Rational traps

Version: 5

We're going to do two things in this unit.

First, we'll discuss some ways our ordinary sense of how to make good decisions goes very wrong.

Second, we'll discuss decisions where the costs and benefits create a trap for rational decision makers. If we don't recognize the trap and do something to break out of it, acting rationally will make everyone — including us— worse off.

It will be really important to resist the temptation to think that the people in the traps are dumb, don't understand the situation, or are otherwise irrational. That's the crazy thing about them. These traps only catch rational decision makers.

Throughout this unit (and this unit is unique in this regard) I'll frequently illustrate things with numbers to help clarify and make concrete what's going on. If you, like me, had terrible math teachers who wrongly convinced you that you're bad at math, don't let that turn you off.[1] You can reason through everything here without the numbers.
There are better and worse ways to make most decisions. Consulting a magic 8 ball for deciding whether to move to a new city for a job is not likely to ensure your future happiness. Once you agree that some ways of making choices are better, you are on the way agreeing that there's an ideal of a rational decision maker. The ideally rational decision-maker always makes choices in the best possible way.

Beyond that, things get controversial. For example, there might be multiple, mutually exclusive, ways of being a rational decision maker. It's also clear that there's an emotional component to good decision-making,

[1] It took until the past few years to realize that I'm actually pretty good at math, just not arithmetic. That's probably true of you too!

but it's hard to pin down how that fits with more cold-blooded comparison of numbers. We don't need to worry about these issues here.

No human being is always perfectly rational. Your brain is both the hardest working and laziest organ in your body; brains love shortcuts, as the massive psychological literature on heuristics and biases shows. Often those shortcuts divert us from making good decisions. But the fact that we tend toward irrationality doesn't show we shouldn't try to reason well. Indeed, the objection that we're hopelessly irrational requires an ideal of rationality —a standard that we fail to meet. Our topic will be certain aspects of that standard.

For our purposes, let's start with a very thin idea about rationality that most writers can accept (they disagree about what needs to be added):

> (R1) You are rational only if there is consistency between your desires, beliefs, and actions.

If you want a taco and (correctly) believe that tacos are the best food, but order a burrito, you are being irrational. If your beliefs are contradictory — if you believe that 5 is greater than 2 and 2 is greater than 5— you are being irrational. Thus we can say

> (R2) An agent is rational only if she chooses the option which she believes will lead to the best expected outcome.

We'll come back to what I mean by 'expected outcome' in a moment (Win in Vegas and Vote).
When you're trying to get the hang of something, its often helpful to think about its contrary. Thus let's talk about some forms of irrationality.
One common way of being irrational involves mistakes about our own preferences.

Suppose I'm engrossed in conversation at the bar. When the bartender asks what I want, I panic, see someone drinking a beer and say 'beer' even though, if I had thought about it, I would've preferred whisky. Since my action (the order) is inconsistent with my real desire (whisky), I've acted irrationally.[2]

Rationality also requires consistency among your preferences. Suppose I have the following preferences: Whisky > Wine > Beer. That is, other things being equal, if the bar has a decent whisky, I will order the whisky. If they have no good whisky, I will order wine. If they have no good whisky or wine, I will order beer. These preferences are consistent.

But we can have inconsistent preferences. There's lots of ways this can happen. One easy case involves intransitive preferences. [3]

Suppose I have the inconsistent preferences: Whisky > Wine > Beer > Whisky. This is bad because I can be money-pumped. It works like this: I order a beer. When the bartender brings it, I notice that they have a wine I like. She offers to swap the beer for a glass of the wine if I give her $1. Since I prefer wine to beer, I agree. When the wine arrives, she offers to swap the wine for a whisky if I give her $1. Since I prefer whisky to wine, I agree. She brings the whisky and offers to swap the whisky for a beer, if I give her $1. Now we're right back at the beginning and my wallet is $3 lighter. Of course, she can then offer wine, and we'll keep going around the circle until I'm out of money.

We'll begin building toward the tragedy of the commons by discussing 2 seemingly unconnected topics:

(1) How to gamble

[2] This example also shows that irrationality doesn't have to be a big deal.

[3] My dissertation advisor in graduate school was (in)famous for arguments that intransitive preferences can sometimes be rational. He convinced me. We don't need to worry about those special cases here.

and

(2) How to make decisions about participating in groups.

Neither of these is strictly necessary for a basic understanding the tragedy of the commons. We will need them when our understanding gets a bit more sophisticated. However, my main motivation for discussing these (in addition to their intrinsic importance) is to help introduce some challenges lurking beneath the surface of seemingly simple decisions.
We'll start with how to rationally gamble, or, more broadly, make decisions under uncertainty. This is both useful in Vegas and illustrates a case in which rational decisions require a kind of fiction —you have to focus on a pretend value rather than the actual amount that will be in your pocket after the bet.

Suppose I offer you the following bet:

Bet 1: We'll flip a fair coin. If it comes up heads, I give you $100. If it comes up tails you give me $5. (NB, for the non-accountants in the room: the parentheses around the value are like a minus sign: ($100) = -$100)

Figure 1

Outcome	Heads	Tails
Payoff	$ 100	($5)

Would you take this bet? That's not really the right question. We all have different risk tolerances. The better question is whether this is a good bet?

What's a good bet? Obviously, if you'll definitely win it's a good bet. But we're interested in situations where the outcome is genuinely uncertain.

Let's say that a good bet is one where you rationally expect to end up better off if you take it.

Is Bet 1 a good bet? Definitely yes. But why? Take a moment to think about your answer before plowing ahead.

To see why, consider a more complex bet. Suppose there are 4 things which can happen: A, B, C, D. I offer you Bet 2:

Figure 2

Outcome	A	B	C	D
Payoff	$ 1,000,000	($100)	($10)	$5

Is Bet 2 a good bet? More importantly, if you had to take either Bet 1 or Bet 2, which is better?

Your answer should be "I don't have enough information." There's a big difference between the two bets, but I set the examples up so that you won't see it at first. There's something you know about the first bet that you don't know about the second. Think about it before reading on.

In Bet 1, I purposely started with flipping a fair coin. You know that the chance of winning is 50%. You probably thus automatically assumed that in the second bet, all 4 outcomes were equally likely to occur. But I tricked you. The actual probabilities are:

Outcome	A	B	C	D
Payoff	$ 1,000,000	($100)	($10)	$5
Probability	0.0001%	98%	1%	0.9999%

Okay. But what should you do with the new information about probabilities?

The answer is that you need to weight each payoff by how likely it is to occur. That is, you need to figure out the expected value of each outcome:

Expected Value = Probability * Actual Value

Here are the expected values for Bet 1:

Outcome	Heads	Tails
Payoff	$ 100	($5)
Probability	50%	50%
Expected	$ 50.00	$ (2.50)

To determine whether taking the bet is a good deal, we add together the expected values for all the outcomes. If the sum is greater than 0, we expect to win (i.e., a good bet). If it is less than 0, we expect to lose (i.e., a bad bet).

The expected value of taking Bet 1 is: $47.50. It's a good bet.

The really important thing is that when you are trying to decide whether to take Bet 1, you know that when the coin lands, your wallet will either include a shiny new portrait of Franklin or be missing a portrait of Lincoln. But you must ignore those numbers. You should be thinking only about $47.50.

To see why that matters, compare the expected values of Bet 1 and Bet 2.

For Bet 2, the expected values of the outcomes are

Outcome	A	B	C	D

Payoff	$ 1,000,000	($100)	($10)	$5
Probability	0.0001%	98%	1%	0.9999%
Expected	$ 1.00	$ (98.00)	$ (0.10)	$ 0.05

Adding together all the expected values of the outcomes gives us an overall expected value of -$97.05. Thus if you take Bet 2, you expect to lose -$97.05.

If I give you the choice between receiving $47.50 or giving me $97.05, you will choose the former. Thus Bet 1 is better than Bet 2.

But, but, but, but, but….. In Bet 2, I could win a million dollars! In Bet 1, the most I can win is $100. What about that?

That temptation is normal. It's how I set the example up to initially trick you. But you must resist it. To make a rational decision you have to ignore how much money will actually be in your pocket (um, suitcase) at the end. You have to focus on a pretend value —the payoff weighted by how likely it is to occur— because the actual amount will lead you astray.
Now for something completely different.

Let's consider how you should make decisions about a group activity where you make only a tiny contribution to the outcome. We will again see how rationality requires thinking in an unnatural way.

More concretely, we'll also find an argument for why you should vote. If you've ever thought about that, you've probably considered the fact that it is extremely unlikely that your vote will make a difference in the election.

You've probably also thought about the obvious response. If the fact that your vote is unlikely to make a difference shows that you shouldn't vote, it follows that no one's vote makes a difference and thus no one should vote.

Of course, there are problems with that response too. The argument seems to prove too much. No farmer can grow enough food to feed everyone. Indeed, if any given farmer stops farming, everything will be fine. Let's hope it doesn't follow that no one should be a farmer.

We're going to attack the problem from a very different direction.
The argument above that you shouldn't vote because your vote won't matter depends on the very reasonable-seeming principle:

> Share of the Total: In a collective act, each person produces her share of the total benefit / harm.[4]

Suppose your team closes a big deal. All 10 members contributed equally to the success. Your boss wants to reward the team with a $10,000 bonus. How much should she give each team member?

Intuitively, it would be unfair to give the whole $10,000 to one team member. Since everyone contributed equally, the fair thing to do is give each member $1,000. Similarly, if one member of the team did half of the work, she should get $5,000 and the other 9 members should split $5,000.

Make sense? That's the Share of the Total principle at work.

Let's take a new example to move from fairness to how you should make decisions about joining group efforts.

[4] In this section, I'm adapting an argument which appears in Derek Parfit's masterpiece Reasons and Persons

Suppose you're required to work on a team but you're allowed to choose which team. There are 3 teams with an open spot; each team is working on a deal which will yield a $100,000 bonus for the team. Let's assume that your only motivation is getting a better bonus (none of the projects is more interesting, no one you hate working with, et cetera). You'll choose the team where your contribution makes the biggest difference —you'll join the 5 person team rather than the 20 person team.

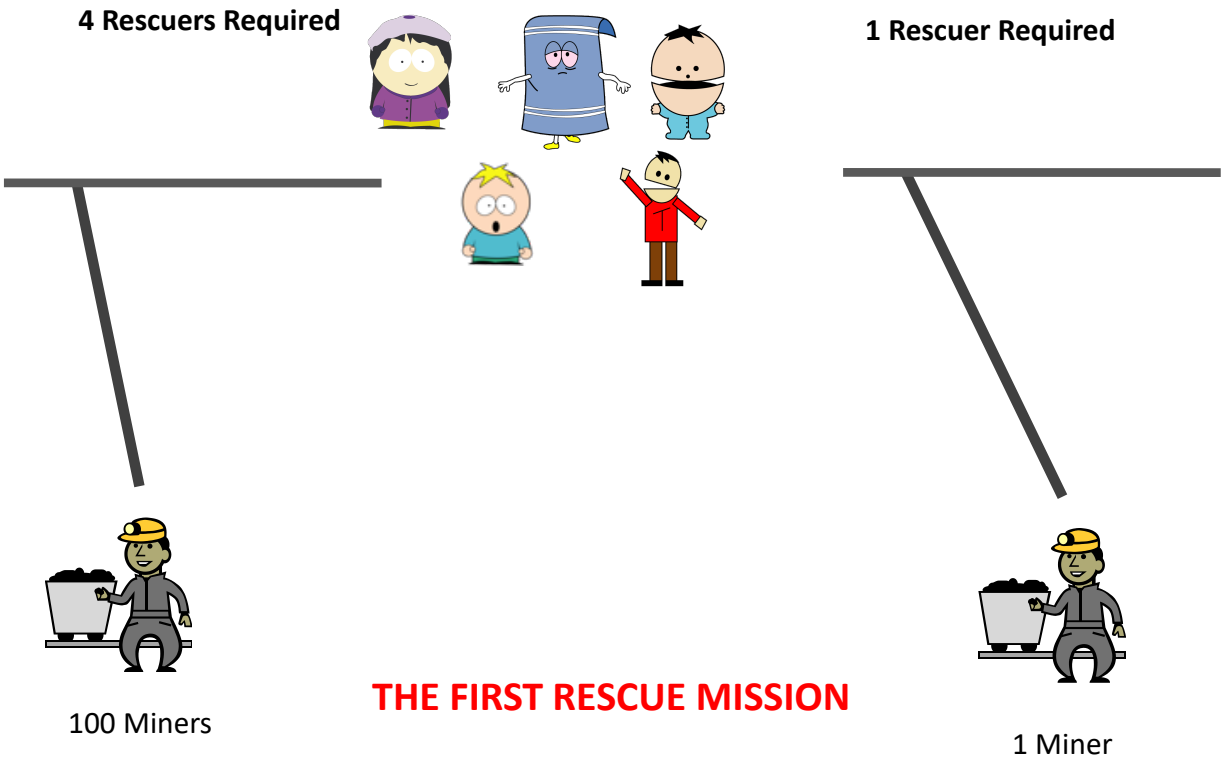
Thus we've seen that (given a lot of assumptions) the best thing for you to do depends on calculating how much you contribute to the total benefit. Hopefully, that makes the Share of the Total principle intuitive.

Unfortunately, the Share of the Total principle is false.
There has been a disaster. In 2 separate locations, miners are trapped underground. The mines are flooding; if they are not rescued, they will die.

Fortunately, the mines have old-fashioned rescue systems. Using a bunch of levers and pulleys, if enough rescuers stand on a platform above ground, the trapped miners can be raised to the surface.

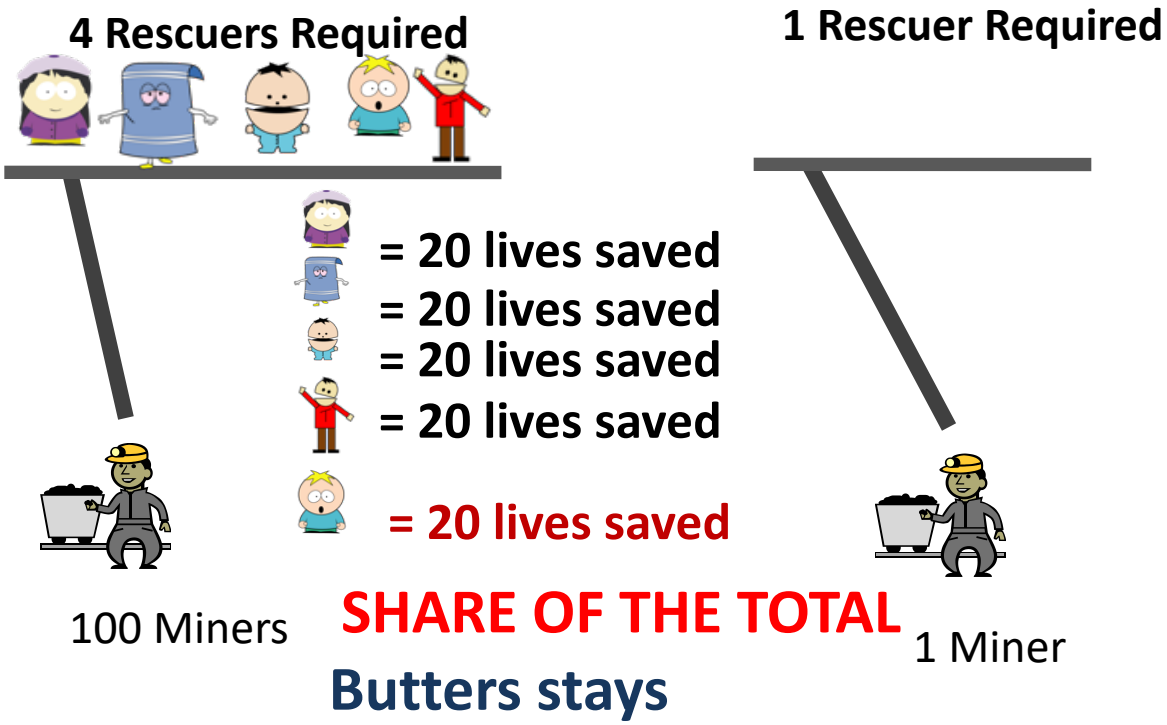
To make this easier, assume that everyone involved are strangers to each other, so there's nothing like friendship that matters. Similarly, all the trapped miners are strangers to you. There's nothing special about any of them. None of them are serial killers or work on a cure for cancer at night. They are all adults (no, wait for it, minor miners). Thus the only thing that matters for your decision is the number of lives you save.

The situation is basically:

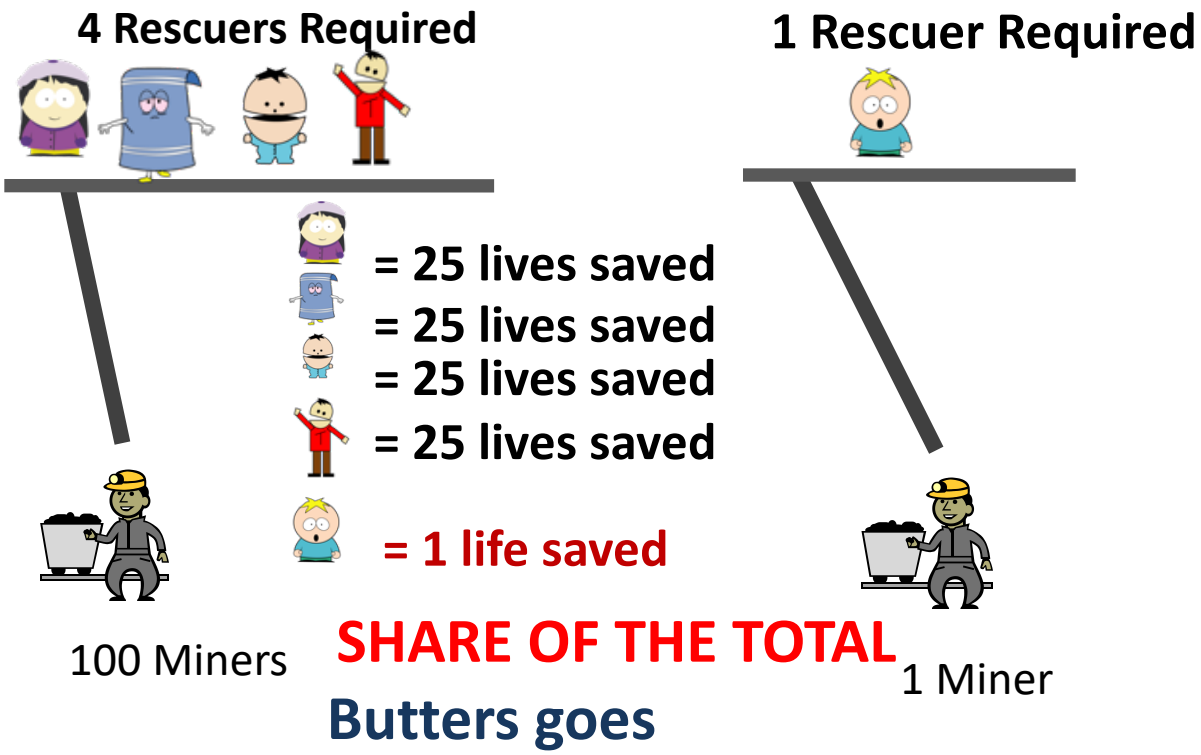


Let's suppose you, represented by Butters, arrive late to the scene. 4 people are just about to hop on the platform to save the 100 miners. What should you do?

If you join the other 4 rescuers, given the Share of the Total principle, each of you will save 20 lives.



Whereas, if you go over to the mine where 1 miner is trapped, you will save 1 life:



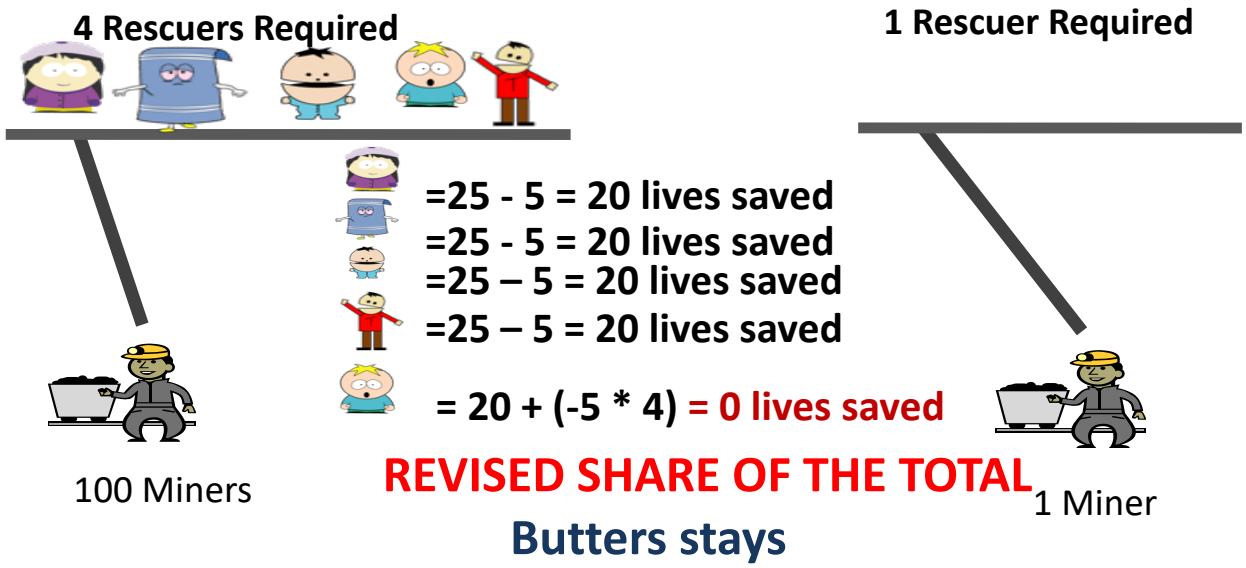
Remember, your goal is to do as much good as possible. Joining the other 4 rescuers means you save 20 lives versus saving 1 life on your own. 20 is greater than 1 so you should….Help save the 100 lives.

That doesn't seem right. The other 4 rescuers would've rescued the 100 without your help. You are superfluous in that rescue. But the 1 miner can only be saved by you. You are essential to that rescue.

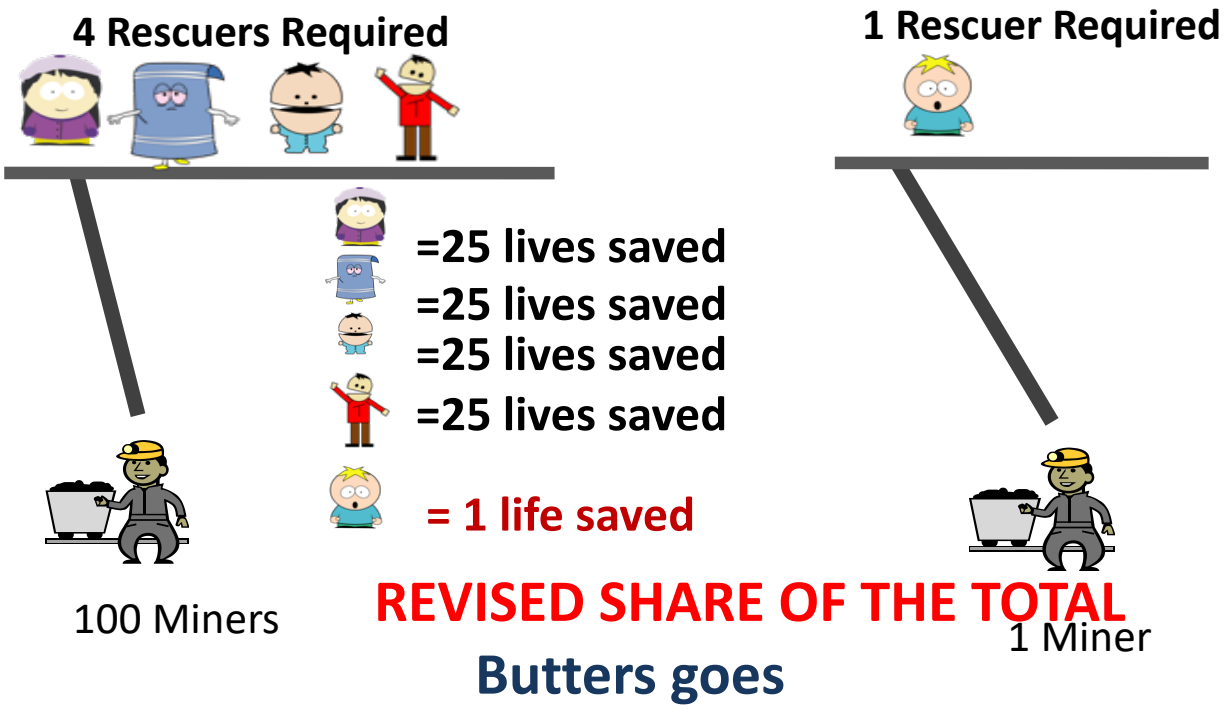
What happened? The Share of the Total principle led us astray. We got the wrong answer because it didn't take into account the fact that you aren't needed in the rescue of the 100 or that you are essential to saving the 1. We need to revise our principle.
Since the problem is that you were extra, we can revise the principle to correct for that:

> Revised Share of the Total: In a collective act, each person produces her share of the total benefit/ harm minus any reduction in the share of benefits/harms which her participation imposes on others' shares

With the Revised Share of the Total, we get the right answer in the first rescue mission. When Butters joins the other rescuers, each of them save 5 less lives than they would without him; him joining causes a 20 life reduction in total. Thus we subtract 20 from his share. If he joins the 4 rescuers, he saves 0 lives.



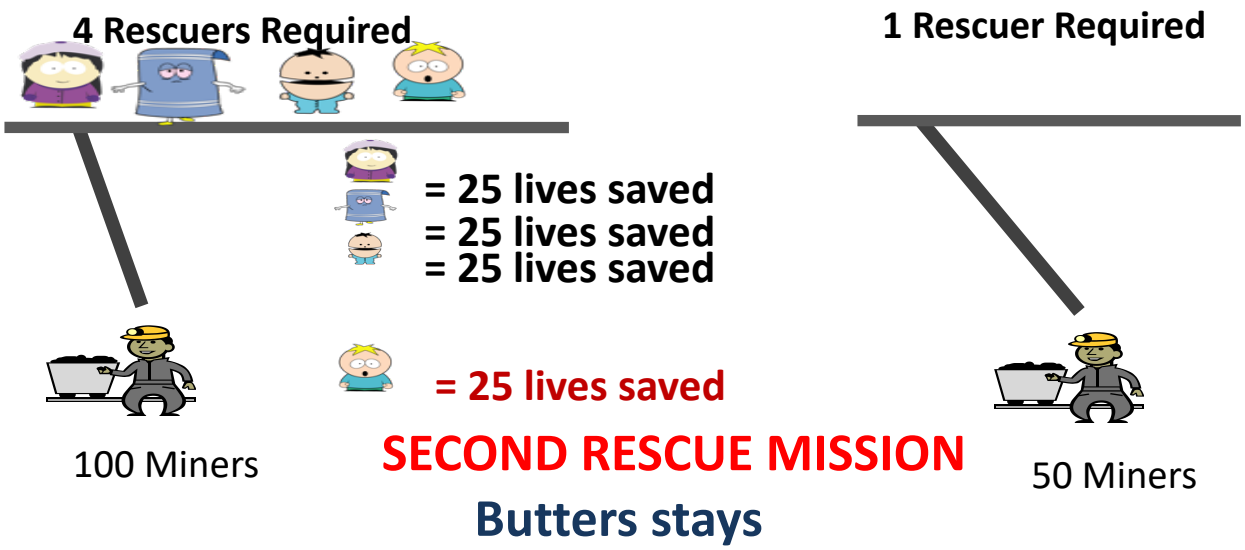
Whereas if he goes to the other mine, he can save 1 life



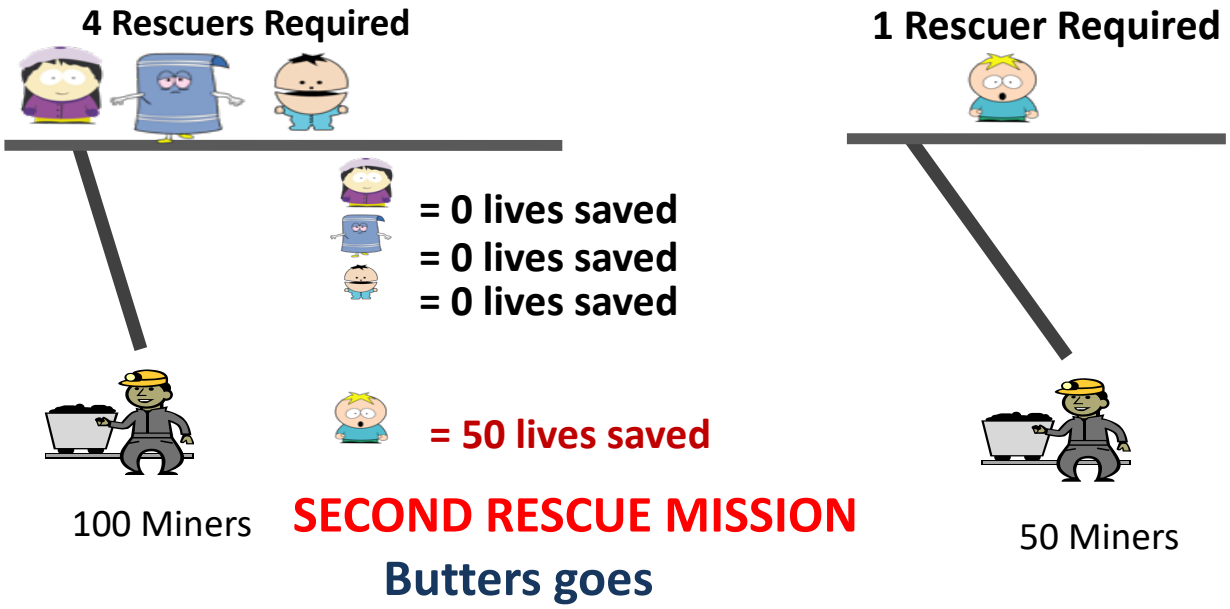
Since 1 is greater than 0, he should go save the 1 miner. That's the right answer. Hooray!

Uh oh. There's been another disaster. This time, Terrance has gone back to Canada, so there are only 4 potential rescuers.

If Butters joins the rest at the mine where 100 miners are trapped, he will save 25 lives.



If he goes to the other mine where 50 miners are trapped, he will save all 50.



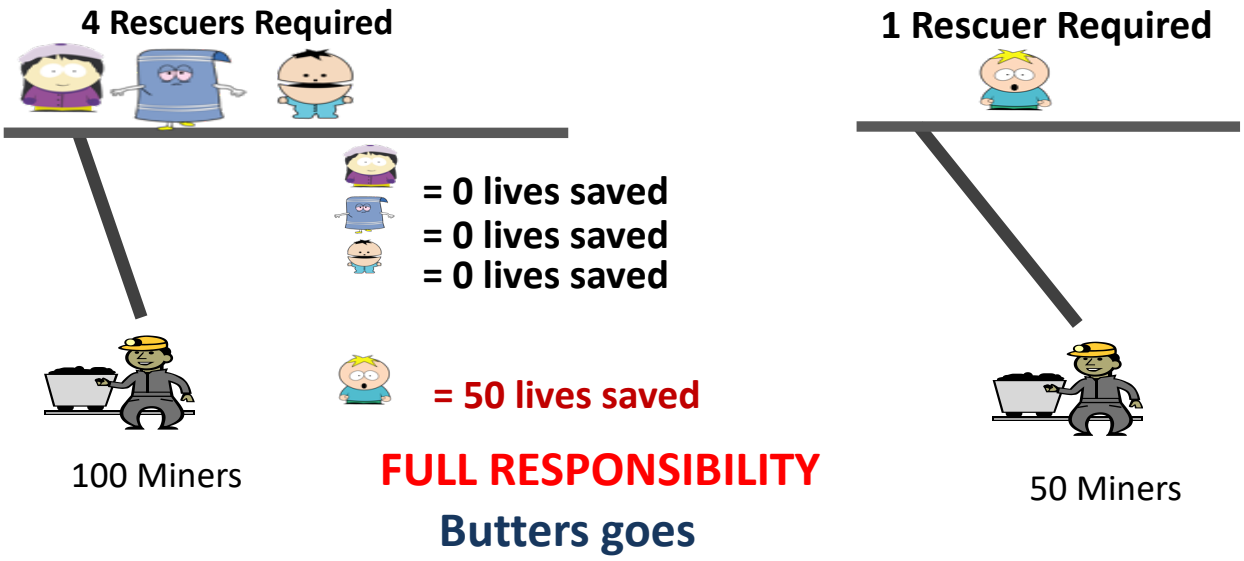
Since 50 is bigger than 25, he will go to the other mine. Thus 100 miners will die and 50 will be saved. Remember, his goal is to save the most lives possible, yet the Revised Share of the Total principle means he will only

save 50 rather than 100. That means we must reject the Revised Share of the Total principle.

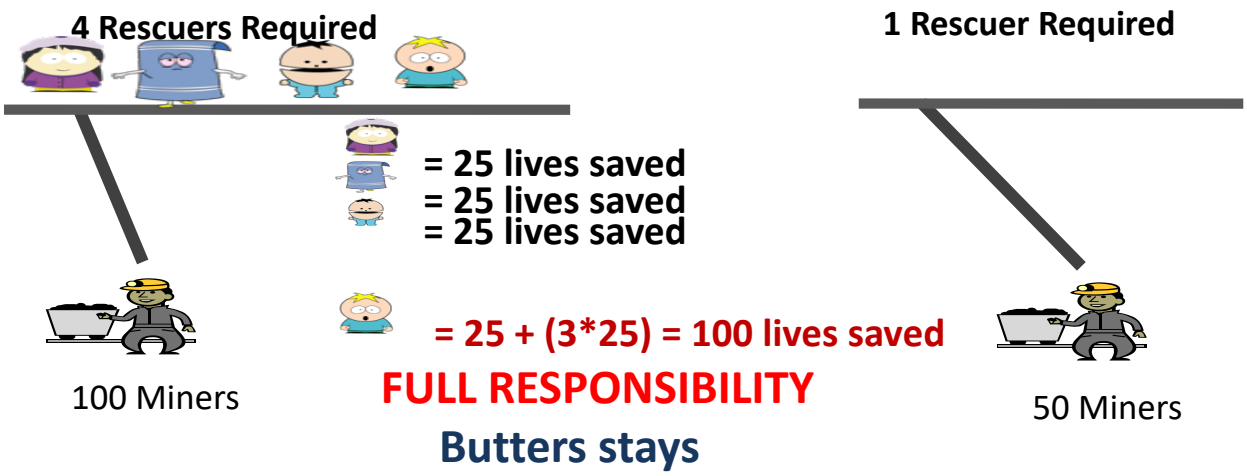
How do we fix this? We can make one more revision to the Share of the Total principle, which effectively replaces it with a new principle:

> Full Responsibility: In a collective act, each person produces her share of the total benefit/harm minus any reduction in the share of benefits/harms which her participation imposes on others' shares, and plus any increase to the shares of others which her participation causes.

Without Butters, the other team members each save 0 miners



But with him, each now saves 25 lives. Thus Butters gets to add 3*25 lives to his total:



That gives us the right answer.

Wait. What just happened?

The Full-Responsibility principle is a complicated way of saying that when you make decisions about participation in a collective activity, you must pretend that the entire outcome is solely up to you. That's more radical than it sounds.

We have rejected the intuitive Share of the Total view(s) and found our way to a much less intuitive claim. After all, when your candidate wins the White House, you don't think that you deserve to be ambassador to Tahiti, right? If 100 million people voted for your candidate, the Share of the Total principle says that you get 1/100million of the credit. That's not even free T-shirt level; definitely not an ambassadorship.

Nonetheless, if you want to avoid the errors caused by the Share of the Total and the Revised Share of the Total views, you need to make your decisions about when to participate in a group activity by pretending that you will be the only person responsible for the outcome.

We now know how to make decisions when we the outcome is not certain and how to determine when it is worth joining in group endeavors. We can now calculate whether it is worth voting.

To do this we need 3 pieces of information: the number of people affected by the choice, the average net benefit of our candidate winning, and the probability that your vote will make the difference.

US population	372,200,000
Average net benefit	$100
Total benefit	$37,220,000,000
Chance your vote makes difference	1 in 100 million
Costs to you	$ (75.00)
Costs to others of you voting	$ (25.00)
Total costs	$ (100.00)

I've chosen a small per person benefit of $100. You can think of that as either the average person getting $100 the candidate wins, or the average person being willing to pay $100 to turn back the clock if the other candidate wins. I've picked a pretty small number to suggest that the preferred candidate is a bit better, but not vastly superior.[5]

The chance that your vote makes the difference is a difficult empirical calculation. There are political scientists who can give you estimates. Note that it is much harder to calculate this for the American political system than elsewhere, given distorting institutions like the electoral college and the 2 party system. A probability of 1 in 100 million is likely on the low side, but it will work for illustration —feel free to do your own research and run this calculation (and let me know what you get!)

[5] To be clear, I'm just doing this for illustration. If you actually wanted to do this calculation, you would need empirical evidence for the value.

If you thought, "Wait. Why am I calculating the average benefit for everyone? I'm part of everyone, so shouldn't the benefit be $100?" you've fallen back into the Share of the Total principle. Remember, the Full-Responsibility principle tells us that we have to reason as though the full costs and benefits of an outcome are up to us, even though the involvement of others is required to create the outcome. Thus the relevant benefit is the average benefit to each American[6]

Drum roll please….

The expected value of you voting for the candidate you prefer is the expected benefit minus the expected cost.

The expected benefit is

$37,220,000,000 * 0.00000001 = $372.20

Assuming that costs are certain (e.g., you have to miss work). The expected cost is

($100) * 1 = ($100)

Thus the expected value of voting is

$372.20 - $100 = $272.20

Therefore you should vote even though the chance your vote makes a difference is tiny!

[6] Really we'd want this number to be the average benefit to the world population. I've restricted it to Americans since this is complicated enough already. But notice that if we use the world population (approx 7.7 billion), the average net benefit can be really tiny and we'll still end up with a net benefit to voting.

If you're eligible to vote in California elections and haven't registered to vote, here's how to do so: https://www.sos.ca.gov/elections/voter-registration/

Go do it now. We'll still be here.[7]

Rationality is pretty great right? Unfortunately, the same principles of rational decision-making which keep money in our wallets and us away from the Vegas roulette wheel malfunction spectacularly when the options' costs and benefits line up in a certain way. We'll talk about 2 related rational traps: the Prisoners' dilemma and the Tragedy of the commons.

This will get complicated. It will help to make some (big) simplifying assumptions that allow us to see the basic machinery at work. Once we've got that, we can start adding complications.

> Simplifying assumption #1: All examples occur in Adam's imaginary world. It's my world, my examples, my details. No adding details!

You will be extremely tempted to add details by saying things like "But what about x?" You will be right to be tempted because the things you have in mind will be the things which help break out of these problems in the real world. But to see why those details help fix the problem, we need to first get clear about the problem. Hence, my world; my examples.

> Simplifying assumption #2: The people in our examples only care about what is best for them.

[7] If this section piques your interest in ethical / technical, and political issues around voting, here are some additional resources:
https://plato.stanford.edu/entries/voting/
https://plato.stanford.edu/search/r?entry=/entries/voting-methods
https://plato.stanford.edu/search/r?entry=/entries/social-choice

I'm going to talk about people who care only about what's best for them. Selfish bastards, in other words.[8] Unfortunately, the rest of us who are partially self-interested and partially other-interested, don't get off the hook. The same traps arise for us too. Though it can be trickier to set out the argument. I'm guessing you'll agree that this is already complicated enough with the selfish folks.

> Simplifying assumption #3: Given a choice between two actions A and B, if the expected value[9] of A is greater than the expected value of B, a rational person will do A.

That just summarizes what we talked about above (Rationality)

Here we go….
Our plane has crashed on an uninhabited[10] desert island. You have taken up residence on the west side near some persimmon trees. I reside on the east side near some kiwi bushes.

One day, we spy each other across a meadow. You holding a persimmon; me holding a kiwi. Warily, we slowly approach each other.

[That music you hear is the beginning of Also Sprach Zarathustra Op.30[11]]

We are now within arms length of each other. I gesture at your persimmon and grunt, holding aloft my kiwi. My meaning is clear: while I like kiwis, I would rather have a persimmon.

---

[8] Being self-interested doesn't have to mean only caring about yourself. On many views, your concern for others counts as your self-interest. If you don't like to see your friends sad, your friends' happiness is part of your self-interest.
[9] Probability of outcome * (Benefits in outcome - Costs in outcome)
[10] No button needs pressing; no mysterious organization; no smoke monster. It's an uninhabited island, not purgatory. Oops. Spoilers….
[11] https://www.youtube.com/watch?v=YU88Awrb0fE

[The tympani are playing now: Dum Dum Dum Dum Duuuumm]

You gesture at my kiwi with your persimmon and grunt in return. You would rather have a kiwi than a persimmon.

[The music builds toward the first crescendo.]

We are about to discover market exchange. You will give me your persimmon. I will give you my kiwi. Both of us will be better off! Capitalism is about to be born!

I reach out my hand. You reach out yours. And then…..

[Record scratch]

Nothing happens. We stare at each other for awhile. Then we shrug and turn back to our respective sides of the island, me still with my kiwi, you still with your persimmon.

What went wrong? Why couldn't we trade?

The answer is that we're in a rational trap in the same family as the (in)famous prisoners' dilemma.

We'll come back to the island in a bit. First, let's go to jail.
Scarlet and Violet are career criminals. They've been brought together Reservoir Dogs style in a city they've never visited by a mysterious criminal mastermind to rob a bank. They will never see each other again afterwards.

Unfortunately, being unfamiliar with Los Angeles, their getaway plan involved speeding to freedom on the 405. (I'll wait for you to stop laughing) Still, being professionals, they managed to ditch the loot before the police lazily walk up to their car on the Sepulveda onramp and arrest them.

The district attorney gets the call that the police have arrested Scarlet and Violet and the news that while they didn't have enough evidence to convict them of armed robbery, there were plenty of lesser charges to work with.

Thinking back to her philosophy 305 class at CSUN, our DA smiles. "No problem." she says, "Put them in separate cells so they can't talk to each other." Hanging up the phone, she turns to her assistant "And put a bottle of champagne in the freezer for me, they're both going to testify against each other." "The freezer?" her assistant calls after her, "Are you sure not the fridge?"

Our DA arrives at the jail, offers both criminals the same deal, collects the signed confessions, returns to the office and enjoys a perfectly chilled glass of bubbly.[12]
To understand how her attention in philosophy class led so quickly to delicious victory, let's represent the situation for our criminals in Figure 3:

Figure 3

	Violet = Silent	**Violet = Rat**

[12] If you'd like a more sophisticated explanation of the problem (with real math), see https://plato.stanford.edu/entries/prisoner-dilemma/

Scarlet = Silent	Scarlet = Possession 5 years Violet = Possession 5 years	Scarlet = Armed robbery 10 years Violet = Trespassing 2 years
Scarlet = Rat	Scarlet = Trespassing 2 years Violet = Armed robbery 10 years	Scarlet = Armed robbery (lenient) 8 years Violet = Armed robbery (lenient) 8 years

(It's important to get the hang of this kind of diagram; you'll need it for the tragedy of the commons.)

Each of our suspects has exactly two choices: They can testify against the other person (i.e., rat) or they can stay silent. That means there are 4 possible outcomes: Both stay silent; Both rat; 1 rats the other stays silent and vice-versa.

Each row represents outcomes which Scarlet has control over. She decides which row they are in by choosing whether to rat. Each column represents

outcomes which Violet has control over. Inside each cell, I've listed the sentences they receive in that outcome.

Our former philosophy student DA meets with each of the suspects and offers them exactly the same deal. "Look." She says "I have more than enough evidence to get both of you for weapons possession. I can put you away for 5 years no matter what." This is the upper left box.

"But I'd like to get at least one of you for armed robbery. For that, I need one of you to testify against the other."

"So here's the deal. If you testify against your partner and they keep their mouth shut, I will charge you with mere trespassing. You'll be out in 2 years. They, on the other hand, will go away for 10 years." These outcomes are represented in the upper right and lower left boxes.

So far, it's unclear what the suspects should do if they want to get the least amount of time for themselves. Arguably, they should both keep their mouths shut. But now our DA slams the gate shut on the trap.

"Of course, if you both testify against each other, I can get both of you for armed robbery. But since you cooperated, I will ask the judge for a more lenient sentence. You'll be out in 8 years rather than 10."

Boom. If they are rational and the only considerations are the amount of time served (my world my example), they are trapped. Both will rat the other out.

That should be surprising. Indeed, it took several years in the decision-theory literature to sort this out. If both keep their mouths shut, they both get 5 years. Instead, they end up with 8.

If this doesn't seem weird to you, you're missing something. I've taught this for years and I still feel the tension.

You might be tempted to think that they are being irrational, giving in to temptation, or short-sighted. Nope. That's why these situations are so tricky: the trap depends on us being rational.
To show that they really are trapped, let's first see why an argument that they should keep their mouths' shut doesn't work. I'll then give you a shortcut version.

Suppose you're Scarlet. You know that Violet is self-interested so she's going to do whatever minimizes her sentence.

You also know that she is rational. That's crucial. Since both of you have the same goal (minimize your own sentence), both have been offered the same deal, and both are rational, that means both of you will come to the same conclusion. The situation is symmetric: If something is rational for you to do, it will be rational for her to do; and vice-versa. That means you can predict what she's going to do.

Don't believe me? Imagine that this isn't true. If Violet is unpredictable, from your perspective her choice is essentially random. There's a 50% chance she'll rat; 50% chance she'll keep quiet. All you can do is calculate the expected value of each outcome and be done.

But knowing that she's rational allows you to know exactly what she's going to do. Here's the same diagram from before (Figure 3) with the ranking of each outcome made explicit:

Figure 4

	Violet= Silent	**Violet= Rat**
Scarlet = Silent	Scarlet = 5 years 2nd Best Violet= 5 years 2nd Best	Scarlet = 10 years Worst Violet= 2 years Best
Scarlet = Rat	Scarlet = 2 years Best Violet= 10 years Worst	Scarlet = 8 years 2nd Worst Violet = 8 years 2nd Worst

If you conclude that you should rat, you know she will also come to the same conclusion. Thus both of you will be doomed to your second worst outcomes (the bottom right cell).

It seems crazy that you would choose this when you can instead choose to work together to get each of your second best outcome (upper left cell). You would think: Ok. I know that she will want to minimize her sentence. She's going to try to figure out what I'm going to do. Since I know that both of us keeping our mouths shut is better for me, she'll conclude that I will keep my mouth shut and therefore keep her mouth shut. Thus I should keep my mouth shut so I can get my second best option.[13]

[13] She'll also know that iocaine powder is from Australia. Australia is a land of thieves. And thieves are not used to being trusted, so I can clearly not choose the wine in front of me….

Unfortunately, this won't work. Think about what you've just concluded. You've reasoned your way into deciding that Violet will keep her mouth shut. You now know which column of our table you are in (the left). But now your choices are between your best and your second best outcome. And how do you get your best outcome? That's right, you rat.

Of course, since she is also rational Violet will come to exactly the same conclusion and decide to rat. If you know that she's going to rat, your choice is between your second worst and worst outcomes. Thus you still should rat. And, of course, vice-versa.

Clever, huh? Our DA deserves her bubbly.
Now that you've seen the our intrepid DA's full genius, we can cut out everything except the last step.[14] Here's the short version of the argument:

You are Scarlet. There are only two possible futures you care about. Either Violet rats or she doesn't.

Suppose she will rat (you're in the right-hand column). What should you do? Staying silent gives you your worst outcome, so you should rat.[15]

Suppose that she will stay silent (you're in the left-hand column). What should you do? Ratting gives you your best outcome versus your second best, so you should rat.

If you don't get this reference, go watch the Princess Bride. Now. I'll still be here.

[14] This shortcut doesn't work for other situations, that's why understanding the long version is important.

[15] The outcome where you stay silent and the other person rats is technically called 'The sucker's payoff'. I'm not kidding. It's sometimes hard to escape the sense that early decision-theorists were jerks….

Therefore, no matter what the other person does, you should rat. That will be true of them too, so both of you will rat each other out.
Let me introduce a bit of terminology that I'll use in what follows.

The prisoners' dilemma pops up in all sorts of situations. Thus to be general, instead of talking about ratting and keeping silent, I'll often talk about defecting and cooperating.

In game-theory, we would say that the strategy of ratting (defecting) is strictly dominant —it's always better to rat in this sort of situation.

Notice that the only thing which matters is how the costs and benefits are structured. If we replace the number of years with the rank of each outcome, we get the abstract structure of the trap as shown in Figure 5

Figure 5

	Violet = Cooperate	**Violet = Defect**
Scarlet = Cooperate	Scarlet = **2nd Best** Violet= **2nd Best**	Scarlet = **Worst** Violet = **Best**
Scarlet = Defect	Scarlet = **Best** Violet = **Worst**	Scarlet = **2nd Worst** Violet = **2nd Worst**

I'll call this the payoff structure. For any decision, if you describe the costs and benefits in each possible outcome, you've described the payoff structure.

If a decision has this payoff structure, in the words of Admiral Ackbar: It's a trap!
When you are in a trap like the prisoners' dilemma, the only way out is to change the payoff structure.

Notice why I've been so adamant about this being my world and you not getting to add information. The things you've been yelling at the screen are exactly the things which change the payoff structure in real situations.

If, as they say, snitches get stitches, the payoff structure is different. The choice isn't between 5 years and 2 years. It's between 2 years plus getting shanked and 5 years. That's a different calculus which makes it rational to keep their mouths shut.

Enough organized crime 101. Let's talk more about how we get out of these traps and then turn to the Tragedy of the Commons —the mother of all traps. To do that, let's get out our Oceana airlines tickets, head back to the island, and see if we can't midwife the market.
How can break out of the payoff structure and obtain forbidden fruit?

First, let's make sure we understand how our trade impasse is like the prisoners' dilemma. Abstracting it into a table like before makes it obvious:

Figure 6

	You: Trade	You: Steal

Me: Trade	Me: 1 persimmon You: 1 kiwi	Me: 0 fruit You: 1 persimmon; 1 kiwi
Me: Steal	Me: 1 persimmon; 1 kiwi You: 0 fruit	No trade[16]

Suppose you think I'm going to hand over my fruit (top row). Stealing (defecting) gives you a fruit for each hand, so you should try to steal.

Suppose you think I'm going to try to steal your fruit (bottom row). Handing over your fruit leaves you with no fruit. So you should steal/ refuse to hand it over.

This seems to show that trade is impossible unless something changes the payoff structure. You've probably already thought of many ways this can happen. Let's talk about a few.
Thomas Hobbes would think that the situation on the island is much worse than I've suggested. As many of you were no doubt thinking, there is an alternative to running away with my kiwi. You could wait until I'm asleep and then kill me with very little risk to yourself. Then it's kiwi-city for you.

As Hobbes noted, the fact that everyone has to sleep sometime means that anyone can kill anyone (we can also be sneaky, set traps, make temporary alliances with other weaker people, et cetera). Once the threat of violence is on the table, we're now in another prisoners' dilemma type payoff structure: The best outcome for each of us is killing the other; the worst is where we disarm and the other person kills us; second best is everyone refrains from killing each other; second worst is us (trying) to kill each

[16] Assume that we simultaneously grab for each other's fruits, struggle a bit, then give up.

other. Life on the island would suck. Or as he put it so much more eloquently[17] in one of my favorite paragraphs in all of philosophy:

> Whatsoever therefore is consequent to a time of war, where every man is enemy to every man, the same consequent to the time wherein men live without other security than what their own strength and their own invention shall furnish them withal. In such condition there is no place for industry, because the fruit thereof is uncertain: and consequently no culture of the earth; no navigation, nor use of the commodities that may be imported by sea; no commodious building; no instruments of moving and removing such things as require much force; no knowledge of the face of the earth; no account of time; no arts; no letters; no society; and which is worst of all, continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short.[18] [Leviathan, XIII]

Hobbes thought the only way out of this situation was an outside power who could keep everyone in check. The book this comes from (Leviathan) is an argument for absolute monarchies. Basically, the argument is that, sure, absolute monarchs can suck, but in anarchy things would be as bad as they can possibly be and without an absolute monarch you're gonna eventually get anarchy.

We don't need to buy his conclusion (that's a different class). But he is pointing us to the first way out of these traps: Laws and regulations. Once you add the threat of punishment to the outcomes, you change the payoff structure and it can become rational to cooperate.

Indeed, we don't need to rely on the state. A community can impose punishments through social sanctions like shame and ostracism. If we catch you stealing fruit, you don't get to come to the luau.
(Hint: keep this mechanism in mind for the assignment on the tragedy of the commons)

---

[17] Yeah, it's old-timey with long sentences. But read it out loud. The guy could write. The first line of his autobiography was "Fear and I were born twins." That's a good line.

[18] My dearly departed cat was named Hobbes because when she was a kitten she was nasty, brutish, and short.

Just FYI, Calvin in Calvin and Hobbes was named after John Calvin, another philosopher.

Ethical principles and moral values do the same job as externally imposed punishments. If there was honor amongst thieves, Scarlet and Violet might keep their mouths shut because their self-respect demands it. Not being able to look at yourself in the mirror is a heavy cost that would have to be weighed in deciding what to do.

Indeed, for someone with a lot of integrity who believes betraying trust is wrong, ratting might not even count as an option. She doesn't have to weigh less years versus betrayal, since betrayal isn't something she'd consider doing.
I set up the prisoners' dilemma with our criminals as complete strangers to undermine one of the most powerful ways we avoid these traps — friendships and personal connections. If you care about someone, you won't want to hurt them. That raises the cost of defecting (ratting; stealing).

If you're not seeing why, consider another trap in this family: the soldiers' dilemma. Think of the individuals in a conscript army. Each soldier can either fight or retreat. The best outcome for you (in terms of your chance of survival) is running away while everyone else stays and fights. Second best is everyone fights together. Second worst is everyone runs.[19] Worst is you stay and everyone else runs.

Historically, the payoff structure has been changed to make it rational for everyone to fight by either making it impossible to run (e.g., chaining soldiers to their posts in WWI; deep phalanx formations in ancient armies) or by raising the costs of running (e.g., the traditional role of officers in shooting deserters).

[19] Historically, the vast majority of casualties in battles happen when one army breaks formation and runs. If this happens, it's pretty likely you'll die. Though probably less likely than if you are the sole rear guard.

Modern armies have a better solution: make soldiers care about those in their units.[20] This is explicitly built into the culture of the US military.[21] If your unit is a band of brothers/sisters, you won't leave them behind anymore than you would your real brothers/sisters. Again, the personal bonds change the payoff structure and make it rational to cooperate.
You may have noticed my emphasis on the 1-time nature of the interaction all the examples. That's because the payoff structure changes when you may interact with the person again. You get what's called an iterated prisoners' dilemma.

On our island, if I steal your persimmon today, you will laugh in my face when I propose a trade tomorrow. Thus my choice is really giving up my kiwi in exchange for the persimmon and having the option of future persimmons versus keeping both the kiwi and the persimmon but having no future persimmons.

Indeed, once we recognize the possibility of future interactions, things like reputations start to matter a lot. When you're trying to decide whether to do business with someone, you better not just look at the details of the deal. You need to look who you're dealing with —their reputation. Sure, their unit costs are lower than the other suppliers. But do they have a reputation for making deliveries on time? Being litigious jerks? We can break out of iterated prisoners' dilemmas when there is a significant reputational cost to defecting. If we're smart when designing policies (for our company or as a government), we'll create mechanisms to ensure that people consider their reputations in deciding how to act.

[20] This isn't a uniquely modern solution. Indeed, you can find versions of this amongst elite units throughout history. The 150 pairs of male lovers comprising the Sacred Band of Thebes (c. 4th century BC) was a very successful version of this logic.
[21] I've been told by social psychologist friends that there's a lot of interesting psychological research behind boot camp / basic training.

Finally, it's interesting that even if we strip out all the human elements, the fact that there will be multiple interactions changes the optimal strategy. Until recently, the optimal strategy in a game where 2 computers repeatedly interact with each other in a prisoner's dilemma type situation is called 'Tit-for-Tat.' It has just 2 rules: (1) On the first move, cooperate; (2) On every subsequent move, do what the opponent did last time.[22] This strategy actually shows up in surprising places like peer-to-peer file sharing algorithms.

Many of the hardest problems humanity faces, including catastrophic climate change, are tragedies of the commons. Like its cousin the prisoners' dilemma, a tragedy of the commons is a rigged game. If you play, you lose. The only way out is to flip over the table and play a different game.

Let me emphasize again that these situations only arise when we are rational —when everyone's decision-making is as good as it gets. We all do dumb or shortsighted things. Ordinary human stupidity gets humanity in lots of trouble. But we know the outlines of the solution: help people be less stupid. Ironically, in a tragedy of the commons, it sometimes might be better for all involved if they were stupider.

One last attempt to convey the strangeness of these situations: My mental image of a tragedy of the commons is all of us packed together in a big group, shuffling towards the edge of a cliff. We all know what's happening. "We should stop" someone says. "Yep. Definitely should stop" comes the reply from everyone else. Yet rationality demands that we keep shuffling forward until we all plunge off the cliff.[23]

---

[22] I say 'until recently' because there are some tricks which do better if you can have teams of players. See https://www.wired.com/2004/10/new-tack-wins-prisoners-dilemma/

[23] The tragedy of the commons has historically had execrable associations with attempts to justify racism, eugenics, mass starvation, and other terrible policy preferences. Indeed, the writer who introduced the term was, to be blunt, a world class asshole. Like knives, ideas can be used for good or for evil. Being

Before we begin, an important warning. The material so far is basically what you'll find if you use Professor Google or Teaching Assistant YouTube. If my explanations have left you wanting, I encourage you to look for other explanations which make more sense to you.

However, Professor Google will betray you if you ask her about the tragedy of the commons. We're going to talk about a more sophisticated version that you won't find unless you get fairly deep into game-theory and economics (particularly the work for which Elinor Ostrom won the 2009 Nobel Prize).[24] Indeed, even if you were familiar with her work, the way I'm going to explain it may seem upside down.

Thus, unfortunately, the materials for our class and our discussions will be almost entirely your sole source for your essay.

If you ask Professor Google about the tragedy of the commons, she'll tell you something like this: If people share a common resource and it is possible to free-ride (take without paying), there is little incentive to maintain it, so it will fall to ruin. This certainly can be a problem: if you have roommates, think about the kitchen sink. The standard example (which is a complete historical fiction[25]) is a village that has a communal grazing area for villagers' sheep. If the grazing area doesn't belong to anyone, everyone will allow their sheep to overgraze it. Tragically the common property gets ruined.

---

aware of the history helps us avoid the misuses. This summarizes some of the concerns: https://blogs.scientificamerican.com/voices/the-tragedy-of-the-tragedy-of-the-commons/

[24] https://www.nobelprize.org/prizes/economic-sciences/2009/ostrom/biographical/ See Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*

[25] See Cox, Susan *No Tragedy on the Commons* http://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/3113/buck_NoTragedy.pdf?sequence=1&isAllowed=y

That is **not** what we are discussing. It is related. But it is very different. Do not be tempted by Professor Google.[26]
Let's start with a story.

We live in a small fishing village by a lake. Our entire way of life is based on catching fish from the lake and selling them to outsiders. We are good at fishing and our fish are delicious, so we catch a lot of them.

Things have been good; they used to be great. But we begin to notice that big catches are getting rarer and rarer.

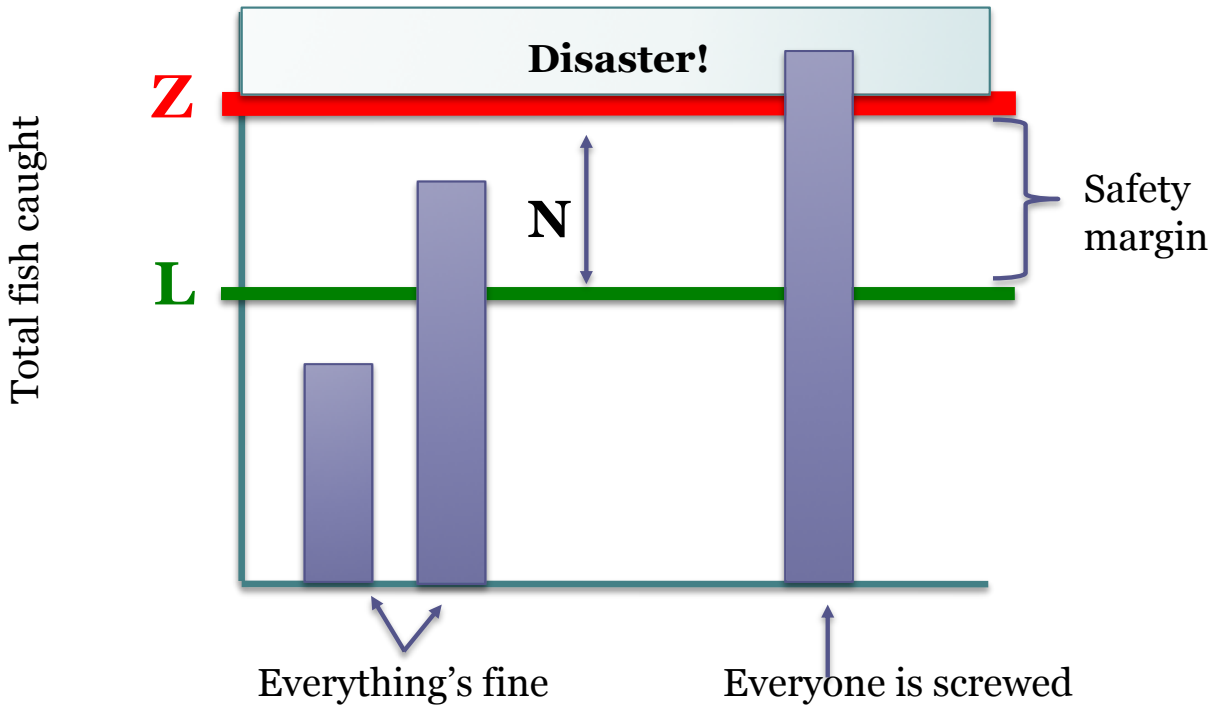
We call in a team of ecologists to study the fish population. I hope this isn't news to you, but to have fish next year, there have to be mommy and daddy fish this year (plus some Al Green and romantic seaweed dinners). They find that we've substantially reduced the number of breeding pairs. We are headed for disaster.

The good news is that the population is okay for right now. They're even able to estimate how many fish we can safely catch before there's no longer enough fish to sustain our economy.

We can illustrate the situation with 7

Figure 7

[26] That's not to say the questions around free-rider/collective action problems aren't interesting. Here are some additional resources: https://plato.stanford.edu/entries/free-rider/



The vertical axis represents the total number of fish caught. Z represents the point of no return. If we catch one more fish than Z, our village economy will collapse. [27]

Being good scientists, our consulting ecologists gave us an estimate of how many fish comprise Z along with an explanation of the uncertainties in their analysis. We think about other things that may affect the fish population and give ourselves a safety margin of N fish.
Thus L is the limit we agree to stay below.[28]

To summarize:

Z	Disaster level. If our catch is > Z, we're screwed.

[27] Realistically there wouldn't be such a sharp cut-off. We also would want to focus on the yearly catch and then translate it into daily limits. Recognizing these points makes the math a bit harder, but doesn't change the situation. Thus I'm going to be sloppy about them to make the story easier to tell.
[28] Again, this helps simplify. We don't actually need an agreed upon limit to have a tragedy of the commons. But without the limit, it is much harder to describe without equations since the usual language of defecting and cooperating doesn't really work.

L	The limit we've agreed upon to avoid going over Z
N	The difference between Z and L — the safety margin

Here's the first really important thing: If we go over the limit L but stay under Z, nothing happens. Everything is fine. That's why I've labeled N the safety margin. You can also think of N as the number of people who can each catch 1 extra fish before we reach the disaster.
Here's what happens. We all head out on our boats. We catch fish until we've reached our share of the limit (L). Then we head back to port.

On our way back, each of us will look out over the side and face the same choice. Right there. Right next to the boat. It's a fish. Clearly the smug little bastard knows he's safe. He's basically taunting you, doing the backstroke, giving you the middle finger (fin?). You could reach out with your net and catch him with no effort. Should you catch him?

In a tragedy of the commons, no one person can cause the tragedy on their own. Any one of us could catch as many fish as we can and not make a dent in the population. The problems only occur when enough —i.e., N— people choose to catch an extra fish.

Indeed, if the number of defectors it takes to cause the disaster (N) is big enough, its arguable that catching the extra fish doesn't make the situation worse at all. This is crucial since it means your choice whether to catch the extra fish is not between the disaster and no-disaster. No rational person

would choose the disaster.[29] Whether there's a disaster is not up to you; it's determined by the collective action of the village.[30]

What should you do?

You might think that the decision requires knowing how many other people will cheat. Actually, it doesn't. That's because of the nature of the payoff structure forces you to the same conclusion, no matter the chances of the outcome.[31]

It will help to put dollar values on the outcomes for illustration. Let's say that you can sell a fish for $100 and that, for you, going out of business is equivalent to a $1 million loss over your lifetime.

Like we did for the prisoners dilemma we can represent your decision with columns being the different futures which you have no control over and the rows being your choices:

	# Caught <= Z No Disaster	# Caught > Z Disaster
Catch extra fish		
Don't catch extra fish		

[29] To put it another way, if we knew that we were right at Z —catch 1 more fish and we're doomed— it would not be rational to catch the extra fish.

[30] Gold star if you detect a strong whiff of the Share of the Total principle operating here. If you didn't, try. It smells a bit burnt, because that principle is toast. This is an avenue for breaking out of the problem.

[31] Okay. That's not completely true. Adding probabilities will shift around the expected values in the outcomes. That might constrain when the tragedy can occur in some cases, but the structure will usually still be there.

Since catching the extra fish brings in an extra $100, we can fill in the rows. Note that this is extra income over and above your usual income; you're not going broke.

	# Caught <= Z No Disaster	# Caught > Z Disaster
Catch extra fish	$100 extra fish	$100 extra fish
Don't catch extra fish	$0 extra fish	$0 extra fish

The loss of your fishing business is a $1 million loss.

	# Caught <= Z No Disaster	# Caught > Z Disaster
Catch extra fish	$100 extra fish $0 loss	$100 extra fish ($1,000,000) loss
Don't catch extra fish	$0 extra fish $0 loss	$0 extra fish ($1,000,000) loss

That allows us to calculate the values of each outcome

	# Caught <= Z No Disaster	# Caught > Z Disaster
Catch extra fish	$100 extra fish $0 loss = $100	$100 extra fish ($1,000,000) loss = ($999,900)
Don't catch extra fish	$0 extra fish $0 loss = No change	$0 extra fish ($1,000,000) loss = ($1,000,000)

Hopefully, you recognize this payoff structure and are hearing ominous music in the background.[32]

To see why we go over the cliff, let's take each possible future separately like we did in the shortcut version of the prisoners' dilemma.

First, suppose there is no disaster

	# Caught <= Z No Disaster
Catch extra fish	$100 extra fish $0 loss = $100
Don't catch extra fish	$0 extra fish $0 loss = No change

What should you do? Catching the extra fish brings in $100. Having $100 extra is better than having $0. Hence, if there's not going to be a disaster, you should catch the fish.

Second, suppose that there is going to be a disaster

	# Caught > Z Disaster
Catch extra fish	$100 extra fish ($1,000,000) loss = ($999,900)

[32] If not: https://www.youtube.com/watch?v=ZvCI-gNK_y4

Don't catch extra fish	$0 extra fish ($1,000,000) loss = ($1,000,000)

Either way, you're going to have a huge loss. But it's better to have a small gain to offset the huge loss.[33] So, if there's going to be the disaster, you should catch the fish.

Therefore, no matter what other people do (whether or not there's a disaster), you should catch the extra fish —as in the prisoners' dilemma, catching the fish is strictly dominant.

Of course, everyone is rational just like you. If the best thing for your to do is catch the extra fish, the best thing for each other person to do is catch the extra fish. Therefore, everyone will catch the extra fish. Our village is doomed.

Go home rationality, you're drunk.
On the count of 3, answer me this question "How do you get out of any tragedy of the commons?"

Ready, 1…..2…..3!

[33] If this strikes you as crazy, you're onto something. There are several issues here which go very deep. First, decision theory (mostly) depends on being able to analyze outcomes independently (i.e., compare 2 boxes). Economics majors may have heard of this as the Independence of Irrelevant Alternatives principle. This (intentionally) loses the bigger picture. But sometimes the bigger picture is crucial. It's arguable that focusing on the $100 loses sight of the fact that it's a disaster.
Second, this decision seems like the mistake Bernard Williams calls 'one thought too many'. If you find yourself thinking that it is better to do something which kills a million people instead of killing a million and one, something has gone terribly wrong in your reasoning. You've had one thought too many. Indeed, this is a place where critiques by feminist and non-western philosophers who reject the whole picture of rationality as a calculus of costs and benefits have a lot of teeth. These issues would require several courses to explore. The CSUN philosophy department offers all of them! Hint hint

CHANGE THE PAYOFF STRUCTURE!

That's right! We need to change the payoff structure. But how?

Abstractly, the answer is simple: we need to make it no longer beneficial to catch the extra fish. That means any solution must do one of two things (or a mix):

(1) Impose additional costs to catching the extra fish

or

(2) Make it more beneficial to not catch the extra fish.

I'm going to leave it to you in your essays to think through how to do this. You'll probably want to look back at Changing the payoff structure for a prisoners' dilemma for some ideas.

There is one tempting solution which will not work, and, worse, potentially lead you astray.

You might think there needs to be an advertising or other educational campaign to let people know the dangers of overfishing. That won't work because the problem has to do with incentives (the costs and benefits of the options), not a lack of information. In fact, you can imagine that an educational campaign might make things worse. Maybe some villagers were blissfully unaware of the danger. Now they realize that they better catch as many fish as they can to give themselves as big a cushion as possible when the disaster hits.

That said, I can imagine ways this might work. The key is that the advertising can't be directed at informing people. Feeling guilty is a cost. Maybe the ad agency that does those really sad PETA ads could put something together that makes people feel so guilty about contributing to overfishing that it's no longer in their interest to catch the extra fish.[34] The effectiveness of this approach is questionable; it certainly won't generalize to all tragedies of the commons.

Let's close on a sobering note. Anthropogenic climate change is the mother of all tragedies of the commons. Worse, it involves several features that human psychologies just aren't good at dealing with.[35]

Reducing greenhouse gas emissions will have an economic cost.[36] Suppose that no single country on its own can cause climate change. Obviously, some countries account for a substantial fraction of greenhouse gas emissions. But if every other country magically stopped emitting, we'd be okay.[37]

---

[34] For the econ majors and other fans of introductory microeconomics in the room: How much guilt? Presumably $101 worth. That is, the amount of guilt you would be willing to pay $101 to not feel. This is a useful way of thinking about human behavior when you're trying to understand how to make it mathy with models. But for the rest of us, including many professional economists, it seems very strange. (An economics professor friend jokes that when you start a PhD program, you finally know enough and have the tools to you need to start unlearning everything you learned as an undergrad.)

[35] To be clear, as with other examples of irrationality, that doesn't mean we can't get over these barriers. It just means it requires effort. Of course, the reward for expending the effort in this case is recognizing the likelihood of mass human suffering. That makes it hard to want to think about it, even for those who recognize that they should.

[36] There are economic gains to be had from an economy that has transitioned to green energy sources. Indeed, we may be underestimating the economic costs of not doing so, e.g., https://www.vox.com/energy-and-environment/2018/6/8/17437104/climate-change-global-warming-models-risks However, the transition incurs costs. From many counties's point of view it may be better to let another country spend the resources developing the technology and just buy it from them. Here's an accessible article which details the scale of the R&D costs involved: https://www.motherjones.com/politics/2019/12/kevin-drum-climate-change-reseach/ Though, admittedly, the costs may be lower, e.g,, https://www.vox.com/2020/1/22/21028914/germany-green-new-deal-solar-power and Nobel Laureat Paul Romer's comments https://twitter.com/NobelPrize/status/1049248005210165248

[37] This is definitely true for every country right now with the possible exceptions of the US and China. Still, if either of those were the only emitter, the problem would be largely solved.

Think about the logic for each country. If there's no climate disaster because enough other countries reduced their emissions, it's better to have not incurred the costs of transitioning. If there is a climate disaster, having more economic growth means more resources for offsetting the costs of the disaster. Thus not reducing emissions is strictly dominant

Unfortunately, in addition to being structurally a tragedy of the commons, anthropogenic climate change layers on several well-known blindspots/ defects of our psychologies. I'll mention just a few.

Nothing in our evolutionary environment prepared our psychologies for dealing with situations where each of us makes a tiny contribution which results in a huge problem.

Similarly, we tend to do a bad job of taking proper account of losses which occur decades in the future.[38]

Our thinking often has a just world bias: we want to believe that things are okay and resist evidence otherwise.

Finally, the fact that those likely to be hardest hit by the disaster live in poorer countries (which aren't the main emitters) plays into the suite of biases that lead us to care less about strangers, especially those far away.

None of this is to say that we should give up and consign our children and grandchildren to a world far worse than our own. No matter what you think should be done about anthropogenic climate change, it is important

[38] Though this seems less of an issue than when I first started giving this lecture 10 years ago….

to have a grip on what creates the problem.[39]

[39] Some writers deny that anthropogenic climate change is really a tragedy of the commons. They claim the problem is rich companies derailing every attempted solution. I think that conflates what the phenomenon is and what makes it hard to deal with. It can both be true that it is a tragedy of the commons and that efforts have been frustrated by monied interests.
See Hale, *Catalytic Cooperation* for the sort of view I'm criticizing: https://www.bsg.ox.ac.uk/sites/default/files/2018-09/BSG-WP-2018-026.pdf