Hume: The Story of a One-Boxer

Josh Brekel

May 5, 2025

What follows is a rough transcription of an experience I had a few years ago. However improbable, the experience brought me face-to-face with none other than David Hume of Edinburgh. Our scene begins with me watching television alongside my college roommates, Terry the Train-lover and Oliver the Opulent. All of a sudden our living room is flooded with a piercing, neon-blue glow. Running out to our front yard, we encounter an extraterrestrial.

Florb. Greetings, Earthly fellas! Newcomb's Problem. Ever heard?

Terry. Who ... what ... who are you?

Florb. Florb, clearly. Now, Newcomb's Problem! Ever heard of it?

Me. I think someone mentioned that in my philosophy class the other day . . .

Florb. Right you are, Josh! It is a classic problem in philosophical decision theory. Reading about the problem amuses me. Indeed, it makes me properly chuffed.

Me. How do you know my name? What are you?

Florb. I am Florb, obviously! I come to pose Newcomb's Problem to you. I am fascinated with how you puny little humanoids reason, and with the art of Southern barbecue. But I have come in peace to put the problem to the test. I have chosen you as one of my test subjects! Be grateful. I have been watching your behavior, Josh. I have you down pat. I have you fine-tuned. I am now ready to predict what you may do.

Oliver. Alright, bucko.

Florb. FLORB!

Oliver. Fine Florb. Whatever. Chill out man. Speaking as Josh's friend and roommate, I'll just say this: we don't take lightly to solicitors. Fly off Florb boy.

Florb. There is money involved.

Oliver. Alright, you can stay. But just because of the money!

Florb. Excellent. Now, let me lay out the terms. Before you are two boxes. Box A is transparent, and you can see that it contains \$1,000. Box B is opaque. It contains either \$1,000,000 or nothing. You have two options: take only Box B, or take both Box A and Box B.

Me. Sounds like I should take both. I mean, if the million's there, I get a million and a thousand. If not, I at least get the thousand.

Florb. Ah, but here's the dry rub. Before placing the money, I predicted what you would do. If I predicted you would take only Box B and be a one-boxer, then I put the million dollars in Box B. If I predicted you would take both boxes and be a two-boxer, I left Box B empty. Also, in case it isn't clear: you express your intentions by which box you touch first. If you put your slimy little human digits on Box B first, then I interpret that as your intention to one-box. In that case, Box A is immediately beamed back to my ship. Alternatively, if you touch both boxes or touch Box A first, then I interpret that as your intention to two-box. In that case, you get to keep both boxes.

Terry. So wait. You've already made your prediction?

Florb. Precisely, Train-lover. I made my prediction weeks ago.

Terry. But if the boxes are already here, and you're done predicting, then who cares what you predicted?

Florb. A fine question. And yet my predictive accuracy is very, very high: 99.99%. I have made thousands of these predictions about Josh, and I am almost never wrong.

Oliver. Almost?

Florb. One time, I thought Josh was going to eat oatmeal for breakfast, but he had yogurt instead. It was a strange day.

Me. It was a Wednesday. I was out of oats. Well. Actually, my oats had weevils. Ugh. Weevils. You know Florb, you look a bit like a big purple weevil. You might want to get that checked out. You look a little sick, buddy.

Florb. Regardless, I have placed the boxes and made the prediction. I will not interfere again. The choice is yours.

Florb's Spacecraft. Snap. Crackle. Pop. Florb-be-gone!

I paused. It felt like a trick, but I couldn't see how. I stared at the boxes. One clear, one opaque. One certainly worth \$1,000. The other—a mystery. I tried to think it through. We need rent money. Majoring in philosophy doesn't pay.

Me. Let's do this properly.

I sat down on the lawn and crossed my legs. We had just started learning about Savage's expected utility theory in my Decision Theory seminar. I pulled that class textbook—Michael G. Titelbaum's *Bayesian Epistemology Volume 2*—out of my pocket, along with my trusty collapsible whiteboard. Slowly, I began to mumble things through.

Me. Hmm. Let's suppose the probability that Florb predicted I would one-box is p, and the probability he predicted I would two-box is 1 - p.

Savage says that you calculate expected utility of an act by multiplying the utilities of the act's outcomes by the probabilities of those outcomes. So let's just take Florb at Florb's word and assume this Florb fella is actually 99.99% accurate. That means:

- If I one-box, then with 99.99% probability, I get \$1,000,000.
- If I two-box, then with 99.99% probability, I get just the \$1,000 from Box A.

That gives us this decision table:

Florb Predicts \rightarrow			
$My\ action\ \downarrow$	One Box	Two Box	$m{E}m{U}_{SAV}$
One Box	\$1,000,000	\$0	$(1,000,000 \times .9999) + (0 \times .0001) = $999,900$
Two Box	\$1,001,000	\$1,000	$(1,001,000 \times .0001) + (1,000 \times .9999) = \$1,100$

So according to Savage's framework, I should one-box. That's weird. The table shows that two-boxing is clearly the strategy that produces a higher utility outcome no matter the state of the world. Two-boxing is definitely the dominant strategy, then. And I thought Savage's framework entailed dominance reasoning! What is going on here? Ugh. Decision theory.

Oliver sat down next to me, shaking his head.

Oliver. That's the wrong way to think about it. Remember they mentioned in class... Here let me find it...

Oliver ripped Titelbaum's book from my hands and started flipping the pages.

Oliver. Ah! Here it is. On page 257, Titelbaum says, "The Dominance Principle—and Savage's expected utility theory in general—breaks down when the state of the world is influenced by which act the agent performs. Savage recognizes this limitation, and so requires that the acts and states used in framing decision problems be independent of each other." So Savage's expected utility theory can't be used in this case. The whole setup is that your choice is evidence about the past—it is evidence about what Florb predicted. Your choice—the act you perform—influences the state of the world. You should use this Jeffrey guy's decision theory. Titelbaum says it was designed to handle these sorts of cases. Jeffrey ...he's like the Ethereum of decision theory.

Me. By the dog, Oliver! You are right. Gah! I hate it when you are right.

Oliver. Yeah, yeah, yeah. In any case, if you decide to one-box, then you should assign high probability to the million being in the box. If you decide to two-box, then you should assign high probability to it being empty.

Me. But that's backwards. I haven't even chosen yet. How can I update on something that hasn't happened?

Oliver. Jeffrey uses conditional probabilities to let you represent uncertainty about your own future actions. And you already said we're taking this Florb guy his word, so we know he's pretty much gonna be correct. I actually feel certain he will be correct. So here is how you would update the decision table:

Florb Predicts \rightarrow			
$My\ action\ \downarrow$	One Box	Two Box	$m{E}m{U}_{EDT}$
One Box	\$1,000,000	\$0	$(1,000,000 \times 1) + (0 \times 0) = \$1,000,000$
Two Box	\$1,001,000	\$1,000	$(1,001,000 \times 0) + (1,000 \times 1) = \$1,000$

Oliver. To get rich quick, you should definitely one-box.

Me. That's ... weirdly convincing. Oliver, usually you only say smart things about sports. Everything else you say is about the stock market, that meathead Joe Rogan, or crypto. What gives?

Terry. Hmph. Not so sure he's being smart even this time.

I turned to Terry. He'd been quiet this whole time. That wasn't like him. Usually, if you said "train" three times in a row, he'd appear out of thin air with a whistle and a historical diagram of the Trans-Siberian Railway. But now? Pretty much nothing. He was not amused. He was not chuffed.

Me. Okay, smart guy. What do you think?

Terry. I think you're both being silly. The money's either in the box or it isn't. You said it yourself. Two-boxing is best no matter what he predicted. And the weird weevil guy is gone. It's not like he's up there fiddling with a remote control.

Oliver. How do you know?

Terry. Because otherwise it wouldn't be a proper philosophical example! You can't test what kind of agent you are if some weevil-faced alien is watching from orbit changing things after you decide! I mean, that possible world certainly exists, but it sure ain't the actual world. The actual world is this one, and he ain't changing the contents of the box here.

Oliver. Ah! Not again. Shut up about possible worlds.

Terry. Ollie! They exist! I am telling you!

Oliver. But as the great Sir Karl Popper said, a hypothesis that is not falsifiable is pretty much meaningless mumbo jumbo.

Terry. That is not even an accurate paraphrase of what he said. You should practice reading.

At that exact moment, the decorative landline in our front yard rang. It wasn't even plugged in. We all stared at it. Hesitantly, I picked up.

¹You have to be *really* careful thinking about trains while in the shower.

Me. You ... reached the residence of two unemployed philosophers and one employed cryptobro.

Florb (on speaker). It seems like you fools are getting sidetracked, just as I expected. To clarify: I am not changing the box contents after your decision. That would ruin the philosophical example. I am a principled purple person. Goodbye.

Click.

A small puff of teal and gold glitter shot out from the receiver. It smelled faintly of Marshmallow Mateys, Windex, and the Belgian techno-anthem, "Pump Up the Jam."

Terry. See! The contents of the box are what they are. Nothing you do now can change that. So you might as well two-box.

Oliver. Terry, you need to stop. We can't keep donating plasma forever.

Terry. Okay look: the contents of that box are already fixed. We know that. Nothing we do now can causally affect what's in there. It reminds me of the extortionist James Joyce talks about.

Oliver. The weird Irish guy?

Terry. No, not the Irish novelist, though he'd probably extort us too if he needed rent money. The philosopher James Joyce. From Michigan. You know?

Oliver and I stared blankly at Terry. He always read more than we did.

Terry. Alright, listen. The guy in this case says something like this to Joyce, "People in this area like to smash windshields, but if you pay me ten bucks, I'll protect your car while you eat." Of course, the idea is that if you don't pay him, this very guy will smash your windshield. Joyce says he actually experienced this while he was in Ohio, by the way. Ohio. What a crazy place. Anyway, so this guy offers his "protection" over your car. And then what? Expected utility theory says you shouldn't pay the guy because that is the dominant act? That's lunacy. Not paying causes your windshield to get smashed, almost like you smashed it yourself. So dominance reasoning can definitely lead you astray, at least in some cases. But this case isn't like that. If you one-box, you're letting yourself be pushed around by

prediction correlations instead of actual consequences of causal relationships. You need to use causal decision theory. You need to think about actual causal relations in the world, and you need to think about causally motivated dominance reasoning. Sure, the inclination to two-box might be evidence there's no money in the box. But the decision to two-box doesn't cause anything! That would be backwards causation.

Oliver. Terry, this kind of reasoning is why we're broke. Remember when you said the rent money was already determined, so there was no point checking your Venmo balance?

Terry. That's not—okay, so that was an unrelated mistake in causal modeling, and it is beside the point here. All I am saying is that if your theory justifies backwards causation, that's evidence your reasoning is backwards. You are being silly.

Oliver. No, you are being silly! This is perfectly in line with the point! I have evidence that your track record sucks.

Me. Guys, guys. Calm down. We have time. Just don't touch those boxes.

Terry. I mean, they fell on my car. We need to move them before I take Ollie to work tomorrow.

Moments later, the sky split open. Florb descended in his spacecraft, like yolk from a cracked egg. Florb now wore sunglasses and blared a bebop remix of some classical music.²

Florb's spacecraft. Pop. Crackle. Snap. Florb-be-here!

Florb. Made your decision yet?

Me. Of course not. This is downright paradoxical!

Florb. I thought you might say that. Here, I brought you this.

Me. Uh, what is this? My mom taught me not to take things from strangers.

Florb. I AM FLORB, NOT STRANGER!

Me. Geesh, okay, fine. A . . . remote control? Florb and his bag of tricks.

²I later learned it was Beethoven's Ninth.

Florb. No tricks here. It's a temporal dial-back device. It lets you travel through time. Specifically, back to speak with your favorite philosopher. It is already tuned for your choice.

Me. Why can't I travel forward and just find out what I chose?

Florb. That would ruin the philosophical example. Clearly. Now, who do you want to visit?

Terry. Wait! You can time travel? How does that work? And how do we know that you, FLOORRBBB, haven't traveled to the future to see what Josh picks?

Florb. Precisely because that would also ruin the philosophical example. You just really don't understand thought experiments, do you?

Me. See, that is the thing about thought experiments, Florb. They are meant to stay in your noggin.

Florb. Bah humbug. In any case, I am here entirely out of curiosity. You think I would sacrifice that just to mess with you? Who do you think I am? Oliver? No! As with all of my test subjects, I have obscured Josh's future from my vision so I could satisfy my curiosity. Plus, forward time-travel violates the terms of my InterGalactic Institute of Philosophical Musings Research Grant. I don't want to lose funding.

Terry. Fine, that makes perfect sense.

Oliver. It does? What about—

Florb. Anyway! Who do you want to visit?

Me. Uh, I quess ... David Hume?

Florb. I knew it! You'll love this guy. His custom demands it.

I pressed the button. The air rippled like heat over stone. The ground tilted sideways—or maybe time did. Then: a cold breeze, cobblestones, and a faint clip-clop of hooves.

Me. Oh boy. Scotland?

A stout man in a powdered wig and a stylish red overcoat stumbled backward from a public house, eyes wide. He was saying something about invisible hands to one Mr. Smith. I tapped him on the shoulder.

Hume. Good God, man! You've near scared the soul from me! What contraption is this?

Me. Sorry, ... Mr. Hume, I take it. I—I'm from the future. I need your help.

Hume. The future, you say?

Hume looked me up-and-down. At that moment, I felt naked before the light of reason. In actuality, I was just some dude wearing 21st-century clothes in the 18th-century. The odor of horses overwhelmed my senses.

Hume. Well, I've always maintained an instinctive belief in the uniformity of nature. Lacking truly rational reason for such belief, I suppose I must allow for surprises to still crop up from time to time. Come, warm yourself, man. Talk slowly.

We entered the tavern. Hume patted Mr. Smith on the back as he passed, mentioning that he has a story to tell him. Hume ordered two drinks (one for him, one for "his curiosity"). I tried to explain Newcomb's Problem—or rather, my Florb problem—as best I could. Hume furrowed his brow.

Hume. Let me see if I understand. There is a being—a godlike predictor—who has filled one box or both based on its expectation of your actions?

Me. Exactly. And the question is: should I take only the one box, or both?

Hume. Actually, I think the question you should be asking is: what in the devil is wrong with this being?

Me. You know, Mr. Hume, I—

Hume. Please, Mr. Hume was my father. Call me "Davey."

Me. Uh, okay. Davey. Thing is, I need money for my rent. So, here. Maybe this will help.

I drew the decision theory matrix and explained the problem again.

Hume. Hmm. A curious puzzle indeed. But tell me—on what grounds do you say your choice has no effect on what's in the boxes?

Me. Well, that's the point of the problem. Florb already made his decision. My choice can't change the past.

Hume. Change the past? Surely not. But you mistake yourself if you think causation runs on such firm rails. You say your act cannot influence what is already done—but influence, sir, is not something observed. It is something inferred. And what do we infer it from?

Me. Intoxication?

Hume. Intoxication? What? Me? No, Never. Regularity. Regularity is that from which we infer influence.

Me. So you're saying...?

Hume. I am saying that the connection between events is not known through reason, but through custom and habit.

Me. But what custom? What habit? This is the first I've ever heard of actual alien encounters, let alone repeated ones.

Hume. Ah, fair point lad. But I gather, from your telling, that others have faced this Florb creature before you? That its predictions have proved reliable?

Me. Yeah. From what Florb says, Florb is almost always right. One time Florb incorrectly predicted my breakfast, but that was because of the weevils.

Hume. Weevils, huh? Tasty little bugs, those. In any part, you have your regularity, then! Perhaps not your own, but passed to you by testimony, by the accumulated experience of others. Surely that may stand in for direct observation. Our trust in most things—history, geography, medicine—rests on the testimonials of others.

Me. I guess that makes sense. But how does that help in this case? I'm confused. I still only face the problem once.

Hume. Suppose you have a die with one thousand sides. On 999 of the sides, the die is marked with two spots. On the last remaining side, the die is marked with three spots. If you were to roll this die, then what would you expect to happen?

Me. Well, I would be shocked if it landed on the side with three spots. I should expect to roll a two.

Hume. There you have it! Imagine facing your situation one thousand times. The mind may build its expectations even from

the fictions it entertains. Suppose you had that repetition—would you not begin to expect a full box when you one-box, and emptiness when you two-box?

Me. Yeah, I think I would.

Hume. There it is: expectation. That feeling is what you call belief. Not a matter of logic, but of the vividness, stability, and coherence of an idea. Over and again, the act of one-boxing immediately precedes great profit. Thus, you feel that one-boxing leads to gain. That feeling, sir, is your sole compass. Where the past is entirely uniform, we cannot help but to expect the future to accord with the usual outcome.

Me. So... you're saying I should follow the feeling? There isn't more than a feeling?

Hume. Boston? Heavens no! I am saying: reason is inert. You lack demonstration or proof in this case, but you possess high probability. Through the inexplicable contrivances of nature, that high probability generates a great sentiment of belief. Belief moves us. If the idea of one-boxing strikes your mind with more force—because it aligns with the pattern, whether lived or imagined—then you believe it will lead to profit. And if you believe that, what else could you do?

Me. I think I get it. Even in a one-off case, I should treat the decision as part of a larger pattern.

Hume. Precisely. There are no necessary connections, only habits of thought. You may not see the tie between choice and outcome—but if your instinct, your custom, your inherited expectation tells you they go together, that's all the world can offer. We never actually see the connections between our choices and outcomes. Yet, custom and habit demand us to posit such a connection. Though that isn't certainty, it's enough to bet your rent. What difference is there, I ask, between a god who places coins in boxes and the sun rising each morning, save the frequency of their habits? From those thoughts, you observe that when people one-box, they bring in the dough. That is all the ground we have for talk of "causation" at all.

Me. Bring in the dough? I...But causal decision theory says—

Hume. Bah! You philosophers of the future! Still building castles in the air. At least ours were made of stone. Causal decision theory presumes we know the "real" cause. But all we ever see are patterns. There is no impression of necessity in any single instance—only the mind's expectation after repeated conjunction. Your fancy modern theories make ghosts of causes. I trust custom more than invisible metaphysics.

Me. So I should one-box?

Hume. If you believe in evidence. You say this Florb creature is near-perfect. Then experience tells you: one-box means full box. And I have always said, 'tis not reason which guides life, but a kind of instinct born of observation. Your two-boxing friend, I suspect, still dreams of metaphysical wires between act and effect. I deal in appearances.

Me. So you think it's rational to one-box, even though the money is already in or out?

Hume. Rationality, sir, is but a slave to the passions. And our beliefs are guided not by certainty, but by the liveliness of our ideas. If the idea of one-boxing strikes your mind with force, then it is belief. And belief is what guides us.

Me. That's ... surprisingly helpful.

Hume. If it helps, I must have misspoken.

Me. I have no clue what that means.

Hume. In any case, it is time for my game of backgammon with Mr. Smith. Off with you, lad. You've got a box—you hear? a box—to choose. And if ever you find yourself in need of wisdom or whisky, you know where to find me.

Me. Ah, drat. Florb didn't say how I get back.

Hume. Barkeep! Bring one of my philosopher's stones for this wanderer!

The bartender rummaged through a drawer and then placed an object in Hume's hand. He tossed it to me. It was a stone with a strange shimmer.

Hume. Just rub this in your hand, press that button, and think about breakfast. Or lunch. Or dinner. Time's more obliging

than we tend to suppose.

Me. Has this ... Has this happened to you before?

Hume. Let's just say you're not the first bright-eyed young philosopher to wander through here. Now go on. Before causation reasserts herself.

Following Hume's instructions, I clutched the stone. The world started to shake, rattle, and roll.

Me. Oh! Davey! I just remembered: Your argument against miracles is not goo—

Before I got the word out, I was back in my front yard. The party moved inside. Opening the door, I see our house in disarray. Oliver is standing on the dining room table, gesturing wildly. Terry is shaking his head and basting a chicken. Florb is sniffing and licking—but not eating—a Pop-Tart.

Oliver.—and I'm telling you, if Florb made the prediction before the causal fork, then the observation can't retroactively influence the outcome!

Terry. Only if you assume temporal locality. But Florb is clearly operating with a non-classical decision-theoretic model—maybe even a fixed-point oracle!

Florb (between licks). Your hypotheses amuse me. Please, continue.

I cleared my throat. They all turned to me.

Me. What did I miss?

Oliver. We were debating how Florb's prediction mechanism fits with causal decision theory. Turns out he lets us look in the boxes. Just not you.

Florb. Yup.

Terry. Which—I mean. It is fair. He says it doesn't influence the outcome, and honestly, I believe him. It's elegant. Also, Florb wants to try chicken.

Florb. Yup. I invite them to observe, cook, and to advise. I do not deceive.

Me. So ... what's in the boxes?

Oliver. Doesn't matter.

Terry. Yeah, doesn't matter.

Me. But guys, what should I do? Davey says I should one-box, but I am not sure what to think. He smelled of Scotch.

Florb & Oliver & Terry (unison). Davey?

Me. Never mind. What should I do?

Terry. You should two-box. Respect trains. Respect causation. Respect trains of causation.

Oliver. This time, I agree with Terry. After seeing the contents of the box, the evidence speaks for itself.

They both looked at me. I glanced at the boxes. Then at Florb. Then at my friends. They glanced at the boxes. Then at each other. Then at me. Then we all looked at Florb. I looked at myself in the mirror. They looked at me.

Against the advice of David Hume, I touched both boxes at the same time.

I two-boxed. I became a two-boxer.

What was inside? Well, that's the thing about philosophical puzzles. Sometimes the answer doesn't matter as much as the choice. The answer certainly doesn't matter as much as the reasoning—the story—that brought you there.

Florb later decided to run the experiment on Terry and Oliver. I think he probably just liked hanging out with them. I know I did.

Terry chose to go back and talk with his favorite philosopher, David Lewis. He says they mostly talked about trains because two-boxing was so obvious to both of them. Now, Terry drives a garbage truck by day and works on the metaphysics of causes at night. We get together about twice a year for old time's sake.

Oliver, on the other hand, asked Florb if he could go hang out with—and I quote—"The founding father of capitalism, Ayn Rand." Florb said that the dial-back device only works for *real* philosophers. Apparently they argued for a long while. Finally, Oliver capitulated and said he would ask Joseph Campbell what he thought. Again, Florb said that "Wasn't a real philosopher." I am not sure where Florb got his philosophical prejudices, but I know they didn't sit well with Oliver's "Philosophy of the bros, by the bros, and for the bros," motto. Frustrated, Oliver muttered something like, "Fine, Seneca or whatever."

He almost died. Turns out Seneca didn't speak a lick of English, and Oliver didn't know any Latin beyond *cogito ergo sum* and *reductio ad absurdum*. He, of course, one-boxed when he made it back. Last I heard, he was sailing around the world with his millions. I don't know how he is doing that, though. There was only *one* million in the box. Boats are expensive.

Later, I found out Florb was detained by the M.I.B. Border Patrol for trying to smuggle cattle off-world. Apparently he lost his research grant and is on interstellar probation until 2128. I miss Florb. He cooked a mean brisket.

As for me, I started a PhD in philosophy. I am still thinking about the choice. But I never said what was in the box.

EPILOGUE

Back in the Edinburgh tavern, a log cracked in the hearth. Mr. Smith set down his backgammon die and looked at his companion.

Smith. What did he mean by "your argument is not goo"? No arguments are goo. That's the entire point of arguments.

Hume. Quite right, Adam. If any of my arguments were goo, I'd have eaten them by now. More likely, he meant "good."

Smith. Ah! Yes, that would make more sense. So, your argument against miracles is not good?

Hume. Apparently not. Though I must say, I find the accusation suspicious, coming from a man who vanished in a shimmer after discussing invisible aliens and transparent boxes.

Smith. Perhaps he had a miracle in mind.

Hume. Then I win.

Smith. How's that?

Hume. If he believes in miracles, and I argue they lack credible testimony, and he shows up as one—well, he disproves his own point. I would be skeptical of him. As is right.

Smith. Davey, that's circular.

Hume. Naturally. Like most habits of thought.

Smith. So what will you do?

Hume. Perhaps I shall revise my argument. Loosen it up a bit. Less goo, more room for cosmic interlopers. But only the charming ones.

Smith. I am still puzzled. I liked that argument.

Hume. Perhaps he meant to say, "Your argument is

not good, but it is persuasive." In which case, I shall take it as a compliment disguised as an insult. The highest form of praise in philosophy.

Smith. Or perhaps he meant: your argument is not the good—merely a good.

Hume (smiling). Well now, Adam, don't get all metaphysical on me. That path leads to Plato, and I've only just recovered from Descartes.

They laughed. The fire popped. Outside, horses clopped along the cobblestones, unaware of Florbs, boxes, or any decisions at all.