.

24.118 – Paradox and Infinity
Problem Set 2: Newcomb's Problem

How this problem set will be graded:

- Assessment for questions 1–5 will depend *both* on whether you get the right answers, and on whether your answers are properly justified. Assessment for questions 6-9 will be based entirely on the *reasons* you give in support of your answers, rather than the answers themselves.
- *No answer may consist of more than 150 words.* Longer answers will not be given credit.

1. Evidential Decision Theory

You have two options: P (go to tonight's party) and S (study for tomorrow's exam). Which one should you pick? Whichever one leads to the best outcome, of course!

But maybe you're not sure which outcome will result from your choice. Maybe the exam will be easy, and you'll pass regardless of whether you study. In that case option P will yield an outcome in which you have fun tonight (value = 15) and pass the exam tomorrow (value = 20). Option S, in contrast, will yield an outcome in which you have a boring evening tonight (value = −2) and pass the exam tomorrow (value = 20). So—on the assumption that the exam is easy—the outcome of P will have a net value of 35, and the outcome of S will have a net value of 18.

But what if the exam is hard? In that case, you'll only pass if you study. So option P will yield an outcome in which you have fun tonight (value = 15) and fail the exam tomorrow (value −40). Option S, in contrast, will yield an outcome in which you have a boring evening tonight (value = −2 and pass the exam tomorrow (value = 20). So—on the assumption that the exam is hard—the outcome of P will have a net value of −25, and the outcome of S will have a net value of 18.

Table 1: Payoff Summary

	Go to Party	**Study for Exam**
Easy Exam	35	18
Hard Exam	−25	18

What option should you pick in a case like this? According to *Evidential Decision Theory* you should pick whichever maximizes *expected value.* The expected value of each of your options is defined as follows:[1]

[1] In general, the expected value of an option A is $\sum_i(v(O_iA) \cdot p(O_i|A))$, where:

Expected value of P:

(Net value of partying the night before hard exam × probability of facing a hard exam given that you've partied) + (Net value of partying the night before easy exam × probability of facing an easy exam given that you've partied)

Expected value of S:

(Net value of studying the night before hard exam × probability of facing a hard exam given that you studied) + (Net value of studied the night before easy exam × probability of facing an easy exam given that you studied)

1. Assume that the probability of getting a hard exam is 0.2, and that the exam will be easy just in case it's not hard. According to Evidential Decision Theory, should you choose P or should you choose S. (3 points)

   (You may assume that your teacher has no idea whether you've been partying or not—and doesn't really care—so the probability of getting a hard exam given that you've partied is no different from the probability of getting a hard exam given that you've studied: they are both 0.2.)

2. As before, but this time you have an evil teacher. Teacher knows whether you've been partying, and is much more likely to make the exam hard if you have. Whereas the probability of a hard exam given that you've studied is 0.2, the probability of a hard exam given that you've partied is 0.7.

   According to Evidential Decision Theory, should you choose P or should you choose S. (3 points)

3. What choice of probabilities would yield the result that your two options have the same expected value? (3 points)

2. Newcomb's Problem

There are two boxes before you: Open and Closed. You can see that Open contains $10, but cannot see the contents of Closed. You are told, however, that either Closed is completely empty or it contains a million dollars.

You are given two choices: you can OneBox or TwoBox. To OneBox is to take the contents of Closed, and leave the contents of Open behind. To TwoBox is to take the

- The O_i are an exhaustive list of relevant ways for the world might be, any two of which are mutually exclusive. (In our example, the O_i consist of O_1 and O_2, where O_1 is Hard Exam, and O_2 is Easy Exam.)
- $v(O_iA)$ is the net value of a situation in which O_i obtains and you do A. (In our example, $v(O_1P)$ is the net value of partying the night before a hard exam.)
- $p(O_i|A)$ is the probability that O_i obtains, conditional on your choosing A. (In our example, $p(O_1|P)$ is the probability of facing a hard exam, given that you partied.)

contents of both boxes. The boxes have been filled ahead of time, and their contents will not be changed. So your decision will have no effect on the contents of Closed.

4. According to Evidential Decision Theory, which of OneBox and TwoBox should you choose? (5 points)

   (You may make any assumption you like about the probability that Closed contains the million dollars. You may also suppose that the probability that Closed contains the million dollars is the same as the probability that it contain the million dollars given that you OneBox, and the same as the probability that it contain the million dollars given that you TwoBox.)

5. As before, but this time we make an additional assumption about the way in which it was decided how much money to put in Closed. Yesterday evening, Predictor was enlisted to make a prediction about whether you would OneBox or TwoBox. If Predictor predicted that you would OneBox, the million dollars was placed in Closed. Otherwise, Closed was left empty. Predictor is known to be accurate 80% of the time. The boxes have now been sealed, and their contents will not be changed. So, as before, *your decision will have no effect on the contents of Closed.* If it now contains the million dollars, it will continue to do so regardless of whether you decide to OneBox or TwoBox.

   According to Evidential Decision Theory, which of OneBox and Two Box should you choose? (5 points)

6. In the predictor case, what do *you* think the right choice is, regardless of what Evidential Decision recommends? (5 points)

3. The Prisoner's Dilemma

Jones and you committed a crime, and are under arrest. The police doesn't have much evidence against you, so if both of you keep quiet you'll each be charged with a relatively minor offense, and spend only 10 days in prison (a payoff of -10). But it is common knowledge that there is an offer on the table. If either of you agrees to sign a statement accusing the other, the defector will be allowed to leave scot-free (a payoff of 0), and the accused will be charged with a felony offense, and sentenced to 10,000 days in prison (a payoff of $-10,000$). Unfortunately, there is a catch. Should Jones and you *both* choose to defect, you will *both* be charged with felony offenses, but given the lesser sentence of 9,000 days in prison because of your cooperation with the police (a payoff of $-9,000$).

Table 2: Payoff Summary

	You Defect	**You Keep Quiet**
Jones Defects	You = $-9,000$, Jones = $-9,000$	You = $-10,000$, Jones = 0
Jones Keeps Quiet	You = 0, Jones = $-10,000$	You = -10, Jones = -10

Jones and you have been placed in separate interrogation rooms, and are not allowed to communicate with each other until after the decisions have been made, and it's too late to make any changes. So neither of you can do anything to affect the decision of the other. Jones and you will never see each other again, regardless of what you decide to do. And your action will have no consequences beyond those that have already been mentioned. (You need not worry about Jones exacting revenge, or about feeling guilt, or about acquiring a bad reputation, or anything like that.)

7. On the assumption that Jones is fully rational, what is the rational thing for you to do? (5 points)

8. As before, but this time Jones is your clone. The two of you are genetically identical, and have grown up in very similar environments. So there is a very high chance that the two of you will go through similar trains of thought, and end up making the same decision.

   On the assumption that Jones is fully rational, what is the rational thing for you to do? (5 points)

9. *Extra Credit:* As in the previous question, but this time Jones and you have been involved in 10 separate crimes. On each of 10 consecutive days a different crime will be considered, and you will face a new decision-situation with the payoffs in the table above. It is common knowledge that before each decision-situation the outcome of all previous decisions will be common knowledge.

   On the assumption that it is common knowledge that Jones and you are fully rational, what is the rational thing for you to do? (5 points)