
NE
E
SE
S
SW
W
NW
N

Intuitive Introduction to Functional Decision Theory

Basic Concepts in Decision Theory

Imagine I show you two boxes, A and B. Both boxes are open: you can directly observe their contents. Box A contains $100, while box B contains $500. You can choose to receive either box A or box B, but not both. Assume you just want as much money as possible; you don't care about me losing money or anything (we will assume this pure self-interest in all problems in this sequence). Let's call this Problem 1. Which box do you choose?

I hope it's obvious picking box B is the better choice, because it contains more money than box A. In more formal terms, choosing B gives you more utility. Just like temperature is a measure for how hot something is, utility measures how much an outcome (like getting box B) satisfies one's preferences. While temperature has units like Celsius, Fahrenheit and Kelvin, the "official" unit of utility is the *util* or *utilon* - but often utility is expressed in dollars. This is how it's measured in Problem 1, for example: the utility of getting box A is $100, while getting B gives $500 utility. Utility will always be measured in dollars in this sequence.

Now imagine I again show you two boxes A and B - but now, they are closed: you can't see what's inside them. However, I tell you the following: "I flipped a fair coin. If it was heads, I filled box A with $100; otherwise, it contains $200. For box B, I flipped another fair coin; on heads, box B contains $50; on tails, I filled it with $150." Assume I am honest about all this. Let's call this one Problem 2. Which box do you pick?

Problem 2 is a bit harder to figure out than Problem 1, but we can still calculate the correct answer. A fair coin, by definition, has 0.5 (50%) probability of coming up heads and 0.5 probability of coming up tails. So if you pick box A, you can expect to get $100 with probability 0.5 (if the coin comes up heads), and $200 also with probability 0.5 (if the coin comes up tails). Integrating both into a single number gives an *expected utility* of 0.5 x $100 + 0.5 x $200 = $50 + $100 = $150 for picking box A. For picking box B, the expected utility equals 0.5 x $50 + 0.5 x $150 = $25 + $75 = $100. The expected utility associated with choosing box A is higher than that of choosing B; therefore, you should pick A in Problem 2. In other words, picking A is the *rational action*: it *maximizes expected utility*.

Note that we could have calculated the expected utilities in Problem 1 as well, and actually kind of did so: there, choosing box A gets you $100 with probability 1. The expected utility would be 1 x $100 = $100. For picking box B, it's $500. Since the probabilities are 1, the expected utilities simply equal the dollar amounts given in the problem description.

Decision Theory is a field that studies principles that allow an *agent* to do the *actions* that give the most *expected utility*. An agent is simply an entity that observes its environment and acts upon it; you are an agent, for example. You observe the world through your senses and act upon it, e.g. by driving a car around. Ducks are also agents, and so is the Terminator. In both Problem 1 and 2, the available *actions* are "Pick box A" and "Pick box B", and they have expected utilities associated with each of them. (For the fans: the Terminator has high expected utility for the action "Shoot Sarah Connor in the face.")

At this point you may wonder why we need a whole field for making rational decisions. After all, Problems 1 and 2 seem pretty straightforward. But as we will soon see, it's

pretty easy to construct problems that are a lot more difficult to solve. Within the field of Decision Theory, different decision theories have been proposed that each have their own recipe for determining the rational action in each problem. They all agree an agent should maximize expected utility, but disagree on how to exactly calculate expected utility. Don't worry: this will become clear later. First, it's time to discuss one of the classic decision theories: *Causal Decision Theory*.

An Intuitive Introduction to Causal Decision Theory

Like any decision theory, Causal Decision Theory (CDT) aims to maximize expected utility; it does this by looking at the *causal effects* each available action in a problem has. For example, in Problem 1, taking box A has the causal effect of earning $100, whereas taking box B causes you to earn $500. $500 is more than $100, so CDT says to take box B (like any decision theory worth anything should). Similarly, CDT advices to take box A in Problem 2.

CDT's rule of looking at an action's causal effects make sense: if you're deciding which action to take, you want to know how your actions change the environment. And as we will see later, CDT correctly solves the problem of the Smoking Lesion. But first, we have to ask ourselves: what is causality?

What is causality?

A formal description of causality is beyond the purpose of this post (and sequence), but intuitively speaking, causality is about stuff that makes stuff happen. If I throw a glass vase on concrete, it will break; my action of throwing the vase *caused* it to break.

You may have heard that correlation doesn't necessarily imply causality, which is true. For example, I'd bet hand size and foot size in humans strongly correlate: if we'd measure the hands and feet of a million people, those with larger hands will - on average - have larger feet as well, and vice versa. But hopefully we can agree hand size doesn't have a *causal effect* on foot size, or vice versa: your hands aren't large or small *because* your feet are large or small, even though we might be able to quite accurately predict your foot size *using* your hand size. Rather, hand size and foot size have *common causes* like genetics (determining how large a person can grow) and quality and quantity of food taken, etc.

Eliezer Yudkowsky describes causality in a the following very neat way:

> There's causality anywhere there's a noun, a verb, and a subject.

"I broke the vase" and "John kicks the ball" are both examples of this.

With the hope the reader now has an intuitive notion of causality, we can move on to see how CDT handles Smoking Lesion.

Smoking Lesion

> An agent is debating whether or not to smoke. She knows that smoking is correlated with an invariably fatal variety of lung cancer, but the correlation is (in this imaginary world) entirely due to a common cause: an arterial lesion that causes those afflicted with it to love smoking and also (99% of the time) causes them to develop lung cancer. There is no direct causal link between smoking and

> lung cancer. Agents without this lesion contract lung cancer only 1% of the time, and an agent can neither directly observe nor control whether she suffers from the lesion. The agent gains utility equivalent to $1,000 by smoking (regardless of whether she dies soon), and gains utility equivalent to $1,000,000 if she doesn't die of cancer. Should she smoke, or refrain?

CDT says "yes". The agent either gets lung cancer or not; having the lesion certainly increases the risk, but smoking doesn't causally affect whether or not the agent has the lesion and has no direct causal effect on her probability of getting lung cancer either. CDT therefore reasons that whether you get the $1,000,000 in utility is beyond your control, but smoking simply gets you $1,000 more than not smoking. While smokers in this hypothetical world more often get lung cancer than non-smokers, this is because there are relatively more smokers in that part of the population that has the lesion, which is the cause of lung cancer. Smoking or not doesn't change whether the agent is in that part of the population; CDT therefore (correctly) says the agent should smoke. The Smoking Lesion situation is actually similar to the hands and feet example above: where e.g. genetics cause people to have larger hands *and* feet, the Smoking Lesion causes people to have cancer *and* enjoy smoking.

CDT makes intuitive sense, and seems to solve problems correctly so far. However, it does have a major flaw, which will become apparent in Newcomb's Problem.

Newcomb's Problem

> A superintelligence from another galaxy, whom we shall call Omega, comes to Earth and sets about playing a strange little game. In this game, Omega selects a human being, sets down two boxes in front of them, and flies away.
>
> Box A is transparent and contains a thousand dollars.
> Box B is opaque, and contains either a million dollars, or nothing.
>
> You can take both boxes, or take only box B.
>
> And the twist is that Omega has put a million dollars in box B iff Omega has predicted that you will take only box B.
>
> Omega has been correct on each of 100 observed occasions so far - everyone who took both boxes has found box B empty and received only a thousand dollars; everyone who took only box B has found B containing a million dollars. (We assume that box A vanishes in a puff of smoke if you take only box B; no one else can take box A afterward.)
>
> Before you make your choice, Omega has flown off and moved on to its next game. Box B is already empty or already full.
>
> Omega drops two boxes on the ground in front of you and flies off.
>
> Do you take both boxes, or only box B?

(Note that "iff" means "if and only if.)

How does CDT approach this problem? Well, let's look at the causal effects of taking both boxes ("two-boxing") and taking one box ("one-boxing").

First of all, note that Omega has already made its prediction. Your action now doesn't causally affect this, as you can't cause the past. Omega made its prediction and based upon it either filled box B or not. If box B isn't filled, one-boxing gives you nothing; two-boxing, however, would give you the contents of box A, earning you $1,000. If box B *is* filled, one-boxing gets you $1,000,000. That's pretty sweet, but two-boxing gets you $1,000,000 + $1,000 = $1,001,000. In both cases, two-boxing beats one-boxing by $1,000. CDT therefore two-boxes.

John, who is convinced by CDT-style reasoning, takes both boxes. Omega predicted he would, so John only gets $1,000. Had he one-boxed, Omega would have predicted *that*, giving him $1,000,000. If only he hadn't followed CDT's advice!

Is Omega even possible?

At this point, you may be wondering whether Newcomb's Problem is relevant: is it even possible to make such accurate predictions of someone's decision? There are two important points to note here.

First, yes, such accurate predictions might actually be possible, especially if you're a robot: Omega could then have a copy - a model - of your decision-making software, which it feeds Newcomb's Problem to see whether the model will one-box or two-box. Based on *that*, Omega predicts whether *you* will one-box or two-box, and fixes the contents of box B accordingly. Now, you're not a robot, but future brain-scanning techniques might still make it possible to form an accurate model of your decision procedure.

The second point to make here is that predictions need not be this accurate in order to have a problem like Newcomb's. If Omega could predict your action with only 60% accuracy (meaning its prediction is wrong 40% of the time), e.g. by giving you some tests first and examine the answers, the problem doesn't fundamentally change. CDT would still two-box: *given* Omega's prediction (whatever its accuracy is), two-boxing still earns you $1,000 more than one-boxing. But, of course, Omega's prediction is *connected* to your decision: two-boxing gives you 0.6 probability of earning $1,000 (because Omega would have predicted you'd two-box with 0.6 accuracy) and 0.4 probability of getting $1,001,000 (the case where Omega is wrong in its prediction), whereas one-boxing would give you 0.6 probability of getting $1,000,000 and 0.4 probability of $0. This means two-boxing has an expected utility of 0.6 x $1,000 + 0.4 x $1,001,000 = $401,000, whereas the expected utility of one-boxing is 0.6 x $1,000,000 + 0.4 x $0 = $600,000. One-boxing still wins, and CDT still goes wrong.

In fact, people's microexpressions on their faces can give clues about what they will decide, making many real-life problems *Newcomblike*.

Newcomb's Problem vs. Smoking Lesion

You might be wondering about the exact difference between Newcomb's Problem and Smoking Lesion: why does the author suggest to smoke on Smoking Lesion, while also saying one-boxing on Newcomb's Problem is the better choice? After all, two-boxers indeed often find an empty box in Newcomb's Problem - but isn't it also true that smokers often get lung cancer in Smoking Lesion?

Yes. But the latter has nothing to do with the *decision to smoke*, whereas the former has everything to do with the *decision to two-box*. Let's indeed assume Omega has a model of your decision procedure in order to make its prediction. Then whatever you decide, the model also decided (with perhaps a small error rate). This isn't different than two calculators both returning "4" on "2 + 2": if your calculator outputs "4" on "2 + 2", you know that, when Fiona input "2 + 2" on her calculator a day earlier, hers must have output "4" as well. It's the same in Newcomb's Problem: if you decide to two-box, so did Omega's model of your decision procedure; similarly, if you decide to one-box, so did the model. Two-boxing then systematically leads to earning only $1,000, while one-boxing gets you $1,000,000. Your decision procedure is instantiated in two places: in your head and in Omega's, and you can't act as if your decision has no impact on Omega's prediction.

In Smoking Lesion, smokers do often get lung cancer, but that's "just" a statistical relation. Your decision procedure has no effect on the presence of the lesion and whether or not you get lung cancer; this lesion does give people a *fondness of* smoking, but the decision to smoke is still theirs and has no effect on getting lung cancer.

Note that, if we assume Omega doesn't have a model of your decision procedure, two-boxing would be the better choice. For example, if, historically, people wearing brown shoes always one-boxed, Omega might base its prediction on *that* instead of on a model of your decision procedure. In that case, your decision doesn't have an effect on Omega's prediction, in which case two-boxing simply makes you $1,000 more than one-boxing.

Conclusion

So it turns out CDT doesn't solve every problem correctly. In the next post, we will take a look at another decision theory: *Evidential Decision Theory*, and how it approaches Newcomb's Problem.

An Intuitive Introduction to Evidential Decision Theory

In the last post, we discussed Causal Decision Theory (CDT). We learned how purely looking at the causal effects of your actions makes you lose at Newcomb's Problem. Evidential Decision Theory (EDT) is an alternative decision theory that tells us to take the action that, *conditional upon it happening*, gives the best outcome. This means EDT looks at the *evidence* each action provides for the different outcomes. It's best to explain what this means exactly by looking at a few known problems. But first, what is "evidence", anyway?

What is evidence, anyway?

"Evidence" may sound like a complicated term, but really, it's extremely simple. Imagine we have a big box with 1000 items in it. 600 of these items are balls; 400 are cubes. Of the balls, 200 are red, whereas 400 are green. Furthermore, 300 cubes are red, and 100 are green. I draw a random item from our big box. What is the probability that the item is a ball?

Well, there are 1000 items, of which 600 are balls. That gives us a probability of 600 / 1000 = 0.6 that a random item is a ball. We write P(B) = 0.6, where the B stands for an item being a ball.

Now, suppose I putt the ball back and again draw a random item from our box. I tell you that it is red. What is the probability that it is a ball?

In total, there are 500 red items (200 red balls + 300 red cubes). Of those red items, 200 are balls. 200 / 500 = 0.4; so that's a 0.4 *probability* that a red item is a ball. We write P(B | R) = 0.4: the probability that an item is a ball *given that* or *conditional on the fact that* it is red equals 0.4. P(C | R) - the probability of an item being a cube *given* that it is red - equals 300 red cubes / 500 red items = 0.6. Note that P(B | R) + P(C | R) = 0.4 + 0.6 = 1 (or 100%). P(B | R) and P(C | R) adding up to 1 makes sense - a red item must either be a ball or a cube.

P(B | R) = 0.4 tells us the *evidence* redness gives us about the item being a ball. That evidence is quite weak: the evidence redness gives for the item being a cube is bigger (0.6). Therefore, we would usually simply say that redness is evidence for the item being a cube. This is simply because there are more red cubes than red balls.

We can easily turn things around. P(R | B) is the probability that an item is red given that it is a ball. There are 600 balls, 200 of which are red. Then P(R | B) = 200 / 600 = $\frac{1}{3}$. P(G) = 500 / 1000 = 0.5 - the probability a random item is green. P(G | B) = 400 / 600 = $\frac{2}{3}$ = 1 - P(R | B).

This is the nature of *evidence* as Evidential Decision Theory uses it. What I gave here is a very basic and very incomplete introduction to *Bayesian* reasoning - if you want to learn more, I recommend An Intuitive Explanation of Bayes's Theorem.

Newcomb's Problem

Remember how in Newcomb's Problem, every two-boxer got $1,000 and every one-boxer got $1,000,000? That means two-boxing is very strong *evidence of* getting $1,000, just like redness is strong evidence of an item being a cube in our example above. Similarly, one-boxing is very strong evidence of getting $1,000,000. Therefore, if we assume Omega is always right in its prediction, *conditional on* you two-boxing, you'd earn $1,000,000; conditional on one-boxing, you'd only get $1,000. One-boxing is then clearly better, and EDT indeed one-boxes. Therefore, following EDT earns you a $1,000,000 in Newcomb's Problem, and EDT'ers do clearly better than CDT'ers here. So how does EDT perform on Smoking Lesion?

Smoking Lesion

In the world of Smoking Lesion, smoking is evidence of getting lung cancer: the problem states smoking is *correlated with* lung cancer. As we know, this correlation is due to a *common cause*: both lung cancer and a fondness of smoking are caused by a genetic lesion. Smoking doesn't cause lung cancer in this world, but EDT doesn't deal with causes: it deals with *evidence*, and P(Cancer | Smoke) is quite high. Smokers get lung cancer more often than non-smokers: smoking is *evidence for* getting lung cancer. As life without cancer is preferred, EDT'ers don't smoke, and simply lose $1,000 in utility: whether or not they have the lesion, smoking doesn't change their probability of developing cancer, and they might as well smoke.

Final Remarks

Where EDT does better than CDT on Newcomb's Problem, CDT wins at Smoking Lesion. EDT's problem in Smoking Lesion is that *correlation doesn't equal causation*. When deciding what action to take, EDT imagines what the world would be like after each action: for each action, it constructs a different world, and *keeps the correlations intact*. So, in Newcomb's Problem, it keeps the correlations caused by Omega's predictions intact, which is correct. However, in Smoking Lesion, the correlation between smoking and getting lung cancer is kept, which is wrong: that correlation is merely "accidental". When considering smoking, one *should* imagine a world in which one enjoys smoking *with* a probability of getting lung cancer; when considering not smoking, one should imagine a world in which one is *not* enjoying smoking *with the same probability of getting lung cancer*.

CDT constructs *its* worlds by keeping only causal effects of the actions intact. This works well in Smoking Lesion, but in Newcomb's Problem, this means cutting the connection between its action and Omega's prediction. This is wrong: CDT acts like Omega's probability of filling box B to be the same regardless of its decision, but that's not the case. The exact way CDT goes wrong will become more clear in the next post, where I finally introduce *Functional Decision Theory*.

An Intuitive Introduction to Functional Decision Theory

In the last two posts, we looked at both Causal (CDT) and Evidential Decision Theory (EDT). While both theories have their strength, CDT fails on Newcomb's Problem and EDT performs poorly on Smoking Lesion. Both problems *do* have a correct answer if you want to get as much utility as possible: one-boxing *results* in earning $1,000,000, and smoking results in an extra $1,000.

Functional Decision Theory (FDT) is a decision theory that manages to answer both problems correctly. Where CDT looks at the causal effects of the available *actions*, FDT considers the effects of its *decision*. It asks:

> Which output of this decision procedure causes the best outcome?

Where the decision procedure is simply the reasoning or "thinking" done to determine what action to do. This might seem like only a superficial difference with CDT, but that really isn't the case. The crucial point is that *the same decision procedure can be implemented by multiple physical systems*. It's best to explain this using an example.

Psychological Twin Prisoner's Dilemma

The Psychological Twin Prisoner's Dilemma (PTPD) runs as follows:

> An agent and her twin must both choose to either “cooperate” or “defect.” If both cooperate, they each receive $1,000,000. If both defect, they each receive $1,000. If one cooperates and the other defects, the defector gets $1,001,000 and the cooperator gets nothing. The agent and the twin know that they reason the same way, using the same considerations to come to their conclusions. However, their decisions are causally independent, made in separate rooms without communication. Should the agent cooperate with her twin?

Looking at this in the most selfish way possible, the best outcome for the agent is the one where she defects and her twin cooperates, in which case she gets $1,001,000. However, note that this situation is *impossible*: the agent and her twin *reason the same way*. If the agent wants the $1,001,000 outcome, so does her twin; in that case, they both defect and both get $1,000, while nobody gets the $1,001,000.

The reader might have noticed there are only two outcomes possible here: either both the agent and her twin cooperate, or they both defect. The agent prefers the first outcome: that's the one where she gets $1,000,000, whereas the defect-defect scenario only gets her $1,000. CDT'ers, however, defect: since the agent's decision doesn't causally effect that of the twin, they reason that *given* the decision of the twin, defecting always gives $1,000 more. This is quite straightforward: given defection by the twin, defecting gives the agent $1,000 whereas cooperating gives her nothing; given cooperation by the twin, the agent gets $1,001,000 by defecting and $1,000,000 by cooperating, making PTPD analogous to Newcomb's Problem. (Indeed, the problems are equal; the names of the actions are different, but the expected utilities are the same.)

CDT fails to incorporate the connection between the agent's and her twin's decision making in her decision making, and reasons as if the two are independent. But that's not the case: just like two identical calculators both return 4 to 2 + 2, two agent's who reason the same way on the same problem come to the same conclusion. Whatever the agent decides, her twin also decides, because, crucially, *her twin has the same decision procedure as she does*. The agent's decision procedure is "implemented" twice here: once in the agent, and once in her twin. FDT asks: "Which output of this decision procedure causes the best outcome?" Well, if the output of this decision procedure is "defect", then all "physical systems" (in this case, the agent and her twin) implementing this decision procedure defect, earning the agent $1,000. If the output of this decision procedure is "cooperate", then both the agent and her twin cooperate, getting the agent a sweet $1,000,000. FDT therefore recommends cooperating!

Note that EDT also cooperates, as it recognizes that cooperating is evidence for the twin cooperating, whereas defecting is evidence for defection by the twin. This is, however, the wrong reason to cooperate: as we saw in the Smoking Lesion problem, mere correlation doesn't provide a good basis for decision making. As we will see later in this post, there are problems where FDT beats EDT and even problems where it beats both CDT *and* EDT.

Newcomb's Problem

We discussed the similarity between Newcomb's Problem and PTPD before, and it might not surprise you to read that FDT one-boxes on Newcomb's Problem. The reason has been briefly touched upon before, and is the same as the reason for cooperating on PTPD: *the agent's decision procedure is implemented twice*. The action of the agent ("you") in Newcomb's problem has been predicted by Omega; in order for Omega to make an accurate prediction, *it must have a model of your decision procedure* which it uses to make its prediction. Omega "feeds" Newcomb's Problem to its model of your decision procedure. If you two-box, so did Omega's model; you'll find nothing in box B. If you one-box, so did the model and you'll find $1,000,000 in box B. This might be very counterintuitive to the reader, but it's no different than two identical calculators both returning 4 to 2 + 2. In that case, the "addition procedure" is implemented twice (once in each calculator); in Newcomb's Problem, its your decision procedure that's doubly implemented. Knowing that Omega's model will have made the same decision she does, the FDT agent therefore one-boxes and earns $1,000,000. (That is, unless she, for some reason, concludes Omega does *not* have a model of her decision procedure, and bases its prediction on something boring as your shoe color. In that case, the agent's decision procedure does *not* affect Omega's prediction, in which case two-boxing is better and FDT does that instead.)

Smoking Lesion

Like CDT - and unlike EDT - FDT smokes in Smoking Lesion. There is no model or otherwise "extra" implementation of the agent's decision procedure in this problem, and FDT (like CDT) correctly reasons that smoking doesn't affect the probability of getting cancer.

Subjunctive Dependence

While CDT is based upon causality, FDT builds on subjunctive dependence: two physical systems (e.g. two calculators) computing the same *function* are *subjunctively dependent* upon that function. In PTPD, the agent and her twin are both computing the same decision procedure; therefore, what the agent decides *literally is* what the twin decides; they are subjunctively dependent upon the agent's decision procedure. In Newcomb's problem, Omega runs a model of the agents decision procedure; again, what the agent decides *literally is* what Omega's model decided earlier, and the agent and Omega are subjunctively dependent upon the agent's decision procedure.

Parfit's Hitchhiker

> An agent is dying in the desert. A driver comes along who offers to give the agent a ride into the city, but only if the agent will agree to visit an ATM once they arrive and give the driver $1,000. The driver will have no way to enforce this after they arrive, but she does have an extraordinary ability to detect lies with 99% accuracy. Being left to die causes the agent to lose the equivalent of $1,000,000. In the case where the agent gets to the city, should she proceed to visit the ATM and pay the driver?

The above problem is known as Parfit's Hitchhiker, and both CDT and EDT make the wrong decision here - albeit for different reasons. CDT reasons that there's no point in paying once the agent is already in the city; paying now only causes her to lose $1,000. Once an EDT agent learns she's in the city, paying isn't evidence for the driver taking her: she's already in the city! Paying now only correlates with losing $1,000. Both CDT and EDT therefore refuse to pay. You may think this is rational, as the agent is indeed already in the city - but note that *any agent following CDT or EDT encountering this problem will never have been taken to the city to begin width*. Neither a CDT nor an EDT agent can honestly claim they will pay once they are in the city: they know how they would reason once there. Both CDT and EDT agents are therefore left by the driver, to die in the desert.

FDT once again comes to the rescue: just like Omega has a model of the agent's decision procedure in Newcomb's Problem, it appears the driver has such a model in Parfit's Hitchhiker. The agent's decision in the city, then, *is* the decision the model makes earlier (with a 1% error rate). FDT then reasons that the expected utility of paying is 0.99 x $1,000,000 - $1,000 = $991,000, because of the 0.99 probability that the driver predicted you would pay. The expected utility of not paying is only 0.01 x $1,000,000 = $10,000, as the driver would have to wrongly predict you'd pay. $991,000 is a lot more than $10,000, so FDT agents pay - and don't get left to die in the desert.

Transparent Newcomb Problem

> Events transpire as they do in Newcomb’s problem, except that this time both boxes are transparent—so the agent can see exactly what decision the predictor made before making her own decision. The predictor placed $1,000,000 in box B iff she predicted that the agent would leave behind box A (which contains $1,000) upon seeing that both boxes are full. In the case where the agent faces two full boxes, should she leave the $1,000 behind?

This problem is called the Transparent Newcomb Problem, and I suspect many people will be sure the rational move is to two-box here. After all, the agent *knows* there is $1,000,000 in box B. She *sees* it. In the original Newcomb's Problem, you don't see the contents of box B, and your decision influences the contents. Now the agent does see the contents - so why leave an extra $1,000 behind?

Because it was never about whether you actually see the contents of box B or not. It's about subjunctive dependence! Your decision influences the contents of box B, because Omega modelled your decision procedure. FDT recognizes this and one-boxes on the Transparent Newcomb Problem - and, crucially, *virtually never ends up in the situation where it sees two full boxes*. FDT one-boxes, so Omega's model of the agent one-boxes as well - causing Omega to put $1,000,000 in box B. Had FDT one-boxed, so would Omega's model, in which case Omega would leave box B empty.

Ask yourself this: "What kind of person do I want to be?" Do you want to be a two-boxer or a one-boxer? I mean, if I ask you now, before you even face the above dilemma, do you want to be someone who would two-box or someone who would one-box? Two-boxers never face the situation with two full boxes; one-boxers do, although (but *because*) they decide to leave the $1,000 behind. You can't be a "one-boxing person" in general and still two-box once you're actually facing the two full boxes - if that's the way you reason once you're in that situation, you're a two-boxer after all. You have to leave the $1,000 behind and be a one-boxing person; this is the only way to ensure that once you're in a situation like the Transparent Newcomb Problem, there's $1,000,000 in box B.

Counterfactual Mugging

> Omega, a perfect predictor, flips a coin. If it comes up tails Omega asks you for $100. If it comes up heads, Omega pays you $10,000 if it predicts that you would have paid if it had come up tails.

So should you pay the $100 if the coin comes up tails?

FDT pays in this problem, known as Counterfactual Mugging, and the reason might sound crazy: your decision procedure in this problem is modelled by Omega *in a counterfactual situation*! The coin *didn't* come up heads, but *had it* come up heads, you *would have* made $10,000 if and only if you pay up the $100 now.

Again, ask yourself: "What kind of person do I want to be?" Rationally, you should want to be someone who'd pay in case she runs across Omega and the coin comes up tails. It's the only way to get $10,000 in the heads case.

Concluding Remarks

This ends this sequence on Functional Decision Theory. The purpose of this sequence was to deliver an *intuitive* introduction; to fully explain FDT, however, we need to go deeper and dive into mathematics a bit more. My plan is to post a more "technical" sequence on FDT in the near future.