

# Phil 4800

## Notes #1: Rational Choice & Framing

### I. Why rational choice theory might predict behavior:

- Rationality helps people pursue their goals. People are rewarded for being rational.
- Competition in the marketplace gives rational agents greater influence.
- The axioms of rational choice are intuitive.

### II. Background: Expected Utility & Rational Choice

*A decision problem has:*

- Options (choices): Things that the agent is choosing among. There must be two or more options.
- States of the World: The way things are descriptively, in the world outside the agent. May affect what happens to the agent. The actual state may be unknown, so there may be many possible states.
- Outcomes: Things that might happen as a result of the agent's choice. The outcome will typically be a function of the agent's choice and the state of the world. Good or bad outcomes are "payoffs".

*The agent has:*

- Subjective probabilities: These represent his opinions about the state of the world. May be understood as what he regards as fair betting odds for each possible state.
- Utilities: How good any given outcome is, from the agent's point of view. (May be his degree of satisfaction if the outcome happens.) The *utility function* assigns a number to each outcome, representing the agent's utility in that outcome.

*A Decision Rule: Maximize Expected Utility*

Perform the action with the highest expected utility, which is defined as

$$\sum_{i=1}^n P(O_i | A) \times U(O_i),$$

where A is an action the agent might perform,  $n$  is the number of possible outcomes,  $O_i$  is possible outcome #  $i$ ,  $U(O_i)$  is the utility of that outcome for the agent, and  $P(O_i | A)$  is the agent's subjective probability that that outcome will happen if he performs act A. Alternate form:

$$\sum_{i=1}^n U(O_i) \times \left[ \sum_{j=1}^m P(S_j) \times P(O_i | A, S_j) \right]$$

### II. Principles of Rational Choice

*Cancellation:* You can ignore ("cancel") states of the world in which your choice would not affect the outcome.

(*Related principle: Independence of Irrelevant Alternatives:* If A is preferred to B given alternatives {A,B}, then B cannot be preferred to A from alternatives {A, B, C}.)

*Transitivity:* If A is preferred to B, and B is preferred to C, then A should be preferred to C.

*Dominance:* If option A produces a better outcome (higher utility) than B given at least one possible state of the world, and does not produce a worse outcome in any other possible state, then A should be preferred to B.

*Invariance:* Different representations of the same problem must have the same solution.

*Completeness:* You have preferences for all alternatives.

*Continuity:* If A is better than B, which is better than C, then there's some probability  $p$  such that B is exactly as good as C combined with a probability  $p$  of getting A.

- If you combine all these assumptions, you get the principle of Expected Utility Maximization.

### III. Violations of Invariance (Framing Effects)

*The Cancer Treatments:*

(Survival Frame) Which is better:

Surgery: 90% live through surgery, 68% through 1 year, and 34% through 5 years.

Radiation: 100% live through treatment, 77% through 1 year, and 22% through 5 years.

(Mortality Frame) Which is better:

Surgery: 10% die in surgery, 32% die within 1 year, and 66% within 5 years.

Radiation: 0% die during treatment, 23% die within 1 year, and 78% within 5 years.

Radiation seems better in the Mortality framing.

*Gain/Loss risk aversion:*

(Gain Frame) Which is better:

a) A sure gain of \$24.

b) A 25% chance of getting \$100 and a 75% chance of getting nothing.

(Loss Frame) Which is better:

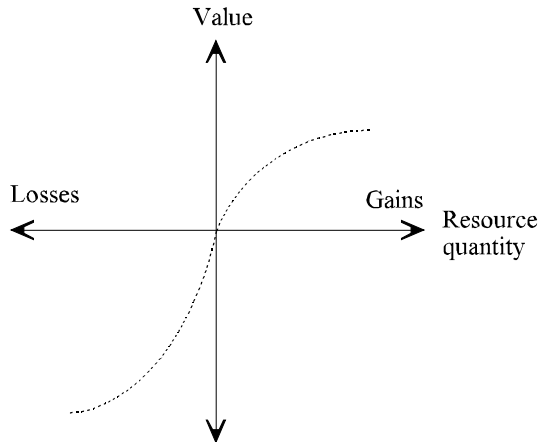
c) A sure loss of \$75.

d) A 75% chance of losing \$100 and a 25% chance of losing nothing.

Most people choose (a) + (d).

### IV. Prospect Theory: The Value Function

*People use an S-shaped value function:*



- People are risk-averse for 'gains', risk-seeking for 'losses'.
- *Loss aversion*: Losses are assigned greater (dis)value than foregone gains.
- Note status quo bias: the 'neutral' position (usually the status quo) is treated as the unique position with highest marginal utility.
- Framing can affect what is seen as the neutral/default position.

*Example*

600 people have a life-threatening disease. Which is better:

- 100% probability of saving 200 of them.
- A 1/3 probability of saving 600, and a 2/3 probability of saving none.

Which is better:

- 100% probability that 400 of them die.
- A 1/3 probability that none die, and a 2/3 probability that 600 die.

## V. Prospect Theory: Weighing Probabilities

*Features of the probability weighting function:*

- Impossible events are discarded.  $B(0)=0$ .
- Certain events are weighed simply according to the value function.  $B(1)=1$ .
- Low probabilities are over-weighted, moderate-high probabilities (below 1) are underweighted.
- Probability differences are treated as more significant for low probabilities than for higher probabilities. The difference between .1 and .2 probability is seen as more significant than the difference between .4 and .8 probability.

When we violate the axioms of decision theory:

- Cancellation and dominance are obeyed when the relations are obvious.
- When the relations are not obvious, people choose in accordance with prospect theory (above), sometimes violating cancellation or dominance.

# Phil 4800

## Notes #2: Pascal's Wager

### To discuss today:

How to evaluate bets.

Pascal's argument: why you should 'bet on' God's existence.

Objections to Pascal's wager.

### I. What Is Pascal's Wager?

- It is an argument intended to show that you *should* believe in God. It is referred to as Pascal's "wager" because Pascal compares believing in God to making a 'wager' (a bet) that God exists.
- Important distinction: 2 kinds of 'reasons for believing' something:
  - *Epistemic reasons*: Reasons that make it likely that the belief is true. An epistemic reason for believing a claim is evidence for the claim.
  - *Prudential reasons*: A prudential reason is a reason why something is in your own interests.
- Pascal proposes that you have a compelling prudential reason to believe in God.

### II. How to Evaluate a Bet

- When offered a bet, you should consider the following factors:
  - a) The probability that you will win.
  - b) How much you gain if you win.
  - c) The probability that you will lose.
  - d) How much you will lose if you lose.
- *Expected value* of the bet: This = [(a) × (b)] - [(c) × (d)]. A bet with positive expected value is good; one with negative expected value is bad.
- Note: the same basic idea applies not just to betting behavior, but more generally to making any decision when you're uncertain of the outcome.
- Example: You want to decide whether you should buy a lottery ticket. Suppose the prize is \$1 million, the probability of winning is 1/5,000,000, and the cost of a ticket is \$3. Then you can view this as a bet:

Probability of winning:	.0000002
Gain if you win:	\$999,997
Probability of losing:	.9999998
Amount of loss:	\$3
<hr/>	
<i>Expected value of bet:</i>	$(.0000002)(999997) - (.9999998)(3) = -2.8$

Hence, the 'bet' is unfavorable.

### III. The 'Bet' on Whether God Exists

- You can either 'bet' that He exists (by believing in Him), or 'bet' that he doesn't exist (by not believing in him).
- Note: Pascal only compares Christianity versus atheism. Doesn't consider other religions.
- According to Pascal:
  - a) If God exists and you don't believe in him, then you go to Hell forever (very bad).

- b) If God exists and you believe in him, you go to Heaven forever (very good).
  - c) If God doesn't exist, and you believe in him, nothing much happens.
  - d) If God doesn't exist, and you don't believe in him, nothing much happens.
  - e) It's about equally likely that God exists as that he doesn't, because there is no good evidence either way.
- Thus, we compare the two possible bets you can make:

Betting on God's existence:

Probability of winning: .5

Gain if you win: 4

Probability of losing: .5

Amount of loss: 0

---

*Expected value of bet:*  $(.5)(4) - (.5)(0) = +4$

Betting against God's existence:

Probability of winning: .5

Gain if you win: 0

Probability of losing: .5

Amount of loss: 4

---

*Expected value of bet:*  $(.5)(0) - (.5)(4) = -4$

Obviously, you should choose the former over the latter.

## IV. Objections

### ***Objection #1:***

The probability of God existing is not .5. It's much lower.

*Reply:* What happens if you substitute different numbers for ".5" in the following equations?

$$(.5)(4) - (.5)(0) = +4$$

$$(.5)(0) - (.5)(4) = -4$$

### ***Objection #2:***

The loss from believing in God is not 0. Believers waste time going to church, etc., and unbelievers have more fun. Plus, there is the potential disvalue of having a false belief.

*Reply:* What happens if you substitute different numbers for "0" in the following?

$$(.5)(4) - (.5)(0) = +4$$

$$(.5)(0) - (.5)(4) = -4$$

### ***Objection #3:***

The argument incorrectly assumes that there are only two possibilities, atheism and Christianity. What about other religions?

*Reply:* This must be granted. The argument favors theism over atheism. But it does not favor Christianity over, e.g., Islam or Judaism.

### ***Objection #4:***

The argument incorrectly assumes that we can choose what we believe.

*Reply:* You can take steps to try to attain belief.

### ***Objection #5:***

The argument assumes that God would punish people for not believing in him. But this is

incompatible with his being all-good.

*Reply:* The argument succeeds if there is any *chance* that this assumption might be true.

***Objection #6***

This might be true: “God hates believers. He will send all who believe in him to Hell, and send all atheists to heaven.” So you should be an atheist.

*Reply:* There is a better chance that God likes believers than that he hates them.

***Objection #7:***

“Anyone who doesn’t give me \$1000 by next week will suffer eternal torment.” This *might* be true, so you should give me \$1000.

*Reply:* There is more evidence for the claim that atheists will go to Hell than for the claim that non-donors of \$1000 to you will go to Hell. (Does this work?)

## Phil 4800

### Notes #3: Self-Torture & Transitivity

#### I. The Self-Torturer

- The torture device has 1001 settings:
  - S0: Off
  - S1: Unnoticeable level of electric current.
  - ...
  - S1000: Horrible torment.
- Each time you turn it up, you get \$10,000.
- The difference between any two adjacent settings is undetectable.

#### *A Counter-Example to Transitivity?*

1. For every  $n$  it is rational to prefer  $S(n+1)$  over  $S_n$ .
2. But it is not rational to prefer  $S_{1000}$  over  $S_0$ .
3. So transitivity of rational preference is false.

#### II. Weak Objections

1. The self-torturer's (ST's) preferences are changing.
  - No, they aren't.
2. We're neglecting behavioral manifestations of electric current.
  - The electric current changes are so small that there is no detectable behavioral difference either.
3. Ignoring measures of discomfort.
  - Discomfort levels are indeterminate.
4. The possibility of 'triangulation'.
  - The changes are too small to be detected even by triangulation.
5. The reversal of preferences: There must be a first stage  $S_n$  such that  $S_0$  is preferred to  $S_n$ . Then he should have stopped at  $S(n-1)$ .
  - Indeterminacy again.
6. The ST's preferences are paradoxical. He should give up some of his preferences.
  - The conclusion does not depend on any empirically false suppositions. (?)
7. The ST should weigh the risk that he might experience greater discomfort without noticing it.
  - This assumes there can be unidentifiable pains or changes in pain.
8. There is a risk that, if one takes the next step, one will subsequently (after further steps) be in unacceptable pain.
  - This doesn't apply if the ST assumes himself to be rational.
9. Why not pick a reasonable stopping point in advance, proceed to there, and then stop?
  - Once he arrives at that point, it will be rational to go one step farther. (The Principle of Strategic Readjustment)

#### III. Quinn's Solution

- Let the ST imagine a filtered series of steps (comprising a smaller number of steps than the original), such that
  - a) His preferences are transitive over the filtered series,
  - b) It has a member better than 0,
  - c) It is the most refined series satisfying these conditions.

- Then the ST should choose a reasonable stopping point in the filtered series.
- Then he should proceed to that step in the series he actually faces.
- Why shouldn't he then move to the next setting?  
    "A reasonable strategy that correctly anticipated all later facts ... still binds."

#### **IV. Compare/Contrast with Gauthier's idea**

- Gauthier thinks it is sometimes reasonable to fulfil an agreement, even when it is disadvantageous at the time.
- But this is due to the agent's limitations: inability to deceive others.
- Quinn's view does not depend on any inabilities on the part of the agent.

## Phil. 4800

### Notes #4: Rachels' "Counterexamples" to Transitivity

#### I. Preliminaries

- Transitivity: If A is better than B, and B is better than C, then A is better than C.
- An analogous claim applies to "worse than".
- Transitivity for 3-element series implies transitivity for n-element series.
- Stipulated sense of "pleasure" & "pain": A pain is an experience inferior to unconsciousness. A pleasure is an experience superior to unconsciousness.

#### II. Why the thesis is not too ridiculous

- Analogy: Rejecting absolute simultaneity (revises conception of time)
- Why we believe transitivity: an inductive argument.
  - But if we find a counter-example, then the inductive argument is refuted.

#### III. Counter-example

- Imagine a series of experiences:

- A 50 yrs of ecstasy
- B 5000 years of pleasure slightly less intense than A.
- C 500,000 yrs of pleasure slightly less intense than B.
- ...
- Z  $5 \cdot 10^{51}$  yrs of muzak and potatoes (barely pleasurable).

1. Each step is better than the previous step.
2. But Z is worse than A.
3. So transitivity is false.

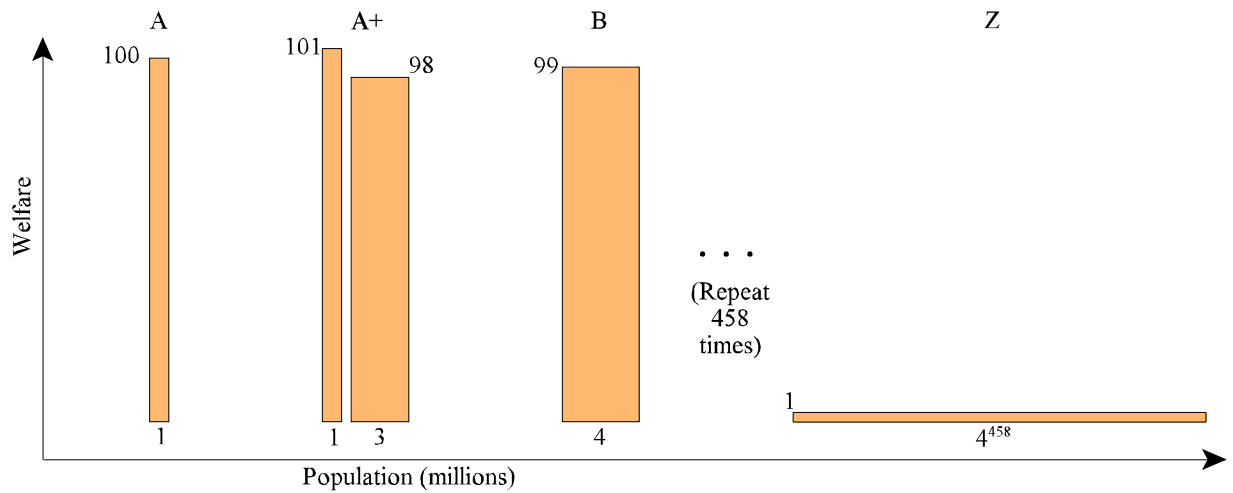
- Similar example can be constructed with pains.

#### IV. Objections

1. It's a sorites paradox.
  - No it isn't. The claim at each stage is equally plausible.
2. The scenario is too far from reality.
  - We have clear opinions about the comparisons & the concepts involved are simple & easily understood.
3. People have trouble grasping long periods. Reduce the pleasure/pain to 3 seconds.
  - Some possible pains are such that 3 seconds of them are worse than eons of mild discomfort.
4. There is some point at which making a mild pain milder but 100 times as long would be better.
  - This is not plausible, if one remembers the stipulated def. of "pain".

#### V. A Population Ethics Example

This is a simpler example than the one in Rachels' paper.



- A+ is at least as good as A.
- B is at least as good as A+.
- So B is at least as good as A. (Transitivity)  
... (repeat 458 times)
- Z is at least as good as A.

Most people think Z is *worse* than A. So transitivity must be rejected.

## VI. Implications

- Transitivity may still work in most cases.
- Value is not a quantity. Thus
  - No such thing as average utility, total utility.
  - Things don't have "more value" or "less value".
  - There is only a 2-place relation, "better than" (like "to the West of").

## Phil. 4800

### Notes #5: Arguments for Transitivity

#### I. Money Pump Argument

*Assume*

- Assume  $A > B > C > A$ .
- You have A.
- Would you trade A + a small amount of \$ for C?
- Would you trade C + a small amount of \$ for B?
- Would you trade B + a small amount of \$ for A?

*Argument's premises:*

1. A rational person would always trade something for something better.
2. A rational person would not be a "money pump".

*Rachels' Reply:*

Denies (1), because the rational person would want to avoid being money pumped.

*My reply:*

- Rachels' reply assumes intransitivity, so does not remove the force of the objection.

#### II. The Dominance Argument

*Premises:*

*(Mereological) Dominance* If  $X_1 > Y_1$  and  $X_2 > Y_2$ , then  $(X_1 + X_2) > (Y_1 + Y_2)$ .

- Assume no organic unities.

*Asymmetry* If  $X > Y$ , then  $\sim(Y > X)$ .

*Example 1:*

- Assume  $x_1 > x_2 > x_3 > x_4 > x_1$ .
- Then  $(x_1 + x_3) > (x_2 + x_4)$  (Dominance)
- Also,  $(x_1 + x_3) < (x_4 + x_2)$  (Dominance)
- But  $(x_2 + x_4) = (x_4 + x_2)$ ! So this state is both better and worse than  $(x_1 + x_3)$ .

*Example 2:*

- Assume  $x_1 > x_2 > x_3 > x_1$ .
- Then  $(x_1 + x_2 + x_3) > (x_2 + x_3 + x_1)$  (dominance)
- But  $(x_1 + x_2 + x_3) = (x_2 + x_3 + x_1)$ , so this state is better than itself.

*Example 3:*

- Assume  $x_1 > x_2 > x_3 > x_1$ .
- $(x_1 + x_2) > (x_2 + x_3)$ . (Dominance)
- $(x_2 + x_3) > (x_2 + x_1)$ . (Dominance)
- But  $(x_1 + x_2) = (x_2 + x_1)$ , so this state is both better and worse than  $(x_2 + x_3)$ .

# Phil. 4800

## Notes #6: Newcomb's Paradox

### I. Newcomb's scenario

*Two boxes:*

- B1: \$1000.
- B2: Either \$0 or \$1 million.

*Your choices:*

- Take B1 *and* B2.
- Take only box B2.

*The decision-predictor:*

- Predicts people's decisions, based on their personality traits, brain state, etc., with 90% accuracy.
- If he predicts that you will take *both* boxes, then he put \$1m in B2; otherwise, he put 0 in B2.
- You know all this, the predictor knows that you know it, etc.
- Sequence: First you hear the decision problem, then the predictor scans you and makes a prediction, then the predictor puts either \$1m or 0 in B2, then you make your choice.

### II. An argument for one box:

*Expected Utility Maximization:*

Let A1 = you take both boxes.

A2 = you take box B2.

Expected money payoff of A1:

$$(\$1,001,000) * P(\text{B2 contains } \$1\text{m} | A1) + (\$1000) * P(\text{B2 is empty} | A1) = (1,001,000)(.1) + (1000)(.9) = \$101,000.$$

Expected payoff of A2:

$$(\$1,000,000) * P(\text{B2 contains } \$1\text{m} | A2) + (\$0) * P(\text{B2 is empty} | A2) = (1,000,000)(.0) + (0)(.1) = \$900,000.$$

Presumably the expected utility of A2 is also greater than that of A1, so choose A2.

*Rationality proven by results:*

Suppose the experiment is repeated many times. Many one-boxers and many 2-boxers play.

The one-boxers wind up a lot richer than the 2-boxers.

This is predictable in advance.

So it's obviously better to be a 2-boxer than a 1-boxer.

### III. An argument for two boxing

*Dominance:*

- Either B2 contains \$1m, or B2 is empty.
- If B2 contains \$1m, then you should take both boxes (gaining \$1,001,000 instead of \$1,000,000).
- If B2 is empty, then you should take both boxes (gaining \$1000 instead of \$0).
- So A1 dominates A2. So choose A1.

*The Friend's Advice Argument*

Imagine a friend is sitting there, looking inside both boxes. He knows what is in B2, but he can't tell you. What would he advise you to do: take B2, or take both?

#### IV. Observations about the Two Principles

- 1) Dominance only applies when the states of the world are probabilistically independent of your action. Otherwise, apply expected utility.
- 2) The probabilities in EU are the prob. of the state conditional on the action.

##### *Example*

	S1	S2
A1	win \$10	win \$100
A2	win \$5	win \$90

- The person working the roulette wheel has been instructed to make S1 happen if I choose A1, and make S2 happen if I choose A2. It is highly probable that he will do so.
- A1 “dominates” A2. But you should choose A2.
- Explanation: S1, S2 are not probabilistically independent of A1, A2 (what you do affects the probabilities of the states).
- Expected utility calculation:  
Assume 90% probability that the roulette operator follows his orders.  
A1:  $(10)(.9) + (100)(.1) = 19$   
A2:  $(5)(.1) + (90)(.9) = 81.5$

##### *Qualification*

- 3) But Dominance *still* applies even if the states of the world are probabilistically dependent on your action, if the probabilities do not reflect any *influence* from the action to the states.

##### *Example*

There is a deadly, genetic disease, which you may or may not have. People with this disease are also more likely to choose academic careers. Should you avoid an academic career, so you will be less likely to have the disease? (No.)

#### V. Nozick's View

- Take both boxes.
- But why does N's problem seem different from the others (like the disease example above)?
  - The explanation of why the \$1m is (or is not) in B2 refers to your decision, but in an intensional context.
  - The idea that the predictor 'can't be' wrong encourages the thought that you have control over what he predicted.

#### VI. Postscript

- Nozick's later view in *The Nature of Rationality*:
  - Most people can be moved from one-box to 2-boxes, or vice versa, by changing the dollar amounts (keeping the example qualitatively the same).
  - There are 2 versions of Expected Utility: Causal EU, and Evidential EU.
  - We should use a weighted sum of CEU and EEU.

## Phil 4800

### Notes #7: The Banker Paradox

A variant on the Ross-Littlewood Paradox.

#### I. The banker game

- Infinite pile of \$\$, infinitely many turns. Bills are labeled “1”, “2”, “3”, etc.
- Round 1: Choose between
  - a) Taking bill \$1.
  - b) Taking bills 1-10, but giving back #1.
- ...
- Round n: choose between
  - a) Taking the next \$1 bill off the stack.
  - b) Taking the next ten bills, but giving back your lowest numbered bill.
- Each round is played in half the time of the last round. So the game completes in a finite time (twice the duration of the first round).

#### II. A paradox

- 1) At every stage, you should choose (b).
- 2) If you choose (b) at every stage, you end up with 0.
- 3) If you choose (a) at every stage, you end up with \$4.

#### III. A variant

- At each stage, instead of giving back your lowest # bill, you switch the labels of the lowest- and highest-numbered bills, then give back the new lowest # bill.
- In this case, you end up with \$infinity.
- This series is qualitatively identical at every stage as the original series. But it has the opposite outcome.

#### IV. What has gone wrong?

- Perhaps the hypothesized series is impossible. Why?
  - Maybe you can't have an actual infinity.
    - There are many counter-examples. Space, the past, the future, numbers, etc.
  - Perhaps it is impossible to complete an infinite series “by successive addition”.
    - Zeno's paradox refutes this. Zeno's series is completed every time an object moves.
  - Perhaps there cannot be an infinite series of dependencies.
    - Another Zeno series refutes this.
  - Finally, perhaps there cannot be an infinite *intensive magnitude*

## Phil 4800

### Notes #8: St Petersburg Paradox

#### I. The St Petersburg game

- A fair coin is flipped until the first time it comes up heads. Let the number of flips =  $n$ . ( $n \geq 1$ )
- Payoff:  $\$2^n$
- Q: How much should you be willing to pay for a chance to play this game?

#### II. Problem:

- Expected value of the game =  $\$infinity$ .
- The chance to play the game does not appear to be worth  $\$infinity$ .
- Also, notice that the actual payoff is 100% certain to be less than the expected value.

#### III. Solutions

- Appeal to diminishing marginal utility of money.
  - Response: Just increase the \$ payouts faster. As long as utility is unbounded, the paradox can be reproduced with some faster schedule of payouts.
- Bounded utility: there is only so much time during which to spend your money.
  - Response: Give payout in time + money.
- Note that the assumption of infinite resources is crucial.
  - If the house has only \$1 million, then the game is worth only \$11.
  - If the house has \$60 billion (Bill Gates), the game is worth \$19.
  - If the house has the entire world GDP (\$55 trillion), the game is worth \$24.
  - As the house's resources increase, the expected \$ value of the game increases without bound, but slowly.
  - Maybe infinite resources are impossible.

# Phil 4800

## Review of Unit 1

At the end of this unit, you should know ...

### These concepts:

'Expected' values  
Subjective probability  
Utility  
Instrumental rationality vs. epistemic rationality

### Principles:

Expected utility maximization  
Cancellation  
Transitivity for preference / value  
Dominance Principle (in decision theory)  
Invariance  
'Mereological' Dominance principle  
Asymmetry  
Diminishing marginal utility of money

### Empirical facts:

How people deviate from rationality

- Loss aversion
- Risk aversion vs. risk seeking, gains vs. losses
- Weighting of small probabilities
- How framing alters judgments, effects of gain vs. loss frames
- Which principles of rational choice people violate, when

### Hypothetical scenarios:

Self-torture  
Rachels' series of pleasures  
Newcomb's problem  
Nozick's case of the deadly genetic disease & academics  
Banker game  
& why you end up with \$0  
St Petersburg  
& expected value of the game  
& response to diminishing-utility-of-money solution

### These people & their main ideas:

Kahneman/Tversky  
Pascal  
Quinn  
Rachels  
Nozick  
Barrett/Arntzenius  
Cowen/High

### Arguments:

Pascal's argument 'for theism'

- Expected utility of 'betting on' God's existence
- EU of betting against
- What if we lower the Pr. of God's existence?
- What if we think being a theist is unpleasant?
- What if we think God wouldn't punish non-believers?

Quinn's argument against transitivity

- Why not rely on estimates of "level of discomfort"
- Why not weigh the 'risk' of having an unnoticed increase in discomfort

Rachels' argument against transitivity

- Why, e.g., there is no such thing as average utility

Money pump argument

- & Rachels' response

Dominance argument for transitivity

Newcomb: argument for one box & Argument for 2 boxes

### Theories:

Quinn's view of how the self-torturer should proceed  
Nozick's answer to Newcomb's problem & when Dominance Pr. applies