Newcomb University: A Play in One Act
Adam Elga
February 26, 2019

Scene 1

Professor Causey (in full tweed) and Ms. Neutra (a potential donor) walk down the hall of a gleaming research building.

Causey: ...and so you see that our experimental causal decision-making lab has state-of-the-art facilities which—pardon me for a moment.

Causey reaches down to what looks like an electrical outlet along the wall and sticks his finger in the socket. A buzzing sound is heard.

Causey: Ouch! Anyway, as I was saying, the facilities are excellent—

Neutra: Sorry to interrupt, but are you alright?

Causey: Oh, that. I’m fine. Just a modest electrical shock from putting my finger in the outlet. No permanent damage. Ah, I see we are approaching another outlet. Excuse me.

Causey puts his finger in another outlet and another buzzing sound is heard.

Causey: Yowza!

Neutra: Why do you keep putting your finger in those outlets?

Causey: To reduce the amount of electrical shocking I experience, of course.

Neutra pauses.

Neutra: I’m confused. It looks like every time you put your finger in an outlet, you get painfully shocked.

Causey: Yes.

Neutra: But I and the other passers-by who don’t touch the outlets don’t get shocked.

Causey: Yes.

Neutra: And you expect that pattern to continue in the future?

Causey: Yes.

Neutra: So why stick your finger in?!
Causey: Didn’t I tell you a moment ago? To reduce getting shocked. Or perhaps you don’t know that we have special outlets in this building? They are called Libet\textsuperscript{1} Ahmed\textsuperscript{2}-Newcomb\textsuperscript{3} outlets, LANs for short. Here’s how they operate. Each outlet has a brain scanner. Whenever someone approaches the outlet, the outlet scans her brain to determine whether she is in a brain state likely to lead to her deciding to stick her finger in. If so, the outlet activates program A. Otherwise it activates program B.

Neutra: What do those programs do?

Causey: Program A has the outlet deliver a shock upon being touched, but if it is not touched then it remotely delivers a slightly stronger shock to the person who was scanned. Program B has the outlet never deliver any shocks, but has it play a pleasing melody if someone touches it. Here’s a decision table for you:

<table>
<thead>
<tr>
<th>Program A</th>
<th>Program B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(LAN predicted touch)</td>
<td>(LAN predicted no touch)</td>
</tr>
<tr>
<td>no touch</td>
<td>worse shock</td>
</tr>
<tr>
<td>touch</td>
<td>regular shock</td>
</tr>
</tbody>
</table>

As you can see, if the outlet is running Program A, one is better off touching it (to get a regular shock rather than a worse shock). If the outlet is running Program B, one is also better off touching it (to hear the nice melody). And by the time one makes one’s decision, the outlet has irrevocably chosen its program. So one is guaranteed to do better by touching the outlet.

Neutra: Fascinating.

Causey: Yes, the original funder of the building, Dr. Newcomb, had some eccentric requirements. Ah, I see we are approaching an outlet. 

Reaches down and is zapped by it. Auggghh! I suppose that now that you know how these outlets operate, you’ll be sticking your finger in?

Neutra: No thank you. Walks past the outlet without incident.

Causey: Oh. Quite irrational of you. A shame, really—you missed out on hearing a wonderful melody.

Neutra: You’ve heard it?

Causey: Actually, no. The brain-scanners in these outlets are infernally accurate (though not perfectly accurate, I’m told). To my knowledge, no one has ever heard the melody. And no one has ever been shocked by the stronger shock.

Neutra: But you expect one of those things to happen soon?

Causey: No.
Neutra: I see. Pauses. Is it not a workplace safety issue, to have outlets around with the potential to remotely shock people? I notice that other students, not to mention children and some pet cats, wander through the building.

Causey: We were worried about that, but it turned out to not be a problem. The cats rarely touch the outlets, and when one does, it gets shocked and takes care to not touch outlets in the future. And incoming freshmen are informed about how the outlets are designed. But they are by and large irrational—they don’t put their fingers in. (Haven’t had the benefit of my Advanced Causal Decision Theory course yet, ha ha!)

Neutra: So children, animals, and freshmen are safe and don’t get shocked. It is just shock-avoidance experts such as yourself who stick their fingers in the outlets and receive shocks.

Causey: Yes.

Neutra: Blinks.

Causey: The strange thing is, so far as their outward appearance and observed behavior goes, these LANs look just like ordinary electrical outlets. In fact, only last week I learned that the outlets one floor up from here are ordinary outlets. Turns out I had been sticking my fingers in them completely unnecessarily! Now I know not to do it.

Neutra: Thanks very much for the tour.

Scene 2

It is one week later. Neutra enters a building marked “Evidential altruism lab” and knocks on an office door.

Evidentia: Come in. Ah, Ms. Neutra. Nice to meet you, I am Professor Evidentia. Welcome! I take it that last week you met Professor Causey and heard about his . . . theories?

Neutra: Yes. He kept shocking himself. Shudders

Evidentia: So sad. Anyway, you’ll find none of that today. Here at the evidential altruism lab we always choose so as to minimize the expected number of shocks people receive.

Neutra: That is a relief. Altruism sounds good.

Evidentia: Speaking of altruism, I will now subject you to a painful shock. Takes out an electric stun gun and prepares to fire it.

5Neutra’s comment here echoes the so-called “Why ain’cha rich?” objection to causal decision theory (Lewis (1981b), Joyce (1999, 151-4), Bales (2017 §3.1), Ahmed (2018)).
**Neutra:** Wait! Stop! What are you doing?

**Evidentia:** Shocking you. For your own sake, of course. To reduce the expected number of shocks you receive.

**Neutra:** Pause. Wait, I get it. You mean that shocking me now will prevent me from getting some even worse shock in the future? If so, then I’m sorry to hear about the whole situation, but I guess I’ll just grit my teeth and bear this one shock for the sake of its future good effects.

**Evidentia:** Adjusting the stun gun. No, you’ve misunderstood. The shock I’m about to administer won’t cause any good effects at all, to you or anyone else. Its only relevant causal consequence will be the pain you feel.

**Neutra:** No good effects? Pure pain for no reason? How can this help me?

**Evidentia:** Perhaps Professor Causey didn’t explain all the features of our building. Do you see that device? *Points up to what looks like a smoke detector on the ceiling.* That is one of our Decision Prediction Units (DPUs). Last week, I noted in my calendar that you would be visiting me today. In response, the DPU scanned our brains to predict whether I would shock you. If it predicted that I would shock you, it left you alone last week. But if it predicted that I would refrain, then it ensured that last week you received twenty shocks, each as painful as the one delivered by this stun gun. Here’s a decision table for you, below. I’ve crossed out those outcomes I’m almost certain don’t obtain:

<table>
<thead>
<tr>
<th>DPU predicted shock</th>
<th>DPU predicted no shock</th>
</tr>
</thead>
<tbody>
<tr>
<td>shock today</td>
<td>1 shock total</td>
</tr>
<tr>
<td>no shock today</td>
<td>0 shocks total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>21 shocks total</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 shocks total</td>
<td></td>
</tr>
</tbody>
</table>

**Neutra:** That is horrifying! But there’s nothing to be done about shocks I may or may not have received last week, so that information is not relevant to your current decision.

**Evidentia:** Actually it is highly relevant. For I know that the DPUs are excellent predictors. So (1) if I shock you now, you will almost certainly end up experiencing one shock in total. And (2) if I refrain now, you will almost certainly end up experiencing twenty shocks in total (delivered last week). So shocking you reduces the total number of shocks I expect that you receive. Rationality, altruism, and simple kindness all demand that I shock you.

---

6These are what might be called “Horgan conditionals”, following Horgan (1981) (though note that Horgan (2015) prefers an argument for one-boxing that does not employ such conditionals).

7If it is thought that “thank goodness that’s over” (Prior (1959), Parfit (1984, 165-7)) considerations enter here, we may suppose that receiving 20 shocks has severe long term after-effects, so that Ms. Neutra would prefer (and Professor Evidentia would altruistically prefer on her behalf) to be about to receive 1 shock over having received 20 shocks already. If concerns about harming someone without permission are thought to be relevant, we may suppose that receiving 20 shocks is such a harmful outcome that it is morally permissible to shock someone once without their consent in order to save them from being shocked 20 times.
**Neutra:** Well, (1) and (2) are not both true. Furthermore I assure you: shocking me now cannot help me!

**Evidentia:** Well, from your point of view I can see why you are saying that. But you must understand that as an evidential altruist, I must act in the light of my evidence.

**Neutra:** OK, I see where this is going—I'm pretty sure what you are going to decide. But let me try to reason with you anyway.

**Evidentia:** Feel free.

**Neutra:** You admit that I have relevant evidence that you lack (memory of whether I was shocked last week)? That I prefer that you not shock me? And that I would

---

8 Here Ms. Neutra assumes that she may in conversation correctly deny a conditional that is false when evaluated with respect to her own evidence. For influential examples illustrating how one's evidence interacts with the conditional claims one may assert, see Gibbard (1981).

9 Ms. Neutra believes what she has been told about the setup and also recalls whether she was shocked 20 times last week (spoiler: she wasn’t). As a result, she is already confident that Evidentia will decide to shock her. Given this, why would she bother trying to persuade Evidentia? That of course depends on further details of the case, but under at least some assumptions Neutra can gain a (small) advantage by trying to persuade—even according to evidential decision theory. Here is a toy model to illustrate this: Neutra’s current credence function is the result of conditioning her prior $P$ on $K$, the truth about what happened to her last week. $P$ counts it as certain that Evidentia’s description of the overall setup is correct. In particular, $P$ counts it as certain that: Neutra got 20 shocks last week iff the DPU predicted that Evidentia would decide not to shock Neutra. Let $C$ be a random variable concerning Evidentia’s dispositions to choose, where $C = 1$ means that Evidentia will shock regardless of what Neutra says, $C = 2$ means that Evidentia will refrain regardless of what Neutra says, and $C = 3$ means Evidentia will shock unless Neutra argues as she does in the text. For $i = 1, 2, 3$, let $C_i$ abbreviate ‘$C = i$’, and assume that $P(C_i) = 1/3$. Let $E$ be the proposition that the DPU’s prediction is in error, and $R$ be the proposition that Neutra argues as she does in the text. Assume that $P$ counts $\{C, E, R\}$ as mutually independent (when $E$ and $R$ are considered as indicator random variables), that $P(E)$, $P(R|K)$, and $P(\neg R|K)$ are each greater than 0, that Neutra’s disutility is proportional to the number of shocks she receives, and that the only alternative to trying to persuade is to do nothing. Then the evidential expected value of attempting persuasion ($V(R)$) is (very slightly) higher than that of doing nothing ($V(\neg R)$) since under these conditions:

$$V(R) = (-1) P(\text{Neutra is shocked exactly once in total}|KR)$$

$$= - P(RC_1E)$$

$$= - \frac{P(RC_1E)}{P(RC_1E) + P(RC_2E) + P(RC_3E)}$$

$$= - \frac{P(R)P(C_1)P(E)}{P(R)P(C_1)P(E) + P(R)P(C_2)P(E) + P(R)P(C_3)P(E)}$$

$$= - \frac{1 - P(E)}{1 + P(E)} + \frac{1 - P(E)}{1 + P(E)/2} = \frac{(1/2)[P(E)] + (1/2)[P(E) + P(E)]}{(1/2)[P(E) + P(E) + P(E)]}$$

$$= - \frac{P(R)P(C_1)P(E) + P(R)P(C_3)P(E) + P(R)P(C_2)P(E)}{P(RC_1E) + P(RC_2E) + P(RC_3E)}$$

$$= (-1) P(\text{Neutra is shocked exactly once in total}|KR) = V(R).$$

Thanks to Arif Ahmed and Adam Kolber for pressing me on this point.
continue to do so even if you shared with me all of your relevant evidence?

**Evidentia**: Of course.

**Neutra**: You agree that we share the goal of minimizing the number of shocks I experience?

**Evidentia**: Yes.

**Neutra**: So if you shared all your evidence with me then someone with your goals and strictly stronger evidence would prefer that you not shock me. Doesn’t that count for something in this case?[^10]

**Evidentia**: No. *Starts to pull the trigger.*

**Neutra**: Wait! This is a long-shot, but let me give you my evidence. Then you might not think shocking me will help me. Last week what happened was—

**Evidentia**: *Puts fingers in own ears.* LA LA LA LA! I’m not listening to you! LA LA LA—

**Neutra**: What are you doing?

**Evidentia**: *Removes fingers from ears.* Sorry, I had to stop you from telling me about what happened last week. If you had told me (regardless of whether you told me that you were shocked or that you weren’t shocked), I would have come to think that shocking you now would serve no purpose, and indeed would be cruel and inhumane. So I wouldn’t have shocked you. So the DPU would have last week predicted that I wouldn’t shock you, and would have ensured that you were last week shocked 20 times. So for your sake, I can’t let you tell me what happened last week.[^11]

**Neutra**: *Stunned silence*

An administrator bursts into the room.

**Administrator**: There has been a mistake! Last week the DPU did not scan Professor Evidentia. It accidentally scanned Ms. Neutra instead! If it predicted that Ms. Neutra would today decide to give a shock to Professor Evidentia, it left everyone alone last week. But if it predicted that Ms. Neutra would refrain today, then it ensured that last week Professor Evidentia received twenty shocks.

**Neutra**: So I must now decide whether to shock her? *Points at Evidentia.*

**Administrator plucks the stun gun out of Professor Evidentia’s hands.**

**Administrator**: Yes. *Hands the stun gun to Ms. Neutra and exits.*

**Neutra**: I don’t want to shock anyone.

[^10]: Compare Gallow (2019, 4).

[^11]: Compare Wells (forthcoming, Section 5) and Gallow (2019, 4), which also make the point (about an analogous situation) that evidential decision theory would enjoin Evidentia to inflict a small additional punishment on Neutra, if that is what it took to keep Neutra from revealing whether she was shocked the previous week.
Evidentia: But if you are to reduce the number of shocks you expect me to receive, you must.

Neutra: This is difficult for me, but I see where you are coming from. If I shock you now, you’ll almost certainly receive a total of just one shock. But if I refrain, you’ll almost certainly receive a total of 20. So if I’m rational and kind, I should shock you. And I shouldn’t let you tell me whether you have been shocked last week.

Evidentia: Now you’ve got the hang of it!

Neutra: OK, if you want me shock you— Points the stun gun at Evidentia.

Evidentia: Hang on, I don’t want you to shock me.

Neutra: Lowers stun gun. Why not? Aren’t we both agreed that the rational, kind thing for me to do is to shock you, for your own good?

Evidentia: Yes, yes, that is all true. But . . . I prefer that you don’t shock me!

Neutra: Why?

Evidentia: Shocking me (slightly) increases the expected number of shocks I suffer.

Neutra: Auuggh! We already agreed I should think: shocking you reduces the expected number of shocks you suffer.

Evidentia: Well, from your point of view I can see why you are saying that. But you must understand that as an evidential altruist, I must prefer and hope in the light of my evidence. I hope that you act irrationally and/or unkindly and refrain from shocking me\footnote{For reasons similar to those described in footnote\textsuperscript{9}, Evidentia is already confident of what Neutra decides, but might count trying to persuade Neutra as doing a tiny bit of good. As Arif Ahmed has pointed out to me, if side-betting is an option Evidentia will by her own lights do better to spend her breath arranging side-bets with Neutra that take advantage of Evidentia’s extra evidence about whether Neutra will decide to administer a shock.}

Neutra: Places gun on the table. Thanks very much for the tour.

\textit{Scene 3}

\textit{Several days later. Neutra, Causey, and Evidentia sip coffees at a restaurant.}

Neutra: I don’t mean to be rude, but you are both totally irrational! Professor Causey, please stop sticking your fingers in LAN outlets. And Professor Evidentia, please stop shocking people for the sake of reducing shocks they may or may not have received by the DPU a full week prior.

Causey: Wait, you think that both of us are irrational?
Neutra: Bonkers, in fact.
Causey: How odd.
Evidentia: Remarkable!

Neutra: Wait, surely almost everyone in your profession thinks that you two act irrationally. That is just common sense.

Causey: Actually, a substantial fraction of decision theorists hold views that say I have acted rationally—\[^{13}\]

Evidentia: —And a substantial fraction of decision theorists hold views that say I have acted rationally.\[^{14}\]

Causey: And almost no decision theorists hold views that say neither of us has acted rationally. Indeed, aside from inessential window dressing, the two choice situations are the same: If you think that one should avoid sticking one’s fingers in LAN outlets, you should also think that you should shock in the DPU situation—

Evidentia: —And if you think that one should refrain from shocking in the DPU situation, you should also think that one should stick one’s fingers in LAN outlets. So we ask: what decision rule do you endorse that says that both of us have acted irrationally?

Curtain.\[^{15}\]

References


\[^{14}\]See for example Ahmed (2014b), Horgan (2015), Jeffrey (1965). For a small-sample on going survey on how decision theorists divide between one-boxing and two-boxing in Newcomb’s problem, see https://philpapers.org/surveys/results.pl?affil=Target+faculty&areas0=1399&areas_max=1&grain=coarse a continuation of the survey described in Bourget and Chalmers (2014).

\[^{15}\]For helpful discussion and feedback thanks to Arif Ahmed, Adam Bales, Alan Hájek, Adam Kolber, Alex Meehan, Timothy L. Williamson, and participants in a graduate seminar (jointly taught with Boris Kment) at Princeton in Fall 2018. It should be noted none of the characters are modeled on particular people, and that the author does not endorse Neutra’s view.


