PHI 523: DECISION THEORY

Spring 2021

Instructor: Boris Kment
email: bkment@princeton.edu

Course description

We will look at selected topics in the recent (and sometimes not so recent) literature in decision theory. We will pay special attention to Newcomb's Problem, evidential and causal decision theory, and questions about decision instability. In addition, we will discuss sundry puzzles, paradoxes, and conundrums related to rational decision making. Participants are welcome to suggest topics for the meetings towards the end of the semester.

As a special treat, the seminar will feature Jack Spencer (MIT), Alan Hajek (ANU), Melissa Fusco (Columbia), and Arif Ahmed (Cambridge) as guest speakers.

The syllabus, including the list of readings, will be distributed during the first session.

Topics

1. Newcomb's Problem

Imagine a game show in which an opaque and a transparent box are placed in front of you. The transparent box can be seen to contain a thousand dollars. The host of the show tells you that the opaque box contain either a million dollars or nothing at all. You can choose between taking the contents of both boxes and taking only the content of the opaque box (whatever it may be). At first, this may seem like a no-brainer—why take less if you can have it all? But suppose you are told that an extremely reliable oracle predicted what you would do. If she predicted that you would take both boxes, then nothing was put in the opaque box. If she predicted that you would take only the opaque box, then a million dollars were placed in that box. It's extremely likely that if you take both boxes, then the oracle correctly predicted this and you'll get only a thousand dollars, whereas if you take only the opaque box, then she likely predicted *that* and you'll get a million dollars. This might may sway you towards taking only the opaque box. But then again, your decision has no influence on what the oracle predicted, and no matter what she predicted, you'll be better off taking both boxes (you'll get a tousand dollars more). Doesn't that mean you should take both?

This decision problem is known as "Newcomb's Problem." Even after half a century of discussion among decision theorist, there is no consensus about what you should do in the situation described. We will look at some of the main arguments in this debate.

2. Evidential and Causal Decision Theory

The disagreement over Newcomb's Problem reflects a fundamental difference between two views about the nature of rational choice. Those who think that you should only take the opaque box typically subscribe to *evidential decision theory*. To simplify somewhat, this is the view that a rational decision maker chooses the option that provides the best evidence for an outcome she values. By contrast, those who advocate taking both boxes typically accept *causal decision theory*. According to the standard formulation of this view, rational agents choose the option that they take to be most likely to causally promote an outcome they value. In most decision situations, the two views designate the same options as rational. That is not true of Newcomb's Problem, which was carefully designed to involve a scenario in which the two views make different predictions. That is what makes Newcomb's Problem so important.

We will read some classic contributions to the debate between evidential and causal decision theory, as well as some recent papers that promise to shed new light on the issue.

3. Decision Instability

Gibbard and Harper describe the following story:

> Consider the story of the man who met death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, "I am coming for you tomorrow." The terrified man that night bought a camel and rode to Aleppo. The next day, death knocked on the door of the room where he was hiding and said "I have come for you." "But I thought you would be looking for me in Damascus," said the man.
>
> "Not at all," said death "that is why I was surprised to see you yesterday. I knew that today I was to find you in Aleppo."
>
> Now suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo.[1]

What should this unfortunate man do after his first encounter with Death? The options are to stay in Damascus and to ride to Aleppo. If he decides to stay in Damascus, he will come to believe that he will be in Damascus the next day, which rationally requires him to change

[1] Gibbard, Allan, and William Harper 1978, "Counterfactuals and Two Kinds of Expected Utility," in A. Hooker, J. J. Leach and E.F. McClennen (eds.), *Foundations and Applications of Decision Theory*, Dordrecht: Reidel, p. 158.

his mind and ride to Aleppo. But if he decides to ride to Aleppo, he will come to believe that he will be in Aleppo the next day, which will rationally require him to change his mind and stay in Damascus. No matter what he decides, he will have to change his mind immediately afterwards. Moreover, he knows that that is the case even at the beginning of his reflections. Or so at least it seems. What should he do?

As we will see, the puzzle as it stands arises for causal decision theorist but not for evidential decision theorists. We will look at some attempts by causal decision theorists to solve the puzzle.

4. Gideon's Paradox

A person whom you trust completely puts a check in the amount of 100 utils in front of you. You can take it (C) or not take it (¬C). "Before you make your choice, there is one thing you should know," the person says. "If you make a decision that is rationally impermissible, you will get an extra 1,000,000 utils."

Puzzle. Assuming that the beliefs of a rational agent are classically consistent, you must assume that one of the following is true.

(1) Both C and ¬C are permissible.
(2) Both C and ¬C are impermissible. (It's a good question whether it's possible for all options to be impermissible, but let's assume that that's possible. Ultimately, nothing hangs on it.)
(3) C is uniquely permissible.
(4) ¬C is uniquely permissible.

Here's a quick and dirty argument for the claim that you can't accept any of these possibilities. If you accept (1), you must conclude that what you do makes no difference to whether you will get the million utils. In that case, C seems uniquely rational—after all, choosing ¬C would amount to declining the 1,000-util check for no reason. That contradicts (1). An analogous problem arises if you accept (2). If you accept (3), you have to conclude that ¬C would be rewarded with a million utils while C would not be. That seems to make ¬C uniquely rational, which contradicts (3). The same problem (with C and ¬C reversed) arises if you accept (4).

This raises two questions: (i) Which options (if any) you should take to be permissible? (ii) What should you do?

This puzzle is sometimes called "Gideon's Paradox."[2] We will consider possible solutions.

[2] Bar-Hillel, M. and A. Margalit (1985), "Gideon's Paradox: A Paradox of Rationality," *Synthese* 63, pp. 139–155.