Arif Ahmed (ed.)
Classical Philosophical Arguments: Newcomb's Problem
Arif Ahmed (ed), Classical Philosophical Arguments: Newcomb's Problem, Cambridge University Press, 2018
Reviewed by Jack Spencer, MIT

Realistic Newcomb problems might arise in monetary policy, first-past-the-post elections, and prisoner's dilemmas, but the most familiar Newcomb problem is fantastical. There are two boxes: one opaque, one transparent. The agent has two options: she can take either only the opaque box or both boxes. The transparent box contains $1,000. The opaque box contains either $0 or $1,000,000, depending on a prediction made yesterday by a predictor who is known to be very reliable. If the predictor predicted that the agent would take both boxes, the opaque box contains $0. If the predictor predicted that the agent would take only the opaque box, the opaque box contains $1,000,000. Assuming that the set-up of the case is coherent and possible, we face *the initial normative question*: What should the agent do? Should the agent one-box or two-box?

Robert Nozick, who brought the problem to attention in 1969, noted the attending disagreement:

> [...] I have put the problem to a number of people, both friends and students in class. To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with people thinking that the opposing half is just being silly. (Nozick 1969: 117)

A recent, more rigorous survey finds a similar divide. David Bourget and David Chalmers (2014) included Newcomb's problem in the survey they ran trying to determine what professional philosophers think concerning thirty central philosophical issues, and, among the roughly 55% of respondents who would took a stand on the matter, about 40% were one-boxers and about 60% were two-boxers.

The staying power of Newcomb's problem is impressive and due partly to the disagreement it engenders. But the interest of Newcomb's problem goes well beyond the initial normative question, as this volume of ten previously unpublished essays attests, and I think the staying power of Newcomb's problem is due primarily to the interesting light it sheds on other philosophical issues. Newcomb's problem raises questions about the nature and normative significance of time and causation. It raises questions about fairness, freedom, agency, and the nature of opportunity. It raises questions about the methodological interaction between social science and rational choice theory, descriptive and normative decision theory. And it serves as a dialectical fulcrum, since the two leading approaches to rational choice theory—evidentialism and causalism—are standardly taken to recommend one-boxing and two-boxing, respectively.

This volume should appeal to a wide audience, not just philosophers, but social scientists, psychologists, and political theorists, too. It should appeal to readers unfamiliar with Newcomb's problem, who can get up to speed quickly by reading the introduction to the volume, written by Arif Ahmed, who also edited the volume. And it also should appeal to readers familiar with Newcomb's problem; for almost all of the essays in the collection move the dialectics they engage with forward.

To provide a sense of the volume's contents, let me mention a few of the questions the volume raises and a few of the ways in which the collected essays engage with those questions.

One question raised by the volume concerns the possibility of Newcomb problems. José Luis Bermúdez argues that it is impossible for an ideal agent to face a Newcomb problem. The rest of the contributors argue or assume that Newcomb problems are possible, and some of the authors—for example, Robert Grafstein—seem to believe that the world abounds with Newcomb problems and that a proper understanding of Newcomb problems is an essential tool in social science.

A second question raised by the volume concerns the granularity of rationality. Most of the contributing authors defend the orthodox view: that rationality applies in the first instance to choices. But some of the contributing authors, including Chrisoula Andreou and Preston Greene, explore the rival view: that rationality applies in the first instance to choice-making dispositions.

Andreou's essay explores the connection between Newcomb's problem, Kavka's (1983) toxin puzzle, Quinn's (1990) self-torturer puzzle, and Andreou's (2008) own Newxin puzzle, with an eye toward determining whether it is ever rationally permissible for an agent to choose an option that they know for certain to be the worst option available to them.

Greene explores a "success-first" approach, in which normative decision theorists attempt "to discover decision theories … and determine their efficacy, under certain idealized conditions, in bringing about what is of ultimate value" (p. 117). As Greene points out, if we apply his success-first approach at the level of choice-making dispositions, then we are lead to a decision theory that not only recommends one-boxing in Newcomb's problem but also recommends one-boxing in a variation of Newcomb's problem in which both boxes are transparent. One-boxing in a transparent version of Newcomb's problem is a particularly vivid case in which an agent chooses an option that they know for certain to be the worst option available to them, but one-boxing in the familiar version of Newcomb's problem is arguably another case in which an agent chooses an option that they know for certain to be the worst option available to them.

A third cluster of questions raised by the volume concern the nature and normative significance of time and causation. Melissa Fusco's essay argues that both evidentialism and causalism give rise to time bias. Reuben Stern's essay explores the various ways that an agent facing Newcomb's problem might be represented using causal graphs and argues that whether agent's should one-box or two-box depends on how agency in Newcomb's problem should be represented. Robert Stalnaker's essay explores the causal structure of game theory and the light that game theory might shed on unstable problems that are thought to make trouble for causalism. And Huw Price and Yang Liu's essay presents a dilemma for causalism, which turns on the nature of causation.

On the first horn of Price and Liu's dilemma is a subjectivist conception of causation. Causalists usually assume that an agent facing Newcomb's problem has no causal control over how much money is contained in the opaque box. But if a subjectivist conception of causation is true, then an agent facing Newcomb's problem can retro-cause there to be $1,000,000 in the opaque box by one-boxing, and evidentialism and causalism then *both* recommend one-boxing. On the other horn of Price and Liu's dilemma is an objectivist conception of causation, which ensures that the agent has no causal control over how much money is contained in the opaque box. But, according to Price and Liu, an objectivist conception of causalism makes it "mysterious why causality should be the arbiter of rational choice, in a way that [causalism] proposes" (p. 161). If an objectivist conception of causation is true, then two-boxing *causally dominates* one-boxing: two-boxing is better than one-boxing, given the truth of any hypothesis about how the world beyond the agent's causal control is. But, according to Price and Liu, if an objectivist conception of causality is true, it's mysterious why there should be any special connection between causal domination and rational choice.

One possible line of response to Price and Liu, drawing on Spencer and Wells (2019), takes the connection between rationality and causation to be indirect. Causalists might first connect rationality to actual value, taking rationality to consist in choosing so as to maximize one's subjective expectation of actual value. Causalists then could connect actual value to causation, arguing that the actual value of an option is (say) the value that would be realized if the agent were to choose the option. This indirect connection between rationality and causation makes the connection between causal domination and rationality unmysterious. If a causal conception of actual value is true, then from the fact that two-boxing causally dominates one-boxing an agent can infer that the actual value of two-boxing exceeds the actual value of one-boxing. And if the agent knows that the actual value of two-boxing exceeds the actual value of one-boxing, then it follows trivially that the agent's expectation of the actual value of two-boxing exceeds the agent's expectation of the actual value of one-boxing.

Whether a causal conception of actual value is defensible remains to be seen. But, interestingly, the claim that rationality consists in choosing so as to maximize one's subjective expectation of actual value can be shown to be inconsistent with evidentialism, even without assuming a causal conception of actual value. (See Lewis (1988; 1996).)

A fifth question raised by the volume concerns "why ain'cha rich?" arguments. Andreou and Greene both address "why ain'cha rich?" arguments indirectly, by exploring the connection between success and rationality. Arif Ahmed and James Joyce both address "why ain'cha rich?" arguments directly.

Ahmed's essay is concerned with the "why ain'cha rich?" argument for one-boxing and the opportunity-based criticism of it. One-boxers almost always get \$1,000,000 upon facing Newcomb's problem; two-boxers almost always get \$1,000; and many one-boxers infer from these two facts that an agent facing Newcomb's problem should one-box. Two-boxers, like Wells (2019), respond to this argument by controlling for opportunities. Following Ahmed, let's say that "an agent has a *C-opportunity* to get a prize *X* at a time *t* if either (a) the agent does get *X* because of something that she does at *t* or (b) if the agent had chosen to act in some other way at *t* the agent would have got, or would have had a non-negligible chance of getting, *X*" (p. 64). As two-boxers point out, if we sort agents facing Newcomb's problem into equivalence classes by their *C*-opportunities, then we can turn the tables and run a "why ain'cha rich?" argument for two-boxing; for two-boxers are always \$1,000 richer than one-boxers who had the same *C*-opportunities.

It's here that Ahmed's essay picks up the dialectic. Ahmed grants that we should control for opportunities when running a "why ain'cha rich?" argument, but he argues that the sort of opportunities that we should control for are not *C*-opportunities, but rather *E*-opportunities, where "a proposition represents an *E-opportunity* for a deliberating agent if her confidence in its truth is not independent of her current intention" (p. 65-6). If we sort agents facing Newcomb's problem into equivalence classes by their *E*-

opportunities (and assume that the predictor is reliable but not perfect), then all of the agents belong to the same equivalence class, and one-boxers are again richer on average.

I’m not convinced that equality of *E*-opportunities entails equality of opportunities. After all, we can improve someone’s *E*-opportunities just by deluding them. Even if it’s *impossible* for poor people in America to become rich, we can give them the *E*-opportunity of becoming rich just by convincing them that they will become rich if they work hard. Similarly, it seems clear to me that agents who choose between $1,000,000 and $1,001,000 have better opportunities than do agents who choose between $1,000 and $0, even if the agents have the same *E*-opportunities. But what it takes for two agents to have equal opportunities is a question that matters not just in rational choice theory, but also in social and political philosophy, so I hope to see the issues discussed in Ahmed’s essay discussed more going forward.

Joyce’s essay defends two main claims: (i) that rationality requires that decision-making agents reach a certain kind of psychological equilibrium, and (ii) that causalism never errs in the recommendations it gives to agents who have reached the relevant sort of psychological equilibrium. One challenge to causalism, which is equally a challenge to Joyce’s causalism, are unstable problems, like Egan’s (2007) *Psychopath Button*, Ahmed’s (2014) *Dicing with Death*, and the following modification of an example from Spencer and Wells (2019):

> *Frustrating Boxes*. There is an envelope and two opaque boxes, A and B. The agent has three options: she can take box A, box B, or the envelope. The envelope contains $40. The two boxes together contains $100. How the money is distributed between the two boxes depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would take A, then box B contains $100. If the predictor predicted that the agent would take box B, then box A contains $100. If the predictor predicted that the agent would take the envelope, then the predictor flipped a fair coin and placed $100 in box A if the coin landed

> heads and placed $100 in box B if the coin landed tails. The agent knows all of this.

According to causalism, if an agent facing *Frustrating Boxes* has reached psychological equilibrium, in Joyce's sense, and is thus 50% confident that box A contains $100 and 50% confident that box B contains $100, then the only rationally impermissible option is taking the envelope: taking box A and taking box B are both rationally permissible. But there is a strong intuition that this prediction is exactly backward: that the only rationally permissible option for such an agent is taking the envelope. And we can undergird the take-the-envelope intuition with a "why ain'cha rich?" argument. Envelope-takers always get $40 upon facing *Frustrating Boxes*. A-takers almost always get $0. B-takers almost always get $0. And, unlike in Newcomb's problem, in which agents have equal *E*-opportunities but unequal *C*-opportunities, envelope-takers, A-takers, and B-takers have equal *E*-opportunities and equal *C*-opportunities. So, if success, holding fixed *E*-opportunities and *C*-opportunities, is a guide to rationality, then, *contra* causalism, the only rationally permissible option in *Frustrating Boxes* is taking the envelope.

Joyce is well-aware of the anti-causalist intuitions that unstable problems elicit and the "why ain'cha rich?" arguments that can be used to undergird those intuitions, and his essay attempts to develop the resources needed to resist them. Trying to determine when a "why ain'cha rich?" argument is sound and whether any sound "why ain'cha rich?" argument can be given for the anti-causalist intuitions elicited by unstable problems will, I suspect, continue to be of topic of interest going forward.

Space prevents me from broaching the many other interesting and important questions that the volume raises. But I hope that this review conveys how and rich and stimulating the volume is.

References

Arif Ahmed. 2014. "Dicing with Death." *Analysis* 74: 587-94.

Chrisoula Andreou. 2008. "The Newxin Puzzle." *Philosophical Studies* 139: 415-22.

David Bourget and David J. Chalmers. 2014. "What Do Philosophers Believe?" *Philosophical Studies* 170: 465-500.

Andy Egan. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116: 94-114.

Gregory S. Kavka. 1983. "The Toxin Puzzle." *Analysis* 43: 33-36.

David Lewis. 1988 "Desire as Belief." *Mind*: 323-32.

------. 1996. "Desire as Belief II." *Mind* 105: 303-13.

Robert Nozick. 1969. "Newcomb's Problem and Two Principles of Choice." In N. Rescher (ed.), *Essays in Honor of Carl. G. Hempel*, 114-46. Reidel.

William Quinn. 1990. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59: 79-90.

Jack Spencer and Ian Wells. 2019. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* 99: 27-48.

Ian Wells. 2019. "Equal Opportunity and Newcomb's Problem." *Mind* 128: 429-57.