Causal And Evidential Decision Theory

Newcomb's problem

- There are two boxes on the table:
    - B1 contains $1000 and you know this
    - B2 contains either $1M or nothing.
- You are now invited to make a choice between the following pair of alternatives:
    - A1: Take both B1 and B2
    - A2: Take only B2
- Imagine a being who is very good at predicting other people's choices.
    - 99 % of all predictions made by him / her so far have been correct.
- You are told that he / she will put $1M in B2
    - if and only if she predicts that you take just B2,
    - and nothing otherwise.
    - He / She knows that you know this.
- Using Dominance Prcinciple
    - Choose A1.
    - You don't have to be worried about the predictor's prediction.
    - She has made her prediction and put either $1M or nothing in B2.
    - When you are to choose, she cannot chnage what she has put in B2.
- Using the principle of maximizing expected utility
    - Choose A2.
    - u(A1) > u(A2)
- The predictor does not has magical powers, or he / she can change the past.
- What we know is only that his / her previous predictions were almost always right.

Causal Decision Theory

- A rational decision maker should
    - keep all her beliefs about causal processes **fixed** in the decision-making process,

        - and always choose an alternative that is optimal **according to these beliefs**
- The causal structure of the world is **forward-looking**, and completetly insensitive to past events.
- Maximize expected causal consequences.
- Example :
        - If eating an apple will cause you to be happy
        - and eating an orange will cause you to be sad
        - then you would be ration to eat the apple.
        - But what if eating a good will cause you to be happy,
        - and eating a bad apple will cause you to be sad
        - but you aren't sure if the apple is good or bad.
- (Only) consider $X\Box \rightarrow Y$
        - Indicative mood $P \rightarrow Q$
            - False when P is T and Q is F.
        - Subjunctive mood $P\Box \rightarrow Q$
            - False when in the closet possible world in which P is T Q is F.

Evidential Decision Theory

- Pressing the button does not cause any psychiatric disase,
- but if you were to press the button you would indirectly learn something about yourself,
    - that you did not already know, namely that you are a psychopath.
- It is not p($X\Box \rightarrow Y$) that should guide one's decision but rather $p((X\Box \rightarrow Y)|X)$.