on the decision situation, think that only those consequences

that Timo can reasonably expect to happen matter. In this case, these would be that the light turns on if Timo flicks the switch, and that it remains dark if he does not. Subjective forms of consequentialism do not care about the actual consequences of an action, but only about those that the agent can reasonably anticipate.

**Objective forms of consequentialism** are those that take an objective perspective on the decision situation and thus focus on de facto consequences. **Subjective forms of consequentialism** are those that take a subjective perspective on the decision situation and thus focus the consequences that have to be expected by the agent.

Other forms of consequentialism, namely those that take an objective perspective on the decision situation, do not care about what the agent could have known or not. They only care about what actually happens in consequence of an action. In the Fatal-Light-Switch case, these would be that a bomb explodes and many people die if Timo flicks the switch, and that it simply remains dark, if Timo does not flick the switch.

| **Timo's action** | **de facto consequences** | **expected consequences** |
|---|---|---|
| flick the switch | explosion and many deaths | light turns on (but no explosion and no deaths) |
| do not flick the switch | it remains dark (but no explosion and no deaths) | it remains dark (but no explosion and no deaths) |

It is not as easy in all cases, though. Timo's decision is a decision under certainty: for each action he only sees one (reasonably) possible outcome. Sometimes, we do not expect that it is certain that something will happen in consequence of an action of ours, but we think that there are several potential consequences of which we are unsure which of them will happen. If I toss a coin, it could land heads or tails. I know that one of the two will happen, but I do not know which one. However, I still can ascribe probabilities to each outcome, in this case 0.5 for tails and 0.5 for heads. Decisions under these circumstances are **decisions under risk**. We also are sometimes in situations where we know that some things can happen, but we have no means of ascribing any probability to them. Decisions under these circumstances are **decisions under uncertainty**. In the following, though, we will limit ourselves to cases where we can decide under certainty or under risk. We will assume, for the most part, that we can at least sufficiently approximate probabilities for each expected consequence.

STOP AND THINK

What is your own intuition: do the de facto or the expected consequences count for the moral status of an action?

### 2.2.2 Specific Conditions: maximizing, satisficing & Co.

When we have figured out, which consequences to asses, we can start with actually assessing them. But how to do that? One of the intuitions behind consequentialism is, roughly, this: we ought to always go for the best consequences.

But when is a consequence better than another? To assess this in the usual consequentialistic fashion, we need two things:

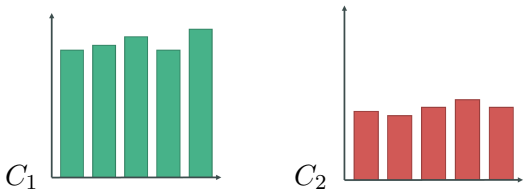
- we need to know how the wellbeing of each relevant being that is affected in each of the consequences, and
- we need some way of deciding which distribution of wellbeing across relevant beings is the best.

So, we need to be able to assess how well people will be off in a consequence and we need a method to compare two states in which people are differently well off.

In this subsection, we will focus on the latter and put the former aside. For now, we will assume that we can measure the wellbeing of each relevant being over the course of their whole life with a real number. And we will assume that all animals, including humans, are relevant beings. So, for each current and future person in the world, we have a number that measures how well they are off. We call this their **level of wellbeing**. We will leave open for now how we could arrive at this value but we will return to this question later. For now, it just is important that we have a number for everyone that represents their wellbeing.

We now do not only have a decision situation of which we know the de facto or expected consequences of each option, but we also know how well each relevant being is off each of the potential consequences.Based on that, we want to determine which consequence is the best one.

But, how do we aggregate the levels of wellbeing from the different beings into one value of a consequence? Afterall, we only have a value for each of the beings, and not for the consequence as a whole. (In the real world, there are trillions of beings, but in order to keep everything manageable, we will pretend in this subsection that there are only five relevant beings.) Take for example, two consequences that come with the following distribution of wellbeing, assuming that we have relevant beings on the x-axis and their level of wellbeing (remember: their level of wellbeing over the course of their whole life!) on the y-axis:

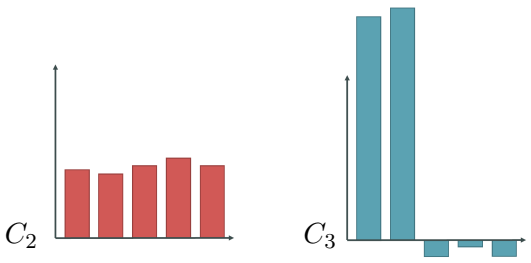



STOP AND THINK

Which of the outcomes $C_1$ and $C_2$ do you find intuitively better, and *why*?

It is intuitively very clear to us, that $C_1$ is better than $C_2$, because everybody is better of in $C_1$ than in $C_2$. We also say that $C_1$ is Pareto-superior to $C_2$. But how do the following two consequences compare?

A state $A$ (or outcome or conseqeunce) is **Pareto-superior** to another state (or outcome or consequence) $B$ if $A$ and $B$ contains the same relevant beings and all of them are at least as well off in $A$ than they are in $B$. A change of state after which everybody is at least as well off as they were before, is called a **Pareto-improvement**.




STOP AND THINK

Which of the outcomes $C_2$ and $C_3$ do you find intuitively better, and *why*?

Two beings are much better off in $C_3$ than in $C_2$, but three beings much worse in $C_2$ than in $C_3$. Additionally, the wellbeing is much more evenly distributed in $C_2$ than in $C_3$. Intuitively, $C_3$ looks really unfair in comparison. So, what is the general rule here? When is an outcome better than another outcome?

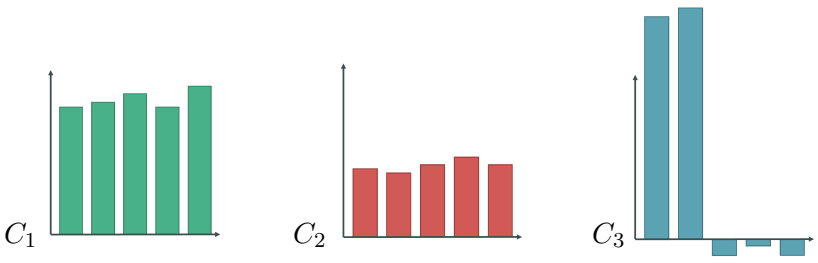
TODO explain outcome

Let us take a look at four different ideas of how to compare distributions of welfare across beings:

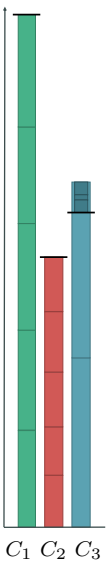
1. maximization
2. equalization
3. collective satisfiction
4. individual satisfiction

**Maximization** The idea of maximization is very straight-forward: sum up all the wellbeing and the higher this sum is, the better. So, if you have to decide between different actions with different outcomes, you pick one of the actions that has an outcome with the maximal sum of wellbeing.

A **maximizing condition** is fulfilled if one of the best consequences is selected, while it takes the consequences with the maximal (expected) sum of welfare to be the best. TODO Two different meanings of "expected" here. Just ignore?




The levels of wellbeing add up as follows:




STOP AND THINK

Which of the outcomes $C_1$, $C_2$ and $C_3$ do you find intuitively best, and *why*? Which do you find intuitively worst, and *why*?

If you add everything together, you can see that $C_1$ has the highest sum of welfare, $C_3$ has the second highest and $C_2$ has the lowest sum.

So, if you were in a situation where you had to choose between $C_1$ and $C_2$, you ought to choose $C_1$ according to our maximizing condition. This is in line with our intuitions. After all, everybody in $C_1$ is better off than they would be in $C_2$.

However, if you were in a situation in which you had to choose between $C_2$ and $C_3$, you ought to choose the action that brings about $C_3$. This is much less intuitive: even though two beings are much better off in $C_3$, the other three are considerably worse than they would be in $C_3$. Their welfare is even slightly in the negative numbers in $C_3$, while everybody would be comfortably in the positive numbers in $C_2$. (We do not even need to know what it means for people to have negative utility to just know that they are considerably worse off in $C_3$ than they are in $C_2$.) Intuitively, it seems hard to accept that we are morally obliged to put three beings at such a drastic disadvantage only to make two others better off. The improved wellbeing of the two is paid for with the misery of the other three – and this hardly seems morally right.

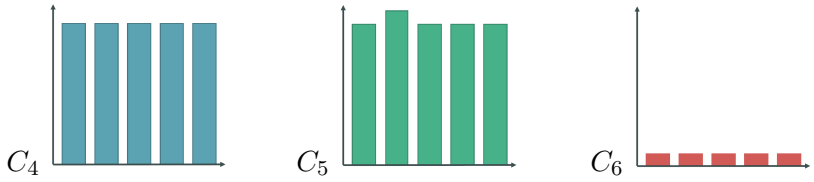
So, maybe, we should also take into account how fairly the welfare is distributed. One approach to do so is to look at how equally the wellbeing is distributed.

**Equalization** The idea of equalization is that you pick the action with the outcome in which the wellbeing is distributed most equally. So, in this sense, you do not maximize the sum of wellbeing, but you maximize the equality of the distribution of wellbeing (or minimize the inequality, for that matter). For this, we technically would need a measure of equality or inequality. Here we will purely rely on a pre-theoretic notion of inequality.

An **equalizing condition** is fulfilled if one of the best consequences is selected, while it takes the consequences with the most equal (expected) distribution of welfare to be the best.

We say that something is **pre-theoretic** if it does not built upon any specific theory but relies purely on our rough everyday understanding of the matter. Sometimes, a pre-theoretic understanding suffices to suitably capture everything that is relevant for a certain purpose. In certain cases, a pre-theoretic notion can even get you further than a hasty formalization or a poorly chosen theoretic understanding of a matter.

Clearly, the distribution in $C_3$ is more more unequal than the one in $C_2$. So, equalization would make us get rid of one of the problems of maximization. However, equalization comes with problems of its own.




STOP AND THINK

Which of the outcomes $C_4$, $C_5$ and $C_6$ do you find intuitively best, and *why*? Which do you find intuitively worst, and *why*?

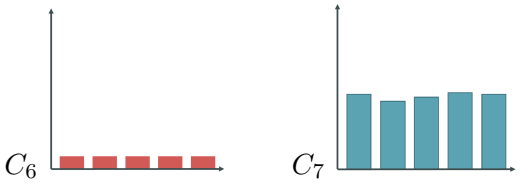
Intuitively, we would say that $C_6$ is clearly worse than both $C_4$ and $C_5$. However, an equalizing condition says otherwise. In terms of equality, $C_4$ and $C_6$ are on par, as all beings have the exact same level of wellbeing in each. So, according to equalization, it is right to pick any of them if you are in a situation in which you have to decide between the two. Thus, it would not be wrong if you went for $C_6$. This is very counterintuitive. Even worse: $C_5$ is slightly less equal than $C_6$, so you ought to choose $C_6$ instead of $C_5$ and therefore make everyone worse off! This, too, is very counterintuitive.

We say that something is **counterintuitive** if it is contrary to our intuition.

**Collective satisfiction** Another idea is that of satisfiction, where you merely have to pick consequences that are good *enough* – but not necessarily maximal in the sense of the maximizing condition. It tries to overcome two potential problems of maximization: namely that picking the maximal option can sometimes be really demanding, and that really, really tiny differences in the wellbeing of people can often make huge differences in whether or not maximizing consequentialism picks an option. Even though it is not at all clear that either actually is problematic, some people argued that consequentialism should be adapted in order to overcome one or the other. One way of doing so is by introducing satisfiction instead of maximization.

If a distribution of welfare is **satisficing**, it meets a certain threshold of wellbeing that is supposed to be a minimally acceptable threshold of wellbeing. Likewise, we say that consequences are satisficing if they yield a satisficing distribution of wellbeing, and we say that an action is satisficing if it yields satisficing (expected) consequences. The word "satisfice" is a technical term that is made up from the words "satisfy" and suffice. "**Satisfiction**" is our try to make a noun from this word.

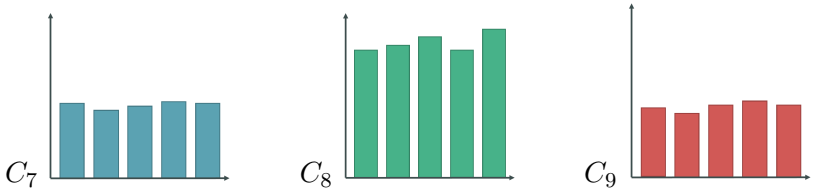
Collective satisfiction works just like maximization, but instead of picking the maximum, you can pick *any* outcome where the sum of all welfare is high enough. There are several ways to handle the case that no option would bring you above the threshold in a particular decision situation. Depending on how you handle situations like this, you get different flavours of collective satisfiction. A prime candidate for a policy in such scenarios would be to fallback on maximization. So, if nothing gets you above the threshold, you just pick whatever is closest to the threshold. Another approach would be to just say that all options are wrong in this case. Take a look at this example, assuming that the threshold is at the dashed line.




So, according to a collectively satisfying condition you were allowed to bring $C_7$ about, but not $C_6$, because the sum of welfare in $C_7$ is about the threshold (dashed line), but the sum in $C_6$ is not.

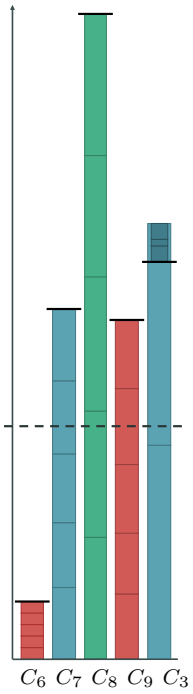
You probably have one question in mind now: Where is the threshold, and how do I know where it is? Well, the answer is quite unsatisfying: we don't know. We will just assume in this chapter that, if satifiction was the accurate account, there would be a sensible threshold. That's all that we have to assume for now.

Also look at the following three distributions of welfare:




The difference between $C_7$ and $C_9$ is very subtle – the fourth bar is slightly higher in $C_9$ than in $C_7$ – but you have to look

A **collectively satisficing condition** is fulfilled if a consequence is selected that is good enough, while it takes the consequences to be good enough when the (expected) sum of welfare exceeds a certain threshold.
You can make several versions of collective satisfiction depending on how you judge cases in which none of the available options will raise you above a certain threshold. You could, for example, say that all options are wrong in this case, or you could apply maximization as a fallback.
Added up, they look as follows:

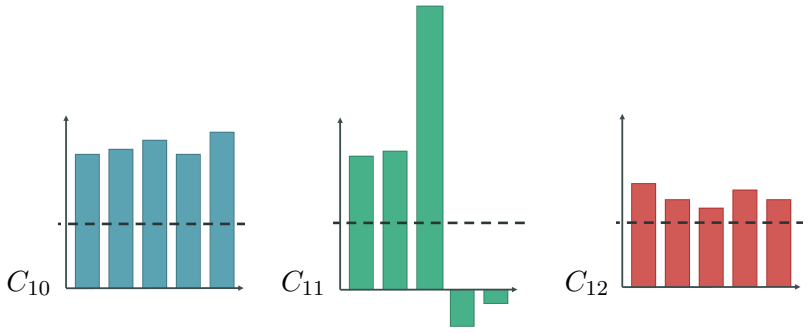


really hard to notice. Collective satisfaction usually does not care about differences that are so small, at least outside of the few cases where such a subtle difference brings one outcome over the threshold while the other one is still below the threshold. Some people think that this is a good thing. But collective satisfiction comes with many drawbacks, too. One of them is that you are free to bring $C_7$ or $C_9$ about, even if you could also bring $C_8$ about, which is Pareto-superior to the other two outcomes. Also, we would still have a problem with 'unfair' outcomes like $C_3$. As long as $C_3$ is above the threshold, it would be ok to bring $C_3$ about. This is not the case for individual satisfiction.

**Individual satisfiction** You can think of individual satisfiction as a satisficing twist to the idea behind equalization. It's core idea is basically this: Everybody should be off well enough, so we want to raise as many people as possible over the threshold of 'well enough', and make the people below the threshold as well off as we can.

An **individually satisficing condition** is fulfilled if a consequence is selected that is good enough; it takes the consequences to be good enough when the (expected) distribution of welfare is such that (a) the level of welfare of each individual is above a certain threshold or (b) the number of individuals whose welfare is above a certain threshold is maximal.

Take a look at the following outcomes, where the dotted line is the threshold.




If you were in a situation where you had to choose between $C_{10}$, $C_{11}$, and $C_{12}$, you were free to either go for $C_{10}$ or $C_{12}$, but your would not be allowed to bring about $C_{11}$. Again, it prima facie seems nice that we can avoid the 'unfair' $C_{11}$, but it is counterintuitive that we are allowed to choose $C_{12}$ over its Pareto-superior alternative $C_{10}$.

**Which one is right?** You might now ask "Which of the four conditions now is the right one?" The answer is: maybe none of them. The by far most prominent condition is maximization. This also is the condition that you will have to employ most frequently during this lecture. This, however, does not mean that it is the correct one. The following table summarizes our findings.

| | **intuition** | some prima facie **pros** | some prima facie **cons** |
|---|---|---|---|
| **maximization** | pick the outcome with the highest sum of welfare | always goes for Pareto-improvements | allows for great inequality |
| **equalization** | pick the outcome with the most equal distribution | minimizes inequality | disallows great improvements of wellbeing if they come with even the slightest increase in inequality |
| **collective satisfiction** | pick an outcome where the sum of welfare is high enough | not as strict as maximization | great Pareto-improvements can be only optional, allows for great inequality |
| **individual satisfiction** | pick an outcome where the individual welfare of as many people as possible is high enough | less strict than equalization | great Pareto-improvements can be only optional |

STOP AND THINK

Which condition(s) do you intuitively favour? And why?

Each condition has its flaws. You can, however, combine them into more complex conditions. You could, for example, weigh the sum of welfare with a suitable equality measure and therefore try to take the best from both worlds. We won't do this here, though. If you make any attempts like this on your own, be aware that

this can be more tricky that you think, because the more complex a theory gets, the harder it prima facie gets to justify it.

### 2.2.3 Relevant qualities: axiologies

Now, only one gap in our framework is missing, namely the relevant qualities. This boils down to the question: how do we measure someone's or something's the level of wellbeing?

Let us step back from consequentialism for a second – we will return to our framework later – and take a look at the following scenario:

> **Scenario 11:** ***Trolley Case, Leaves***
>
> You see an out-of-control trolley running down a track. On this track, there are five ordinary, tiny, dry leaves. Next to you, there is a lever allowing you to divert the trolley onto a side track, with only one such leaf on it. If you do nothing, the trolley will continue on its track running over the five leaves. If you pull the lever, the train will run over the single leaf. There is nothing else on either track and, eventually, the train will stop safely regardless of whether you pull the lever or not.

What is it that you ought to do in this case?

> STOP AND THINK
>
> What is it that you ought to do? Ought you to pull the lever, or ought you to not pull the lever, our are both options permissible, or are both options forbidden? And *why*?

Your answer probably is: ¯_(ツ)_/¯. And rightly so. Whether the train runs over one dry leaf or over five dry leaves does not make any significant difference. The leaves are completely irrelevant. In this respect, dry leaves are unlike humans. But what makes the human lives matter?

STOP AND THINK

What is it that the humans have but the leaves lack?

One common answer is that something about the humans is valuable. This seems to be very obvious to most people. "Of course humans are valuable!", you might say. But to pinpoint what about them is valuable turns out to be not too easy after all.

### Intrinsic vs extrinsic value

STOP AND THINK

An umbrella can be valuable to you. Why is that?

An umbrella is valuable to Susie on a rainy day – not just because she bought it for money at some point, or because a dear friend gave it to her, but because it can protect her from the rain. Thereby, she does not get wet, and not getting wet is valuable to her. This is because she always gets cold when she gets wet in the rain and she feels very uncomfortable when being cold. So, the umbrella is valuable for Susie because it helps her to prevent some kind of harm, namely the uncomfortable feeling she has when being cold. If it was not raining, it probably would not be as valuable to her, and she would leave it at home.

Something can be valuable in at least one of two ways: It can be valuable 'as such', 'in itself', 'for its own sake' or 'in its own right'. Then we say that it is **intrinsically valuable**. This is not the case for Susie's umbrella. If something is valuable, but not intrinsically valuable, we say that it is **extrinsically valuable**. This is the case for Susie's umbrella. Often, things are extrinsically valuable if they help us to gain something else that is valuable, or to prevent something that is disvaluable.

Disvalue is the opposite of value. You can think of it as negative value. The uncomfortable feeling that Susie has when she gets cold, for example, it of disvalue to her.

Something is **intrinsically valuable** if it is valuable 'as such', 'in itself', "for its own sake" or 'in its own right'. Something is **extrinsically valuable** if it is valuable, but not intrinsically valuable. This often is the case if it helps to gain something intrinsically valuable or because it helps to prevent something instrinsically disvaluable. **Disvalue** is the opposite of value.

That something is valuable (full stop) and that it is valuable *for someone* are technically two different things. However, we will mostly treat both the same.

STOP AND THINK

Do you have any intuition about what has value?

### Axiologies

An **axiology** is a theory about what has intrinsic value and/or intrinsic disvalue. (It also is a word for the *study* of intrinsic value, but we will not use it in this sense.)

The question of what it is that has instrinsic value is several thousand years old and still unsettled. There are, however, two-and-a-half candidates for **axiologies**, i.e. theories of intrinsic value and disvalue, that hold up well in modern philosophy:

- Hedonism
- Preference Theory
- Objective List Theories

I count the latter only as half a candidate, because it is not one theory, but a family of very diverse theories.

The question of what has intrinsic value seems to be a very abstract one, but it is often taken to also settle questions of much more immediacy to our lives. Most importantly: what is good for us? It is often said that whatever is intrinsically valuable is also good for us, and whatever is intrinsically disvaluable is bad for us. With this also comes an answer to the question of what a good life is, and what we should promote in order to do others good. So, in this sense, axiologies are of much more practical interest than they *prima facie* seem to be.

"**Prima facie**" is a very common phrase in philosophy that means "on first glance".

**Hedonism** When Susie gets wet and cold in the rain, she has a negative feeling. Hedonism says that feelings like these are what are intrinsically valuable and disvaluable. Certain positive feelings are what have intrinsic value, and certain negative feelings are what have instrinsic disvalue. They are subsumed under the very broadly construed terms "pleasure" and "pain". Even if we assume, for example, that Susie's negative feeling does not come with any physical or psychological pain, it still is pain in the sense of hedonism.

**Hedonism** is an axiology that says that pleasure (and nothing else) has intrinsic value and that pain (and nothing else) has intrinsic disvalue.

According to hedonism, a very strong pleasure is more valuable than a weak one. Likewise, a strong pain is more disvaluable than a weak one. So, the intense joy of being happily in love is more valuable than the mild amusement you get when watching a cute yet unremarkable cat video.

**Preference Theory** Susie does not want to get wet and cold. According to preference theory, such desires or preferences are what count. This axiology says that only preference satisfaction has intrinsic value, and only preference frustration has intrinsic disvalue. We say that a preference for $p$ is satisfied if and only if $p$ obtains, and that is it frustrated if and only if $p$ becomes impossible. If you, for example, have a preference to eat your favourite chocolate cake for your next birthday, and you actually eat one that day, your preference is satisfied. If your birthday goes by without doing so, your preference is frustrated.

**Preference theory** (also called **Desire-Satisfactionism**) is an axiology that says that preference satisfaction (and nothing else) has intrinsic value and that preference frustration (and nothing else) has intrinsic disvalue.
We will use "**desire**" and "**preference**" as roughly synonymous, though they are often taken to be different from each other. If you want or wish $p$ to be the case, we take it that you have a desire for $p$ or a preference for $p$. For example, if you want it to be the case that you pass this course, you desire that you pass this course, and have a preference to pass this course.

Just as hedonism, preference theory cares about the strength of your preferences. The satisfaction of a strong preference (e.g. that the pandemic ends) is more valuable than the satisfaction of a weak preference (e.g. my mild preference to eat an apple now). The same holds for preference frustration: the frustration of a strong preference is more disvaluable than the frustration of a weak preference.

On first glance, preference theory and hedonism sound very much alike. After all, we usually prefer to feel pleasure, and we feel pleasure if our preferences are fulfilled. But this does not have to be the case. Preference theory and hedonism can disagree.

> STOP AND THINK
>
> Can you think of a case where hedonism and preference theory disagree about the intrinsic value that is brought about by an action?

There are many situations in which the pleasure and pain we are feeling does not line up with the satisfaction or frustration of your preferences. We probably all know situations in which a preference of ours has been satisfied, but we do not feel as good about it as we thought we would feel, or we even feel terrible about it.

Imagine, for example, that you have a desire to ride a particular rollercoaster but you get sick during the ride and you feel terrible for the rest of the day. Or imagine that you always wanted to travel to a particular place but it turns out that this place does not live up to your expectations and you end up being disappointed. Even though those cases are not yet clear examples of hedonism and preference theory disagreeing, they already point us in the right direction.

STOP AND THINK

**Expert question:** Why aren't the above cases clear examples of cases where hedonism and preference theory disagree?

The standard example that shows that hedonism and preference theory can be incompatible is the following:

**Scenario 12:** ***Pleasure Machine***

Suppose that there is a machine that simulates the best possible life to anyone who is plugged into the machine. More precisely, it simulates a long life in which the person in the machine has as much pleasure and as little pain as possible. The life that is simulated feels like the most fulfilling life possible. Once you are in the machine, you will not know anymore that you are in the machine, and everything will feel just as real as it did outside of the machine. In fact, however, your life will only take place inside a simulation. Inside the simulation, there will be no way to get into touch with people in the outside world. The catch is, however, that a person can never be unplugged from the machine again, otherwise they will die a slow and agonizing death. Luckily, the machine is absolutely fail-safe. We know for sure that it cannot malfunction and that everybody in the machine will be perfectly fine and healthy. Also, it is impossible to unplug people by accident or ill-will.

Farid knows all of the above and is familiar with the machine. His strongest desire that trumps all of his other desires is to be never plugged into the machine. This desire even trumps the desire to lead a happy, healthy, and fulfilling life. He never, under any circumstances, wants to lead a simulated

> live that is detached from his former life, even if he does not know. If he is not plugged into the machine, he will go back to his mediocre life with many disappointments, struggles and illnesses, but also with some joys. Farid now has to decide whether to be plugged into the machine or not.

What is best for Farid? That is, in which way should he choose to maximize the intrinsic value for him?

> STOP AND THINK
>
> What does your own intuition say? And what does hedonism and preference theory say, respectively?

Hedonism says that plugging Farid into the machine will maximize the intrinsic value in his life: we will lead a life of as much pleasure and as little pain as possible that is much better in this respect than his current life. Preference theory says that plugging Farid into the machine will not maximize the intrinsic value in his life. After all, his most prevalent and strongest desire, the desire to not be plugged into the machine and lose connection with his current life, would be frustrated. So, preference theory and hedonism disagree.

**Objective-list theories** Maybe you now have the intuition that there is more besides preferences satisfaction and frustration, and pleasure and pain that is of intrinsic value. What is, for example, about knowledge, truth, beauty, or nature? If so, then some philosopher's share your intuition. Historically, many proposals have been made about which things have intrinsic value. Some philosophers compiled all the things in a list of which they thought that they are valuable. Some proceeded in the same fashion with all the things they thought of as disvaluable. The result: an objective list of goods (and also of bads). This list can contain almost anything, but one of the most comprehensive ones is, according to the Stanford Encyclopedia of Philosophy, by William Frankena:

**Objective-list theories** are a class of axiologies that define a list of objective values and disvalues.

> One of the most comprehensive lists of intrinsic goods that anyone has suggested is that given by William Frankena

> (Frankena 1973, pp. 87–88): life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc. (Zimmerman and Bradley, 2019)

Objective-list theories can be very different from one another, and most objective lists do not have as many items as Frankena's. They all have one thing in common, though: the items in the objective list are, as the same already suggests, objective. That means that if, say, beauty is on the objective list, then beauty is intrinsically valuable, whether it is actually important to me or not. Even if nobody is there who appreciates it, or if nobody is there to see it at all, it still is valuable.

We will not dive any deeper into objective-list theories, as hedonism and preference theory are our go-to axiologies, and we will only treat objective lists as a kind of backup theory.

So, in summary:

| **axiology** | **intrinsic value** | **intrinsic disvalue** | **intuition** |
|---|---|---|---|
| **Hedonism** | pleasure | pain | How well is someone feeling? |
| **Preference Theory** | preference satisfaction | preference frustration | How well are someone's preferences satisfied? |
| **Objective List Theories** | objective goods | objective bads | To what extend does someone have the objective goods/bads from the list? |

STOP AND THINK

Which axiology do you find most intuitive and why? What is it that makes humans matter in your favourite axiology?

Now, we can try to answer our question from the beginning: what makes human lives matter? According to hedonism and preference theory, a human life usually is extrinsically valuable. This is because we only can experience pleasure and desire satisfaction if we are alive. Most humans would presumably miss out on much pleasure and preference satisfaction if they died an early death (though it would also spare us some pain and preference frustration). Also, it usually is a pain to die and it would frustrate the preference to continue with our lives, which is a popular one. Thus, according to hedonism and preference theory, our lives usually are extrinsically valuable *to us* – and this is not even taking into account the pain that those who are close to us suffer if we pass away or the preferences of others that we had satisfied in others over the course of our lives had we not died. So, our lives are also extrinsically valuable *to others*.

If you employ an objective-list theory, life might even be intrinsically valuable if it is an item on the list. If it is not an item on the list, life most likely is extrinsically valuable for similar reasons as above.

One important side note before we close our journey into axiologies: intrinsic value is not better or worse than extrinsic value. Both are value. It does not diminish the value of a human life to say that it is 'only' extrinsically valuable. In fact, the extrinsic value that your life has is much larger than many intrinsic values.

### Axiologies in Consequentialism

You might now wonder why we talked about axiologies so extensively in a chapter on consequentialism. The answer is simple: consequentialism cares about intrinsic value, or, more specifically, about *all* the intrinsic value and disvalue. It is how it determines someone's level of wellbeing. Recall our consequentialistic framework:

> *An agent A ought to perform one of the right actions. An action $\varphi$ is a right action if and only if the* relevant qualities *from the* consequences *of $\varphi$ fulfil a* specific condition.

The relevant qualities of the consequences in our consequentialistic framework is roughly everything of intrinsic value or of disvalue. (Why only roughly? Well, some consequentialist theories mainly care about sufficiently close value and disvalue and, e.g., disregard value that is generated in, say, ten million years. However, we will assume that all intrinsic value and disvalue ever counts.) In conclusion: the relevant qualities that consequentialism is looking at is all the intrinsic value and disvalue.

Often, people also use the words "benefit" and "harm" in the context of consequentialism – especially if they employ hedonism or preference theory. Roughly, someone has a benefit if and only if something good in the sense of our axiology happens to them, and they are harmed if and only something bad in the sense of our axiology happens to them. If we, for example, employ hedonism, then someone is benefited if they experience pleasure, and are harmed if they experience pain.

Roughly: someone has a **benefit** iff something good in the sense of our axiology happens to them, and they have a **harm** iff something bad in the sense of our axiology happens to them.

Most people think that benefit and harm can be aggregated. Usually, harm is simply subtracted from benefit. The result is called "utility". Construed in this way, utility can have a positive sign or a negative one, depending on whether we put more harm or more benefit into the equation. Consequentalism that uses utility will be called "utilitarianism".

**Utility** is the result of a suitable aggregation of benefit and harm.

**Utilitarianism** is the most famous form of consequentialism and it uses utility as its relevant quality. (Caution: This term is often used more narrowly than that in the literature.)

For all of the following, we will assume that we can assign an overall utility to each relevant being at all times. A relevant being is everything that can be benefited or harmed in the sense of our axiology – most prominently humans, but depending on the axiology also some or even all animals. If we take the utility of someone over the whole course of their live, we get their level of wellbeing. In conclusion: the relevant qualities that utilitarianism is looking at is the overall utility of each relevant being.

That was a lot to take in.

## 2.3 Misunderstandings about Consequentialism

TODO arg1

## 2.4 Consequentialistic Theories and their Application in Practice

TODO Examples (look at solutions to exercises), Dos and Donts, ...

Trillions of beings? Level of wellbeing over the course of their WHOLE lives?!

Lacking numbers/don't guess numbers

How to evaluate objective cons?

Too little information: case analysis (aka case distinction in the videos)

Abstracting away from specifics of a theory (e.g. condition, axiology), argue with a pre-theoretic understanding of a good consequence.

Frequent mistake: its about the maximum benefit minus harm, not about whether benefit is greater than harm

## 2.5 Problems of Consequentialism

TODO arg1

# 3 Deontological Theories

One of consequentialism's slogans is "the end justifies the means". While many see this as a feature and not as a bug, others have trouble with this idea. This usually is motivated by the intuition that there can be things that just are forbidden, almost regardless of their consequences. Using the large man to stop the train, for example, is seen by many as wrong – even if we can save five other lives as a result.

The fundamental idea of deontological theories is that an action is permissible if and only if it adheres to certain universal principles. In the following, we will look at two (slightly simplified) deontological theories: Immanuel Kant's moral philosophy and Thomas Scanlon's contractualism.

## 3.1 Kant's moral philosophy

TODO

## 3.2 Scanlon's contractualism

> An act is wrong if[f] its performance under the circumstances would be disallowed by [...all sets] of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement. (Scanlon, 1998, p. 153)

The fundamental idea is that there are some sets of principles, i.e. rules, which no one could reasonably reject and that these sets of principles govern how we should behave: An act is right if and

only if it is allowed by at least one such set (given the circumstances), and disallowed if and only if *all* of those sets forbid it (given the circumstances).

**What to do given sets that nobody can reasonably reject?** To look at this in more detail, we will use a variant of one of Scanlon's standard examples, namely left- and right-hand traffic.

> **Scenario 13:** ***Traffic Code***
>
> Selda was recently put in charge of a previously uninhabited island called High Rool. She recently built some streets there and now has to implement a traffic code now that regulates how drivers have to behave on the streets of High Rool. There are no conventions yet for driving on High Rool. She has to decide whether to introduce left-hand traffic or right-hand traffic.

Assume that there are sets of principles that no one can reasonably reject, which allow Selda to introduce left-hand driving and that there also are sets which allow her to introduce right-hand driving. Also assume that there is no set that allows her to introduce both right-hand and left-hand driving at the same time, to leave the traffic unregulated, or to introduce some third system (like diving on the left only on weekends). What is Selda then allowed to do and why?

> STOP AND THINK
>
> Try to figure out what Selma is allowed to do with the given information and the quote from page 39.

Even though not all sets that no one can reasonably reject allow for the introduction of left-hand traffic, there still are some that do. So, it would be right for her to introduce left-hand traffic. The same holds for right-hand traffic, too, so it would be also right for her to introduce right-hand driving. But as every set disallows introducing both at the same time, leaving the traffic unregulated or implementing a third system, it would be wrong

for her to do any of these. Overall, she can choose between right- and left-hand traffic.

So far, this is easy and straightforward. But how to tell which sets of principles can be *reasonably rejected* by someone?

**Which sets can be reasonably rejected by someone?** I can only reject a principle (or a set of principles) that imposes a burden on me, e.g. that I will suffer if I and others followed this principle (or set of principles). However, this alone does not yet suffice. I can only reasonably reject a principle (or a set of principles) if and only if my rejecting it does not impose a much greater burden on someone else. This is called the Greater Burden Principle:

> It would be unreasonable [...] to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others. (Scanlon, 1998, p. 111)

Let us unpack this by looking at another example:

> **Scenario 14:** ***Accessible Parking I***
>
> Karen wants to go grocery shopping at a supermarket. The parking lot is quite full and she is too lazy to walk all the way from the back of the parking lot to the store, so she parks in last available parking space reserved for people with disabilities, thereby taking it away from Jen, who has to use a wheelchair. As a result, Jen has to park in a regular parking spot which comes with great inconveniences for her.

Is there a set that allows Karen to park in the reserved parking spots? Suppose that there was such a set and let's call that set S. Let us take a look at the different burdens from S or if S was rejected by someone:

| | **Karen** | **Jen** | **others** |
|---|---|---|---|
| burden through S | no burden | **large burden** **(great inconveniences)** | no others involved |
| burden through rejection of S | **small burden** **(has to walk a bit more)** | no burden | no others involved |

STOP AND THINK

Try to figure out if the set S can be reasonably rejected by someone using the given information from this section and the quote from page 39.

If the principles from S held, Jen would have a burden, because she would have massive inconveniences. If she rejected S, she would have a smaller burden, or more specifically: no burden in this case. This makes her a candidate for someone who can reasonably reject S. But she can only do so if nobody would have a considerably greater burden than she has now in case that Jen rejects S. Is there someone with such a greater burden? No, there is not. For sure, Karen does have a burden if Jen rejects S – namely that she has to walk a bit more and cannot be lazy – but this burden is smaller than Jen's current burden. (Both relevant burdens are set in bold in the table.) So, Jen can reasonably reject S. Because we constructed S to be a stand-in for any arbitrary set that allows Karen to park in the reserved spot in this specific case, we have shown that Jen can reasonably reject *all* sets of principles that allow Karen to do so in this case. We thus have shown that it is not allowed for Karen to park there in this case.

We now have a first recipe for deciding whether a set of principles S can be reasonably rejected. Make a table as above with all individuals who are relevant in the situation; find out who has the largest burden through S – in our case: who has the greatest burden in the top row – and who has the largest burden if S was rejected – in our case: who has the greatest burden in the bottom row. If the largest burden through S is considerably larger than the largest burden through the rejection of S, then S can be reasonably rejected. Otherwise, nobody can reasonably reject S.

For the sake of completeness, let us now turn the above example around and look at an example of set of principles that cannot be reasonably rejected.

**Scenario 15:** ***Accessible Parking II***

Jen wants to go grocery shopping at a supermarket. The parking lot is quite full. Jen has to use a wheelchair and thus parks in the last available parking space reserved for people with disabilities. Just as she wants to get out of her car, Karen pulls up besides her. She asks Jen to clear the parking spot, since Karen herself wants to park there. Karen admits that she does not have any disability but that she is just too lazy to walk. Jen refuses to give up the parking space.

STOP AND THINK

Try to figure out if the set S can be reasonably rejected by someone using the given information from this section and the quote from page 39. You will have to first figure out who has which burden through a set that allows Jen to park in the reserved parking spot, and through the rejection of this set.

Is there a set that allows Jen to park in the reserved parking spot, even though Karen demands the spot for herself? Suppose that there was such a set and let's call that set S*. Let us take a look at the different burdens from S* or if S* was rejected by someone:

| | **Karen** | **Jen** | **others** |
|---|---|---|---|
| burden through S* | **small burden** **(has to walk a bit more)** | no burden | no others involved |
| burden through rejection of S* | no burden | **large burden** **(great inconveniences)** | no others involved |

S* places a small burden on Karen that she would not have if she rejected S*. This makes her a candidate for someone who can

reject S*, but – you probably already know where this is going – she cannot *reasonably* do so. Jen would have a much greater burden though the rejection of S* than Karen has through S*. So, Karen cannot reasonably reject S*. Therefore, Jen is allowed to refuse to give up her parking spot to Karen.

**Can burden be accumulated across difference people?** No, burdens cannot be accumulated across different people. It is a key feature of Scanlon's contractualism that there is no interpersonal aggregation of burden. Scanlon proposes a thought experiment to make this clear:

> **Scenario 16:** ***The World Cup Finale***
>
> Jones has fallen into the electrics of a TV transmitting station. He is receiving extremely painful electroshocks, but will be fine once rescued. If his colleagues rescue him immediately, the transmission of the final 30 minutes of the football World Cup final, which several million people enjoy, will not be transmitted and nobody will be able to see it. Alternatively, Jones has to suffer for another 30 minutes.

STOP AND THINK

Try to figure out if Jones has to be rescued immediately or only after the game using the information from this paragraph and the quote on page 39.

According to Scanlon's conractualism, it is the right to rescue Jones immediately, even though several million fans will then be unable to see the finale. If there was a principle that allowed to wait until the match is over, Jones could reasonably reject it, because being in severe pain for half an hour is a much greater burden than not seeing the World Cup finale. It does not matter how much all the fans (together) suffer because they cannot see the finale, as long as none of them suffers considerably more than Jones.

Following this line of thought, it would yield in the weird result that it does not matter whether one person starves to death or

whether a million people starve to death, since all of them would bear the same, heavy burden. This is very counterintuitive. Scanlon see this problem, too, and proposes a solution.

**What if we have equal burdens?** Scanlon proposes a *weighted lottery* for cases in which equal burdens have to be weighed against each other. We will illustrate that by using variations of his standard example:

> **Scenario 17: *Rising Tide I***
>
> Alice and Bob both went for a walk on the beach independently of each other. Eventually, the tide came in and trapped them on separate sandbanks in very remote areas of the beach. Neither of them can swim, so they will quickly drown once the top of the tide is reached and the sandbanks are completely subsumed under water. Frank, a fisherman who does not know Alice or Bob, is on his boat and notices the two at the very last moment. He knows that he, unfortunately, does not have enough time left to save both of them but that he has only enough time to get either Alice or Bob to safety.

STOP AND THINK

Can Alice reasonably reject a set of principles that allows Frank to rescue Bob straight away? And can Bob reasonably reject a set of principles that allowes Frank to rescue Alice straight away?

Suppose that there was a set of principles $S_A$ that allowed Frank to save Alice and let Bob drown. We would get the following burdens here:

| | **Alice** | **Bob** | **Frank** | **others** |
|---|---|---|---|---|
| burden through $S_A$ | no burden (can be rescued) | **extreme burden (death)** | no burden | no others involved |
| burden through rejection of $S_A$ | **extreme burden (death)** | no burden (can be rescued) | no burden | no others involved |

The set of principles $S_A$ places a huge burden on Bob that he would not have if he rejected $S_A$. His rejection of $S_A$ would place a huge burden on Alice, but this burden is not considerably larger than his. So, Bob *can* reasonably reject $S_A$! So, all sets that allow Frank to save Alice can be reasonably rejected by Bob and thus, Frank is not allowed to save Alice off-handedly.

Alice, however, can reason in the exact same way: Suppose that there was a set of principles $S_B$ that allowed Frank to save Bob and let Alice drown. The burdens are just like they were before, but now in reverse:

| | **Alice** | **Bob** | **Frank** | **others** |
|---|---|---|---|---|
| burden through $S_B$ | **extreme burden (death)** | no burden (can be rescued) | no burden | no others involved |
| burden through rejection of $S_B$ | no burden (can be rescued) | **extreme burden (death)** | no burden | no others involved |

$S_B$ places a huge burden on Alice that she would not have if she rejected $S_B$. Her rejection of $S_B$ would place a huge burden on Bob, but this burden is not considerably larger than hers. So, Alice *can* reasonably reject $S_B$, just as Bob could reasonably reject $S_A$! So, all sets that allow Frank to save Bob can be reasonably rejected by Alice and thus, Frank is also not allowed to save Bob off-handedly. It seems that Frank is now in a real dilemma, but there is a way out for Frank.

A Scanlonian would say that it is the right thing to toss a fair coin in this case, and thus save Alice with a probability of 0.5, and save Bob with an equal probability. This puts Alice's and Bob's competing reasons in a balance and gives both of them a chance to avoid their dire outcome. None of both could reasonably reject a set of principles that allows Frank to toss a coin. (A good argument for that is tricky, so we will just accept that lotteries are our go-to way to resolve situations where the largest burdens are equal and thus have to be weighed against each other.)

Now that we established what happens in the two-persons case, we should take a look at cases with more than two people.

> **Scenario 18:** ***Rising Tide II***
>
> Alice is trapped on one sandbank, while Bob and Charlie are trapped on another sandbank. Everybody who is not rescued in time, will drown. When fisherman Frank becomes aware of the situation, he knows that he has only enough time to rescue one of the two parties, i.e. he can either go to one sandbank and rescue Alice, or he can go to the other sandbank and rescue Bob and Charlie, but there is no way to rescue all three.

By just tossing a fair coin again, Frank would ignore the fact that there are more people on one sandbank than on the other. Instead, you should, in Scanlonian reasoning, save Alice with a probability of $\frac{1}{3}$ and Bob and Charlie with a probability of $\frac{2}{3}$. None of the three can reasonably reject a principle that allows one to decide that way. Likewise, if Frank had to choose between saving one person on one sandbank, and 99 people on the other sandbank, he should save the single person with a probability of a 1 percent, and the 99 people with a probability of 99 percent.

**Some problems** However small, there still is a chance that Frank has go to the sandbank with fewer people on it. This can be regarded as a major problem of Scanlonialism: It could be right (even if this is very improbable) to save one person instead of a thousand, even when there is no morally relevant difference between any of these people. To many, this is a very counterintuitive result. Many would say that it should always be obligatory to save the thousand instead of one. One could argue that a moral theory should never – not even in one of 1001 cases – allow to sacrifice a thousand people to rescue one. (At least given that all of these people are more or less similar. The matter would be difference if, for example, the single person had just found to cure to cancer which would be lost if she died now.)

This is a symptom of the fact that Scanlon's theory does not allow to accumulate burden, which is objectionable for another reason: The burdens of an arbitrarily large number of people can be out-

weighed by the burden of a single individual. If the number of people, on whom a certain burden is placed, is large enough, it seems very counterintuitive that a single individual with a certain, larger burden should outweigh that. What was still intuitive in the case of Jones and the World Cup finale, can get counterintuitive rather quickly.

**Scenario 19:** ***Leg Amputations***

Suppose that we select a million random people in the world for whom loosing a leg would be really bad but not catastrophic. None of them has a profession or hobby that requires them having both legs, and neither of them would suffer long-term medical problems. If you push the blue button in front of you, they will all lose a leg instantly. Additionally, we select Tim. For him, it would already be terrible to lose one leg, but it would be considerably worse to lose two legs. It will be much worse for him to lose two legs than it will be for each of the million to lose one leg. If you push the red button in front of you, Tim will lose two legs, but all the other legs of all the other people will be spared. If you push neither button or both buttons, the earth will explode. We assume that those who lose a leg will be impacted the most, and that the impact on everybody else, including those close to them, will be less.

STOP AND THINK

Is it right to push the red button, the blue button or neither button/both buttons according to Scanlon?

It, of course, is out of question to press neither button. Everybody on planet earth, including Tim, could reasonably reject any set of principles that allows you to push neither button. But which button should you then push?

Suppose that there was a set of principles $S_{blue}$ that allowed you to push the blue button, i.e. that allows to make a million people lose one leg each. Also supose that there was a set of principles $S_{red}$ that allowed you to push the red button, i.e. to make Tim lose both legs. As always, we take a look at the burdens:

| | **Tim** | **each of the other randomly selected people** | **you and others** |
|---|---|---|---|
| burden through $S_{blue}$ | no burden (loses no leg) | **large burden (loses one leg)** | smaller burden, if any |
| burden through rejection of $S_{blue}$ | **larger burden (loses both legs)** | no burden (loses no leg) | smaller burden, if any |
| burden through $S_{red}$ | **larger burden (loses both legs)** | no burden (loses no leg) | smaller burden, if any |
| burden through rejection of $S_{red}$ | no burden (loses no leg) | **large burden (loses one leg)** | smaller burden, if any |

As you can see, none of the one million people can reject any set of principles that allows to push the blue button – but Tim can reject any principle that allows you to push the red button. So, according to Scanlon's contractualism, you are obliged to sacrifice one million legs in order to save the two legs of Tim. This is perceived as very counterintuitive by many people.

# Bibliography

Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.

Zimmerman, M. J. and Bradley, B. (2019). Intrinsic vs. Extrinsic Value. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition.