COMPSCI 323 - Computational Microeconomics

Notes for recitation, March 30: Voting and Newcomb's problem

Caspar, Hyoung-Yoon, Jiali, Vince

1 Voting methods

The example used in this section is from Pacuit (2019).

Consider a scenario where you are leading an intergovernmental organization (IGO) and trying to setup the voting rules that favor your state; or when you promised to grant universal suffrage to a colony and handover to civilian leader, but decide to manipulate the election by choosing a voting rule that elects a political puppet of you.

# Voters	Ranking			
3	A	B	C	D
5	A	C	B	D
7	B	D	C	A
6	C	B	D	A

Assume the votes are given above, now you are going to choose voting rules such that:

1.(a) Candidate A wins

Plurality

1.(b) Candidate B wins

Borda, STV

1.(c) Candidate C wins

(C is a condorcet winner) Slater, Copeland, Maximin, Kemeny. To see that candidate C wins under Kemeny and Slater, draw a pairwise election graph, which should be acyclic. (No need to invert any edges to compute winners.)

1.(d) Candidate D wins

None of the methods we learned in class

2 Newcomb's problem

Newcomb's problem was first published by Nozick (1969). The following version (which I think is a bit easier to understand) is from Chapter 5.1 of *The Foundations of Causal Decision Theory* by James M. Joyce (1999):

> Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand dollar bill, and says: "You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited $1,000,000 into your bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept. The money is already either in the bank or not, and nothing you now do can change the fact. Do you want the extra $1,000?" You have seen the psychologist carry out this experiment on two hundred people, one hundred of whom took the cash and one hundred of whom did not, and he correctly forecast all but one choice. There is no magic in this. He does not, for instance, have a crystal ball that allows him to "foresee" what you choose. All his predictions were made solely on the basis of knowledge of facts about the history of the world up to Friday. He may know that you have a gene that predetermines your choice, or he may base his conclusions on a detailed study of your childhood, your responses to Rorschach tests, or whatever. The main point is that you now have no causal influence over what he did on Friday; his prediction is a fixed part of the fabric of the past. Do you want the money?

This decision can be represented by the following payoff table:

	$1m deposited	$1m not deposited
Take the $1,000	$1m+$1k	$1k
Not take the $1,000	$1k	$0

As with the Sleeping Beauty problem discussed in lecture, (philosophical) decision theorists disagree about what is the rational choice this problem. The two main arguments are:

- Your choice does not causally affect whether the money was deposited. Regardless of whether the money is deposited, you get an extra $1,000 if you take the $1,000. Hence, you should take the $1,000. (causal dominance principle, causal decision theory)

- If you take the $1,000, then the psychologist likely will have predicted that you would take the $1,000. Hence, if you take the $1,000, you will likely only end up with $1,000. If, on the other hand, you reject the $1,000, the psychologist likely will have predicted this and so you will likely end up with $1,000,000. Hence, you should reject the $1,000. (evidential decision theory)

Other immediate reactions to the problem are about free will. Obviously, free will has been discussed very widely by philosophers. Unfortunately, Caspar (who is writing this) doesn't know this literature very well, so we don't say too much about this topic here and assume a more AI/CS/Econ perspective.

One reason to be interested in this kind of problem is that it might relate to strategic interactions. For example, when two people play Rock-Paper-Scissors, they will also try to predict each other. As noted in the lecture, it also relates to the Sleeping Beauty problem.

In the following, we sketch some other arguments about what should be done in Newcomb's problem and provide a few pointers to the literature. Vince and Caspar both know this literature somewhat well and are doing research related to it (see below). Feel free to contact us with questions and ideas! You don't need to know about Newcomb's problem for the final exam, but hopefully you find it interesting!

2.(a) Learning

Let's say you faced Newcomb's problem every day (maybe with smaller payoffs, so that you continue to care about the outcomes). Then on days on which you reject the $1,000, you will usually receive the $1,000,000. On days on which you take the $1,000, you will usually not receive the $1,000,000. So if you just blindly take the action that has been succeeded by high rewards in the past (à la (model-free) reinforcement learning or operant conditioning), you will end up taking one box.

On the other hand, what happens if you perform a kind of randomized controlled trial, i.e., if you flip a coin every time you face the problem. Which option will empirically look better?

2.(b) What if someone told you whether you had the money?

Imagine that before you make the decision, your bank calls to inform you about your current account balance. Unless you're very rich, this will tell you whether the psychologist sent you the $1,000,000. It seems that regardless of what the bank tells you, you will at that point want to accept the $1,000. But then it seems like the bank doesn't even have to call you – you already know what you will do after the call. So why not take the money even without the call?

2.(c) Regret

In algorithms and machine learning research, we often aim to design algorithm that minimize *regret*, where regret is the difference in utility between the best option and the option selected by an algorithm in question. In most algorithms and ML applications, the reason why regret is usually positive even for very good algorithms is that the algorithm doesn't know with certainty the value of all the different options. In Newcomb's problem, we it seems that rejecting the $1,000 always incurs a regret of $1,000 – in some sense, after the outcome is revealed one walks away with $1,000 less than one could have walked away with. If you take the $1,000, on the other hand, you will incur a regret of $0. A principle similar to regret minimization from philosophical decision theory is known as ratificationism, see for example Weirich (2016, Section 3.6).

2.(d) "Why Ain'cha Rich?"

If people faced Newcomb's problem in the real world frequently, then those who reject the $1k would be much richer on average. So they might ask those accepting the $1k: "If you're so smart, why ain'cha rich?"

2.(e) Medical cases

This is probably the most influential argument about Newcomb's problem.

In the words of Judea Pearl – a computer scientist, famous for his work on causality – from the 2nd edition of his book *Causality* (2009, Section 4.1.1):

> The paradoxes that emerge from this fallacy [evidential decision theory; rejecting the $1,000] are obvious: patients should avoid going to the doctor "to reduce the probability that one is seriously ill" (Skyrms 1980, p. 130); workers should never hurry to work, to reduce the probability of having overslept; students should not prepare for exams, lest this would prove them behind in their studies; and so on. In short, all remedial actions should be banished lest they increase the probability that a remedy is indeed needed.

Many find this type of argument convincing. The main counterargument is known as the tickle defense. For an overview, see, e.g., Section 4.3 of Arif Ahmed's 2014 book *Evidence, Decision and Causality*.

2.(f) An adversarial offer

From Oesterheld and Conitzer (unpublished):

> Two boxes, B_1 and B_2, are on offer. A (risk-neutral) buyer may purchase one or none of the boxes but not both. Each of the two boxes costs \$1. Yesterday, the seller put $\$3$ in each box that she predicted the buyer not to acquire. Both the seller and the buyer believe the seller's prediction to be accurate with probability 0.75. No randomization device is available to the buyer (or at least no randomization device that is not predictable to the seller).

At least one box contains money. That means that the average box contains more than it is worth. If you use causal decision theory (the theory that takes the \$1,000), then you don't take into account that the box you buy will likely be empty and you will therefore want to buy one of the boxes, thus voluntarily losing money. Voluntarily losing money seems problematic.

(Everyone would agree that if you can randomize in this example, you should! But it may be that you have no coin at hand or that the psychologist/seller also predicts whether you randomize and fills none of the boxes if you do randomize.)