Introduction

Arif Ahmed

This introduction sets out what Newcomb's Problem is, why it matters, and some things people have said about it. The appendix sets out some formal details of decision theory insofar as these are relevant to Newcomb's Problem.

1 What It Is

1.1 Nozick's Original Version

Credit for Newcomb's Problem should arguably go to Michael Dummett.[1] But Robert Nozick's eponymous 1969 paper is what set off the enormous debate that followed. Nozick states that he learnt the problem from the physicist William H. Newcomb of the Livermore Laboratory. Nozick puts it as follows.

> **Standard Newcomb**
> You must choose between taking (and keeping the contents of) (i) an opaque box now facing you or (ii) that same opaque box *and* a transparent box next to it containing $1000. Yesterday, a being with an excellent track record of predicting human behaviour in this situation made a prediction about your choice. If it predicted that you would take only the opaque box ('one-boxing'), it placed $1M in the opaque box. If it predicted that you would take both ('two-boxing'), it put nothing in the opaque box.[2]

He goes on: '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 large numbers thinking that the opposing half is just being silly'.

[1] This is Dummett's problem of the dancing chief (Dummett 1964). One reason to hesitate over this attribution is that in Newcomb's Problem it is stipulated that acts are causally irrelevant to correlated states, whereas the central question of Dummett's paper is over whether this is even possible.
[2] Nozick 1969: 207.

That disagreement matters in part because *we* may face versions of the problem. I'll discuss these at section 2.1. But it also matters because there are arguments for either side, resting on principles that until 1969 had seemed harmonious as well as compelling. Put very simply, these are as follows:

Causal Principle: A rational agent does what she thinks will *cause* her to realize her aims.

Evidential Principle: A rational agent does what constitutes her best *evidence* that she will realize her aims.

The Causal Principle seems to recommend two-boxing. You can't now *make* any difference to the contents of the opaque box, which were settled yesterday. Two-boxing therefore guarantees an extra $1K. The Evidential Principle seems to recommend one-boxing. One-boxing is, and two-boxing is not, excellent evidence that you are about to get $1M. So the Causal Principle and the Evidential Principle cannot *both* be right. I'll discuss these principles more formally at section 2.2.

1.2 General Form of the Problem

The problem invokes three features that are common to all decision problems – acts, outcomes and states; and four that are specific to it – stochastic dependence, causal independence and two kinds of dominance. I'll describe these in turn.

If you are choosing what to do, then your choice is between *acts*. Their *outcomes* are the possible consequences that matter to you. What outcome an act obtains depends on the *state* of nature at the time. In *Standard Newcomb* you choose between the acts of one-boxing and two-boxing. The outcome is monetary and depends on what you were predicted to choose, the latter being the state. Table 1 summarizes all this.

The column headings represent the possible states S_1 and S_2, the row headings represent your options A_1 and A_2, and the interior of the table

Table 1: Standard Newcomb

	S_1: *Predicted* A_1	S_2: *Predicted* A_2
A_1: Take only the opaque box	$1M	0
A_2: Take both boxes	$1M + $1K	$1K

indicates the outcome, in terms of your payoffs, in each of the four act/state combinations. These are the general features of the problem.

Let me turn to its specific features. First, there is *stochastic dependence* between act and state. When anyone takes only the opaque box, the predictor has almost always predicted this; when anyone takes both boxes, the predictor has almost always predicted *this*. This has two consequences. (a) One-boxers almost always end up millionaires, and two-boxers almost never do. (b) You are very confident that *you* will end up a millionaire if and only if you now take only the opaque box.[3]

Second, states are *causally independent* of acts. Whether you take one box or two *makes* no difference to the prediction. This combination of causal independence with stochastic dependence illustrates the saying that correlation is not causation. There is, e.g., a correlation between weather forecasts and subsequent weather, but weather forecasts have no causal influence on subsequent weather events, nor could any weather event have any (retroactive) influence on prior forecasts of it. The correlation exists because forecasts of weather and actual weather are effects of a common cause, i.e., the atmospheric conditions, etc., that precede both. In Newcomb's Problem, choices might similarly be correlated with predictions because they share a common cause, for instance, some previous state of the agent's brain.

The third feature is that one option *dominates* the other. Given either state – the state of predicted one-boxing, or of predicted two-boxing – you are \$1K better off if you two-box. I'll call this feature *horizontal dominance*.

The final feature is that the worst outcome in one state (that of predicted one-boxing) is better than the best outcome in the other. If the true state is S_1, then whatever you do, you are guaranteed a better outcome than you could possibly get in S_2. Right or wrong, a prediction of one-boxing makes you much better off than a prediction of two-boxing could. I'll call this feature *vertical super-dominance*.

These four specific features of Newcomb's Problem are responsible for the tension that it creates. Stochastic dependence and vertical super-dominance seem to rationalize one-boxing. After all, almost everyone who one-boxes ends up a millionaire, and almost nobody who two-boxes does, including people who have reasoned just as you might be reasoning now. Why expect to

[3] Strictly speaking, neither point logically *follows* from the assumption that the predictor has a good track record; the reasoning involves inductive inference from the latter. In Newcomb's Problem as in everyday life, we waive skeptical concerns about induction. We take it for granted that the predictor's past track record *is* strong evidence (a) of his future track record and (b) that he predicted you right on this occasion.

buck this trend? Causal independence and horizontal dominance seem to rationalize two-boxing. After all, either the $1M is already in the opaque box, or the opportunity to put it there is past, and nothing that you do can make a difference to whether it is there; and either way you are better off two-boxing.

2 Why It Matters

Newcomb's Problem is an interesting intellectual exercise, but so are many other things that have attracted less expenditure of thought and time. I think there are two reasons for the intense interest that *this* problem continues to provoke. (i) Its abstract structure seems to apply to cases that really do, or easily could, arise in real life. (ii) It motivated a profound shift in the way we think about rational choice. I'll take these points in turn.

2.1 Realistic Newcomb Problems

Here are two Newcomb Problems that could easily be, or probably are, real.

(a) *Fisher smoking case.*[4] Suppose that what explains the correlation between smoking and lung disease is not (as everyone now thinks) that smoking causes lung disease, but rather that both have a common cause: an innate predisposition towards lung diseases that also, and separately, predisposes its bearers to smoke. Suppose you are wondering whether to smoke, but you don't know whether you have the predisposition. You know that you would like smoking, but you like good health very much more.

The decision involves all four factors that distinguish a Newcomb Problem. Smoking and lung disease are stochastically related but causally independent. Smoking dominates non-smoking: whether or not you have the predisposition, you are better off smoking than not. And because you care a lot more about lung disease than about smoking, the absence of the predisposition super-dominates its presence. In this version, not smoking corresponds to one-boxing and smoking corresponds to two-boxing.

(b) *Voting in large elections.*[5] In a large election it is almost certainly true of you, as of any individual voter, that your vote won't affect the outcome. On the other hand, you might think your voting is symptomatic of whether others like you, in particular supporters of your candidate, will

[4] Jeffrey 1983: 15. [5] Quattrone and Tversky 1986: 48–57.

vote. If you expect turn-out to be decisive, your voting for your preferred candidate may be evidence of your preferred outcome.

If so, the case is a good approximation to Newcomb's Problem: good enough, that is, to raise the same problems. Let S_1 and S_2 be possible outcomes of the election – either your candidate wins or she does not – and let A_1 and A_2 be the options of voting for your candidate and not voting at all. The problem satisfies causal independence, nearly enough: it is practically certain that your vote *makes* no difference to the outcome of the election.[6] It satisfies horizontal dominance: given that your candidate wins, or given that she doesn't, you are better off not incurring the small opportunity cost of voting. And it satisfies vertical super-dominance, if it matters greatly to you that your candidate wins.

It is less obvious that elections involve stochastic dependence, but there is evidence that they do. From a purely statistical perspective, this is not surprising: if we consider all choices whether to vote, amongst Republican supporters from every US Presidential election that took place in the twentieth century, we should expect a correlation between a choice's having been to vote (rather than abstain) and the Republican candidate's having won. More importantly, this correlation has a subjective counterpart. Many people *do* think of their own choices as symptomatic of the choices of people like them, including in the context of large elections.[7] Any such person therefore faces a real-life Newcomb Problem in any large election in which he (i) can vote and (ii) has a strong interest. In this version of the problem, voting corresponds to one-boxing and not voting corresponds to two-boxing. (For more discussion, see the chapter by Grafstein in this volume.)

Those are two examples of Newcomb's Problem. The literature notes many others.

(c) The choice between vice and virtue in the context of Calvinist predestination.[8]
(d) Macroeconomic policy choice in the context of rational expectations[9] (but see the chapter by Bermúdez in this volume).

[6] For instance: in the UK since 1832, there have been five elections for Parliamentary representatives, out of approximately 30,000 such, in which the margin of victory has been zero or in single figures. This gives a frequency of about 0.05% of cases in which the margin of victory was in single figures.
[7] For evidence that they (i) think this and (ii) do so reasonably, see Ahmed 2014a section 4.6.
[8] Resnik 1987: 111; Ahmed 2014a: 9ff.; for historical details, see Weber 1992; Tawney 1998.
[9] Frydman, O' Driscoll and Schotter 1982.

(e) The choice whether to engage in some mildly unpleasant activity that is symptomatic of cardiac health.[10]
(f) The choice whether to smoke, when present smoking indicates future smoking.[11]
(g) Bets about experiments involving non-causal quantum correlations.[12]
(h) Choices in the Libet experiment, where experimenters can predict the agent's decision before she becomes consciously aware of it.[13]
(i) Bets about the prior state of the world in the context of determinism.[14]
(j) Decisions made by autonomous vehicles in an environment containing many similar agents.[15]
(k) Prisoners' Dilemma (from game theory) also realizes Newcomb's Problem, *if* each prisoner is confident enough that both reason alike[16] (but see the chapter by Bermúdez in this volume).

So despite its typically science-fictional presentation, the basic structure of Newcomb's Problem is arguably realistic, and its realizations may be very widespread.

2.2 Causal and Evidential Decision Theory

Probably the most important philosophical insight to have arisen from Newcomb's Problem is the distinction between two systematic ways of thinking about practical rationality. Perhaps the best way to understand the difference between these is in terms of two possible responses to the ancient philosophical problem of *fatalism*.

That problem itself arises from an overextension of a natural principle of rationality, which we may call the principle of *dominance*, a very simple version of which we may write as follows:

Dominance: For any two acts A_1 and A_2, if for each state the outcome of A_2 is better for you in every state than the outcome of A_1, then it is rational to choose A_2 over A_1.

Dominance looks perfectly reasonable: if, for instance, investing in gold gets you a better return than investing in land if the Republicans control the Senate after the next election, and gold also gets a better return than land if they do not, then it is sensible to invest in gold rather than land.

[10] Quattrone and Tversky 1986: 41–8. [11] Monterosso and Ainslie 1999; Ahmed forthcoming.
[12] Cavalcanti 2010; Ahmed and Caulton 2014. [13] Slezak 2013. [14] Ahmed 2014a section 5.2.
[15] Meyer, Feldmaier and Shen 2016. [16] Lewis 1979.

But the fatalist argument shows that in the absence of restrictions, the principle of Dominance leads to absurd consequences. Here is Cicero's report of one such case:

> So their [the Stoics'] argument goes: 'If you are destined to recover from this illness, whether you were to call in a doctor or not, you would recover; furthermore, if you are destined not to recover from this illness, whether you were to call in a doctor or not, you would not recover—and either one or the other is destined to happen; therefore it doesn't matter if you call in a doctor.'[17]

Dominance seems to validate this argument, and if calling a doctor carries any cost then it appears to recommend *not* calling a doctor. This is apparent if we lay out the acts, the states, and notional values for the outcomes as follows:

Table 2: Fatalism

	S_1: You recover	S_2: You don't
A_1: Call the doctor	5	0
A_2: Don't	6	1

In either state you are better off having not called the doctor than having called the doctor. Dominance seems to recommend not calling the doctor in this situation, however ill you are. More generally it seems to recommend the fatalist strategy of never taking a costly means to *any* end, however desirable.

Intuitively, the flaw in this reasoning is that it overlooks any *connection* between act and state. More specifically, we might expect dominance to fail when one of the acts in some sense makes the state more likely than does the other.

But this diagnosis, whilst correct, leaves room for two interpretations of 'making more likely'. We might say (i) that Dominance fails only if (as it seems to the agent) the acts have a *causal* influence on the state; or we might say (ii) that Dominance fails if the acts are (again from the agent's perspective) *evidentially* relevant to the state. Both (i) and (ii) are enough to undercut the fatalist argument, since calling the doctor is *both* a cause of, *and* evidence for, your recovery. But generalizing these diagnoses into a principle of action

[17] *De Fato* 28–9.

gives rise to theories of rationality that differ elsewhere, and in particular over Newcomb's Problem.

The natural generalization of (i) gives rise to *Causal Decision Theory* (CDT). This theory of rationality has various formal realizations that are not precisely equivalent; but what they all have in common is the idea that the rational act is whichever available one is most likely to *cause* what you want to happen.[18] The natural generalization of (ii) gives rise to *Evidential Decision Theory* (EDT), according to which the rational act is whichever available one is the best evidence of what you want to happen.[19]

EDT and CDT agree that you should call the doctor in *Fatalism*. And they agree wherever one's options are between acts that are evidence for exactly those states that they causally promote. But they do not agree over cases where one's acts are evidence for states that they do not causally promote, and this is exactly the situation in Newcomb's Problem. One-boxing is evidence that you will get $1M because it is evidence of the state in which you were predicted to one-box; EDT therefore recommends one-boxing. Two-boxing brings it about that you are $1K richer than you would otherwise have been; CDT therefore recommends two-boxing. Newcomb's Problem measures the distance between the thought that rational choice must pay special attention to *causal* dependences of states on acts and the thought that it need only be sensitive to the extent to which acts are *evidence* of states. The dispute over Newcomb's Problem is therefore one aspect of the more general epistemological question concerning the place of causality itself in our conception of the universe. (For technical details of EDT and CDT, see the Appendix to this Introduction. For discussion of alternative versions of CDT, see the chapter by Stern. For the connections between CDT and game theory, see the chapter by Stalnaker.)

3 The Debate Since 1969

In light of its apparently wide application, one might have expected discussion of Newcomb's Problem to have flourished in all those branches of science that

[18] The theory originated with Stalnaker 1972. For other versions, see Lewis 1981; Sobel 1986; Joyce 1999. All of these theories agree (a) that causal beliefs play a central role in rational decision-making; (b) that you should two-box in Newcomb's Problem. Spohn's (2012) and Price's (2012) versions of CDT both accept (a), but they reject (b), for different reasons. The Appendix to this Introduction spells out the relatively simple and early version attributed to Gibbard and Harper (1978).

[19] Jeffrey 1965 remains the classic exposition of Evidential Decision Theory. (Jeffrey 1983, the second edition of that book, modifies the theory so that it recommends taking both boxes in Newcomb's Problem – see Jeffrey 1983: 15–25.)

deal with choice, including economics, psychology and political science as well as philosophy. As it happened, professional discussion of the problem in the 1970s and early 1980s was largely conducted amongst philosophers. But there has since the 1980s been increasing (though still hardly mainstream) interest in the problem within these other areas and more lately in robotics and computer science. The following is a necessarily partial summary of some highlights of the philosophical literature.

3.1 Are Newcomb Problems Possible?

The "tickle defense" purports to show that, appearances to the contrary, nobody ever faces real Newcomb Problems. Informally, the idea is as follows. Newcomb's Problem arises only when you think that your act is evidence of a causally independent state. But this can only happen if either the state either itself causes you to act in some way or is a side effect of some prior cause of your act. Either way, this prior cause of your act must be mediated by your *motivations* – your desires and beliefs. At the time of acting, you know what your motivations are. But if you *know* that, you won't regard the act they produce as *further* evidence of its distal cause, nor therefore of the state. All the evidential bearing that your act could have on the state is already available from your known motivations.[20]

Thus, consider the Fisher smoking example described at 2.1(a). Smoking is supposed to indicate a predisposition that causes it. And it's plausible that learning that someone *else* smokes is evidence for you that she has that predisposition. But this is not so clear when it comes to *your own* smoking. If you are predisposed to smoke, then presumably you already like the idea of smoking (you have a "tickle" or urge to smoke), and whether you do is something that you already know. But the predisposition only makes you smoke by making you like the idea, and since you already know about that, your actual choice reveals no more about the presence or absence of the predisposition. From the perspective of the agent herself, smoking is therefore not any sort of evidence of a state that it doesn't cause. The Fisher smoking case is therefore not a Newcomb Problem.

Several issues arise here. We might question the quasi-Cartesian assumption that you know your own motivational state. The contrary idea, that subconscious desires and beliefs can play the same role in motivation as familiar conscious ones, is familiar from Freud, and whatever you think of

[20] Eells 1982: ch. 6, 7.

that, you might also think that a degree of muddle about what you think or want is a human imperfection that we cannot simply assume away.[21]

A second assumption of the tickle defense is that your acts only correlate with non-effects of them that are either their own causes or share some common cause with them. But cases of quantum entanglement cast doubt on this assumption, and it is possible to construct quantum cases that create a Newcomb-like clash between EDT and CDT but are immune to the tickle defense.[22]

But there is a second reason to doubt the possibility of Newcomb Problems. It is crucial that the state is *causally* independent of what you now do. But what *is* this causal relation? After all, nobody ever observes a relation of causality between distinct events, in the way that one observes, say, the relation of harmony or discord between distinct musical tones. So what is it? One possible answer, suggested by Berkeley but developed more thoroughly by Menzies and Price, is that A causes B when there is a correlation between an agent's *directly bringing about* A and the occurrence of B.[23]

This makes Newcomb's Problem impossible. We are told that the contents of the opaque box are correlated with what you bring about, i.e., whether you choose to take one box or two. But on the present view, this means that your choice *causes* the opaque box to contain $1M, or to contain nothing. That contradicts the stipulation that the contents of the opaque box are causally *independent* of what you do.

In response, one might think that there is a strong independent reason to doubt this "agency theory of causation." It may be that the account cannot be generalized to cover all impersonal causal relations without draining it of content.[24] It may be that the idea of an agent is itself causal in some way that makes the theory objectionably circular.[25] And it is a disturbing consequence of the theory that it appears to sacrifice the asymmetry as well as the temporal directedness of causation: since it is doubtless true, whether or not Newcomb's Problems are possible, that human actions have causes with which they are correlated, the theory is committed to saying quite implausibly that human actions are the causes as well as the effects of their own causes. (For further discussion, see the chapter by Price and Liu.)

[21] Cf. Lewis 1981a: 311–2. [22] For details, see Ahmed and Caulton 2014.
[23] Berkeley 1980 [1710]: sect. 25ff.; Menzies and Price 1993.
[24] For this and other criticisms of the agency theory, see Woodward 2003: 123–7.
[25] For discussion of various such objections, see Ahmed 2007.