Newcomb's Problem, edited by Arif Ahmed

J. Dmitri Gallow †

Fifty years ago, Nozick (1969) introduced the philosophical world to a puzzle which he learnt from the physicist William Newcomb. With some superficial alterations, the puzzle goes like this: An evil genius provides you with an opaque envelope containing a cheque. This cheque is yours to keep. Now: the decision you face is whether to give *her* £1,000 or not. Remember, she's an *evil* genius, so you want to deprive her of as much money as possible.[1] So why would anyone think about giving her £1,000? Because: unbeknowst to you, several years ago, she took a scan of your brain and made a prediction about what you would do in this very situation. If she predicted that you would hand over the £1,000, then she placed a £1,000,000 cheque in your envelope. If she predicted that you would keep your money, then she only wrote you a cheque for £1,000. You know all of this. You also know that these predictions of hers are quite reliable. Given that you pay her £1,000, you're 60% sure that this is what she predicted you'd do. And, given that you don't pay her anything, you're 60% sure that this is what she predicted you'd do. (She's played similar games with thousands of others. About 60% of the time, she predicts correctly, and her reliability is independent of whether they paid or not.)

Presented with this puzzle, some advise you to pay the evil genius £1,000—doing so makes it more likely that she wrote you the cheque for £1,000,000! Others advise you to keep your money—you've *already got* the cheque, so paying her now accomplishes nothing! Within philosophy, the puzzle has attracted an incredible amount of attention, and for good reason. It sits at the intersection of a cluster of rich philosophical questions about the nature of instrumental rationality and its relationship to free will, predictability, probability, causation, success, restraint, planning, and much and more besides. Each of these questions is well covered in this new collection of essays, edited by Arif Ahmed. It deserves to be read by anybody interested in any of these topics.

Almost everything about Newcomb's problem is debated and controverted, and most of these debates and controversies are well represented in *Newcomb's*

† Thanks to Daniel Drucker, James Joyce, and Reuben Stern for helpful conversations and feedback.

1. I'll also suppose that you value each new pound of which she is deprived just as much as you valued the previous one—that is, I'll suppose that your values are linear in sterling. I'll continue to make this assumption throughout.

Problem. Some deny that cases with this structure ever actually arise. This position is taken up in José Luis Bermúdez's contribution. If you are at all like this reviewer, you will find this position initially surprising. As Ahmed notes in his introduction,[2] you may think that consequentialist voters face Newcomb's problem each election. If you find yourself going to the polls, this gives you evidence that others like you are also going to the polls, and therefore, it gives you evidence that your candidate will win, though you're *astronomically* unlikely to make any difference to the outcome of the election. Calvinists face a kind of Newcomb problem with each new temptation—for resisting temptation makes it more likely that they are predestined for heaven, though it was settled long ago whether their name is written, and nothing they do now can change that.[3] Bermúdez doesn't deny that people face these choices, but he does deny that these choices are Newcomb problems. The trick is that Bermúdez only treats a decision as a Newcomb problem if it is a case in which two prominent decision theories, *evidential* decision theory (EDT) and *causal* decision theory (CDT), make different predictions.

EDT, by the way, tells you to take whichever action gives you the best news about the world. It says to choose whichever action has the highest *conditional* expected value. Conditional on what? Conditional on the action being chosen. So, if Pr is your probability function, and V is your *value* function, specifying how desirable a proposition is, then EDT says to perform whichever act, A, maximises the expectation $\sum_i \Pr(S_i \mid A) \cdot V(S_i \,\&\, A)$, where $\{S_i\}_i$ is a set of mutually exclusive and jointly exhaustive states of nature.[4] CDT, on the other hand, tells you to take whichever action does the most to *improve* the world. It says to choose whichever option has the highest *unconditional* expected value. That is, it says to perform the act, A, which maximises $\sum_i \Pr(S_i) \cdot V(S_i \,\&\, A)$.

The standard—but by no means universal—understanding is that EDT advises you to pay the evil genius the £1,000, while CDT tells you to keep your money. And this is what Bermúdez denies. He thinks that, in realistic cases like this, EDT will advise you to keep your money, since, in realistic cases like this, you will be in a position to obtain all readily available information about what was predicted *before* making your decision. Then, the act you ultimately choose won't provide you with any new information about the prediction—or, more generally, $\Pr(S_i \mid A)$ will be equal to $\Pr(S_i)$ for all i—so that EDT and CDT will agree.[5]

2. Similar points are made in several other places—*e.g.*, Joyce (1999, p. 55).

3. Calvinists theology may be wrong, but this doesn't change the fact that Calvinists regularly face this decision, for Calvinists believe the theology.

4. I'm going to suppose throughout that the S_i are *causally* independent of how you act. I am using 'A' ('S_i') for both the act (state) and the proposition that you've performed that act (that the state obtains).

5. This is closely related to Eells (1982)'s 'tickle defense', which Eells understands as a

As I say, nearly everything about Newcomb's problem is controversial—even what Newcomb's problem *is*. This particular controversy hides beneath the surface of this collection of essays, as different contributors use the name 'Newcomb's problem' in inconsistent ways. (This will no doubt perplex newcomers to Newcomb.) As we've already seen, Bermúdez stipulates that you face Newcomb's problem only if EDT and CDT give different recommendations about how you should choose. In his contribution, James M. Joyce provides 5 conditions which your decision problem must meet in order for you to be facing Newcomb's problem. These criteria do imply that EDT and CDT will diverge in their recommendations, but Joyce's characterization is nonetheless logically independent of Bermúdez's. Picking up on ideas from Meek and Glymour (1994), Reuben Stern suggests in his contribution that, if you think you shouldn't pay the evil genius, then you should think that, maybe, the evil genius can't predict your choice.[6] Given Joyce's definition, if you think that your actions are or may be unpredictable in this way, then you are not facing Newcomb's problem.[7] In their contribution, Huw Price and Yang Liu argue that, in Newcomb's problem, you should regard a decision to pay as *causing* the evil genius to predict that you would pay. In that case, EDT and CDT would agree. So the decision which Price and Liu call 'Newcomb's problem' is definitionally not a Newcomb problem, as that term is used by Joyce and Bermúdez.

One of Joyce's contentions is that authors have not been clear enough about what decision problem they are discussing when they talk about Newcomb's problem and related decisions. After finishing this collection of essays, it's difficult to disagree. In part, these different uses of 'Newcomb's problem' are symptomatic of the wide variety of philosophical issues and interests raised by the puzzle. Different authors aim to address different philosophical questions when they discuss Newcomb's problem, and this leads them to understand the decision differently.

For Stern, the primary question he aims to explore with Newcomb's prob-

defense of EDT, not an argument against the existence of Newcomb problems.

6. How is the evil genius so accurate in their predictions, then? Stern holds that you should think that, to the extent that the norms of practical rationality apply to you, you are not predictable. So, to the extent that the norms of practical rationality apply to you, you are special—you are one of the few who have intended to not pay in a genuinely unpredictable way. (In the jargon: to the extent that the norms of practical rationality apply, you are one of the few whose decision to not pay is an *intervention*.)

7. Joyce's first criterion for counting as a Newcomb problem is: "For each possible action A there is an associated 'type-A' brain state, and you are subjectively certain that you will do A at t_1 iff you occupied the type-A state at t_0". The second condition is that (you believe that) the evil genius's prediction correlates with your brain state at time t_0. So, if you think that your intentions to act causally necessitate your action, then you must think that your intentions to act are also correlated with the evil genius's predictions.

lem is about the nature of choice—whether, when we form an intention to act in a certain way, our actions will be unpredictable. In the jargon, his question is whether our intentions to act may constitute *interventions*. He believes that this question can help settle the debate between EDT and CDT over whether or not to pay the evil genius. When viewed through the lens of graphical causal models, Stern says that this debate "appears to turn on how exactly agency should be causally represented",[8] and that "the nature of the disagreement between evidential decision theorists and causal decision theorists is *not* best treated as a disagreement about the fundamental normative principles that govern rational choice, but rather as a disagreement about the nature of choice."[9]

As an aside: while Stern's contribution is full of insight about the causal structure of choice and its relation to probability and decision, I think we should reject this claim about the debate between EDT and CDT boiling down to a disagreement about the nature of choice. What Stern convincingly demonstrates is this: if you think that you have some way of choosing what to do which is not predictable by the evil genius, then EDT will tell you to choose to keep your money in this way (in the jargon: it will tell you to *intervene* so as to keep your money). Whereas, if you think that, no matter how you decide what to do, your decision will be predictable by the evil genius (in the jargon: if your decision is not an intervention), then EDT will tell you to hand over the £1,000. This shows that we can achieve either verdict about whether to pay by combining EDT with an appropriate assumption about the nature of choice. But this doesn't show that the debate between EDT and CDT has anything to do with the nature of choice. It shows only that whether your act is predictable or not matters for evidential decision theorists.[10] For *causal* decision theorists, predictability is irrelevant. They will tell you to keep your money *whether or not* this decision of yours is predictable.[11] So I don't believe that the debate between evidential and causal decision theorists is a debate about whether your choices are predictable. While the evidentialist's verdict depends upon the answer to this question, the causalist's verdict does not.

8. p. 202–3.

9. p. 203, footnote 4. Here, Stern is following Meek and Glymour (1994).

10. This is slightly obscured by the fact that Stern doesn't accept EDT, but rather a hybrid of EDT and CDT which he calls 'generalized interventionist decision theory' (GIDT). Given any hypothesis about the world's causal structure, GIDT evaluates acts as EDT does, paying attention to the evidence your act provides you about factors outside of your control within that causal structure. However, unlike EDT, and like CDT, GIDT ignores evidence your act may provide you about the world's causal structure. In the concrete cases he considers, there is no uncertainty about the world's causal structure, so GIDT agrees with EDT.

11. For those familiar with the terminology: keeping your money *causally dominates* paying the £1,000, whether or not you are predictable, and whether or not your decision constitutes an intervention.

For Price and Liu, the primary question is what they term 'the Euthyphro question for causalists': "Is it the causal connection between C and E that makes it rational to do C to achieve E? Or does the rationality of the latter somehow constitute or ground the fact that C causes E?"[12] Price and Liu argue for the second answer: facts about causation are grounded in or constituted by facts about rational choice. Their interest in Newcomb's problem comes from the agency theory of causation, according to which the relation of causation is a kind of secondary quality which rational agents project onto the world.[13] Agency theorists say that C causes E when it would be rational to choose to bring it about that C, were E the only thing you desired. Newcomb's problem, then, is of interest because it may provide a case in which we project a *backwards* causal relation onto the world. If paying the evil genius is an *effective strategy* for getting a £1,000,000 cheque, then the agency theory would say that paying the evil genius *causes* them to have written you a £1,000,000 cheque in the past. This is why Price and Liu understand Newcomb's problem differently from some other contributors. If Newcomb's problem is to present a case of backwards causation, then we had better not *define* Newcomb's problem to be a case in which there are no backwards causal relations, as Joyce does.

For Joyce, the primary question in play in discussions of Newcomb's problem is which theory of rational choice is correct. His central goal is to defend his preferred version of CDT against some putative counterexamples to the theory.[14] To illustrate the kind of counterexamples Joyce considers: suppose that you must drink one of two vials of liquid which the evil genius has placed in front of you. One contains an odourless, flavourless lethal poison, while the other contains harmless water (though you know not which). If the evil genius predicted that you'd choose the left vial, then she put the lethal poison on the left. If she predicted that you'd choose the right vial, then she put the lethal poison on the right. Once again, she is 60% reliable. If you choose to drink the vial on the left, L, you'll be 60% sure that it contains the poison. And if you choose to drink the vial on the right, R, you'll be 60% sure that it contains the poison. But you are given a third alternative: you may, if you wish, pay a pittance to use a coin flip to determine which vial you drink. While the evil genius is 60% reliable at predicting how you'll choose, she's unable to predict the outcome of the coin flip. If she predicted that you'd follow the coin flip, F, then she simply flipped her own coin to decide where to put the poison. Following the coin, you'll think that the vial you drink from is as likely

12. p. 160, with minor notational changes.

13. See, *e.g.* Menzies and Price (1993).

14. Roughly, these are the decision problems from Egan (2007), Ahmed (2014), and Spencer and Wells (2019).

to contain the poison as not, 50%.[15]

Faced with this decision, it's natural to reason as follows: choosing to follow the coin will give me a 50% chance of living; whereas, if I choose either left or right on my own, I'm 10% more likely to choose the poison. A 10% chance of life is worth a pittance to me, so I should pay. This reasoning has some pull, but CDT says that it is impermissible to pay the pittance. Joyce defends CDT by objecting to the reasoning above and attempting to disarm the intuition. The defense is compelling. As a lapsed causal decision theorist who still longs for the hymns and incense of causalist high mass, I wanted to be convinced. Unfortunately, I have some reservations—let me briefly explain them here.

On Joyce's treatment, you should deliberate until you are equally likely to select the left and the right vial (and *certain* to not pay the pittance to follow the coin). At that point, you will be indifferent between the two options L and R, and uncertain which you will ultimately choose. This state of indifference is one from which it is difficult to rationally extricate yourself. If you attempt to choose the left vial then you'll give yourself evidence that the poison is on the left, and so the utility of L will drop below the utility of R. And if you attempt to choose the right vial, then you'll give yourself evidence that the poison is on the right, and the utility of R will drop below the utility of L. So, even though you're indifferent between L and R, pursuing either option would give you new information which would change your rational evaluation of the options. So it appears that CDT does not allow you to rationally choose either L or R. Joyce deals with this as follows: from the deliberative perspective in which you think you're just as likely to choose L as R, you should (irrevocably) *pick* one of the two. *Picking*, for Joyce, is a special way of choosing. He explains: "'Pick' is a term of art for a choice process which selects one from a set of equally good acts in a way that is *not* sensitive to differences in utility. (Think Buridan's ass!) Picking is inherently arational."[16] Thus: when deciding between L, R, and F, rational deliberation ends in a draw between L and R. At that point, rational deliberation hands over causal responsibility for decision-making to an arational picking process. Because rational deliberation has ended, you are no longer deciding on the basis of options' utilities, so when you find yourself inclining towards picking left, this may lead you to evaluate right as as the more choiceworthy option, but it won't lead you to reconsider picking left (as it would if you were choosing rationally).

Joyce notes that CDT gives different advice if, instead of choosing between L, R, and F, you are asked to choose between following the coin, F, and not

15. This decision is based on the case from Ahmed (2014). See also *The Frustrator*, from Spencer and Wells (2019).

16. p. 150.

following the coin—which is just a choice between L and R, $\{L, R\}$. Presented with this second decision, $\{F, \{L, R\}\}$, CDT will tell you to choose F. He suggests that, when we feel pulled to pay the pittance, we are confusing the decision $\{F, L, R\}$ with the decision $\{F, \{L, R\}\}$. I'm not sure about this—but rather than disputing the source of the intuition, I'd like to focus on CDT's disparate treatment of the two cases. I don't want to criticize CDT for treating a sequential decision like $\{F, \{L, R\}\}$ differently from an all-at-once decision like $\{F, L, R\}$. Instead, I'm interested in the *reason why* it treats the two cases differently. When making the choice between F and $\{L, R\}$, CDT evaluates the arational pick between L and R *in prospect*—in the decision $\{F, \{L, R\}\}$, picking between L and R is not something that your current self may do, but rather, something that your *future* self may do. When CDT evaluates acts *in prospect*, it evaluates them the same way as EDT does. That is, even though, when it evaluates a *current* option, CDT ignores correlations between the option and states beyond your control, when CDT evaluates a *future* option, it takes into account correlations between that option and states beyond your control. And your pick, unlike the coin flip, is positively correlated with poison. By choosing $\{L, R\}$, you may save yourself a pittance, but you also *cause* yourself to be more likely to end up with poison. On the other hand, by choosing F, you lose yourself a pittance, but you also *cause* yourself to be less likely to end up with poison. Choosing $\{L, R\}$ makes you the kind of person who's 60% likely to drink poison. Choosing F makes you the kind of person who's only 50% likely to drink poison. And making yourself a person like that is worth a pittance.

How do things change when you face the decision $\{F, L, R\}$? The causal consequences of F remain unchanged. The coin flip is still causally downstream of your choice, and it is still uncorrelated with poison. So choosing F still makes you the kind of person who's only 50% likely to drink poison. And the arational process of picking between L and R still has a 60% probability of culminating with you drinking poison. Your picks are still biased towards poison. Still, when making the choice $\{F, L, R\}$, even though you know that picking between L and R is 60% likely to lead you to death, *picking between L and R* is not one of your options. Your options are L, R, and F. And while one of L and R—the one you will actually choose—has a 60% chance of leading you to death, one of L and R—the one you won't actually choose—has only a 40% chance of leading you to death. Before picking, you won't know which is which, since you'll think that you're equally likely to pick either; so you'll think both L and R are 50% likely to lead you to death. As neither costs a pittance, both are preferable to F.

On Joyce's view, in the decision $\{F, L, R\}$, rational deliberation concludes in a draw between L and R. Before abdicating responsibility, rational deliberation's final act is to entrust the arational picking process with the responsibility of deciding between L and R. In the decision $\{F, \{L, R\}\}$, on the other hand, rational deliberation never steps down; it concludes by choosing to not

allow your future self to pick between L and R. What I have a hard time understanding is why rational deliberation should conclude differently in these two decisions. I can see that the resolution of rational deliberation in the second decision is a *choice*; whereas, in the first, rational deliberation concludes before a choice is made. But I can't see why this difference should make a difference. In both decisions, rational deliberation may or may not enable an arational picking process which is biased towards poison. In both cases, whenever rationality reigns, this picking would take place in the future. If rational deliberation should evaluate picking *in prospect* in the decision $\{F, \{L, R\}\}$, then it seems to me that it should similarly evaluate picking *in prospect* in the decision $\{F, L, R\}$. True, in the second decision, picking commences immediately after rational deliberation concludes, during what we may wish to call *the same* decision; whereas, in the first, picking commences only after another round of rational deliberation, during a *separate* and *later* decision. But in both cases, whenever rationality holds the reins, picking lies in the future. And in both cases, picking is correlated with death. So I have a hard time seeing the rational grounds for treating picking differently in the two cases.

To dramatize my confusion, suppose you've reached the deliberational perspective in which you think you're just as likely to pick L as R. You ask Joycean CDT: 'What do I do now?'. It says: 'Pick between L and R.' You reply: 'But I thought that *picking between L and R* wasn't one of my options. Moreover, even if *picking between L and R* were one of my options, I definitely shouldn't choose it, since it has a 60% probability of leading me to death.' CDT: 'Of course you shouldn't choose to have your future self pick between L and R. If that were an option, you should avoid it like the plague! But it's not one of your options. In contrast, both L and R are options, and both of them are permissible. So picking between L and R will lead you to a permissible option. So you may permissibly pick between L and R—though, just to reiterate, by no means should you choose to put yourself in a position to pick between L and R.' At the end of the day, I just have a hard time making sense of this advice. And that is why I still have reservations about the view, despite Joyce's compelling defense.

Unfortunately, there is so much more of interest in *Newcomb's Problem* than I am able to discuss in any depth in this brief review. Chrisoula Andreou relates Newcomb's problem to Kavka's toxin puzzle and Quinn's puzzle of the self-torturer in a discussion of whether rationality could require you to bring about an outcome which you disprefer to another outcome you could have had instead. Ahmed defends paying the evil genius the £1,000 by noting that those who do so tend to end up with more money than those who don't. Melissa Fusco has an ingenious discussion of *time bias* in causal decision theory. She notes that causalists will gladly keep their money, but will be distressed to learn that they kept their money. Fusco uses some delightful examples to operationalize these preferences before defending this form of time-bias. Preston

Greene defends paying the evil genius even when the envelope is see-through, and you can see that she has written you a cheque for only £1,000. Robert Stalnaker has a fascinating discussion of the relationship between causal decision theory and game theory. And there's still more. The collection is a must-read for anyone interested in any of these topics.

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge, UK: Cambridge University Press.

Eells, Ellery. 1982. *Rational Decision and Causality*. Cambridge, UK: Cambridge University Press.

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116, no. 1: 93–114.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kavka, Gregory S. 1983. "The Toxin Puzzle." *Analysis* 43, no. 1: 33–36.

Meek, Christopher and Glymour, Clark. 1994. "Conditioning and Intervening." *The British Journal for the Philosophy of Science* 45: 1001–1021.

Menzies, Peter and Price, Huw. 1993. "Causation as a Secondary Quality." *The British Journal for the Philosophy of Science* 44: 187–203.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In "Essays in Honor of Carl G. Hempel," , edited by Nicholas Rescher, 114–146. Dordrecht: D. Reidel.

Quinn, William. 1990. "The puzzle of the selftorturer." *Philosophical Studies* 59: 79–90.

Spencer, Jack and Wells, Ian. 2019. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* 99, no. 1: 27–48.