Puzzles and Paradoxes from Decision and Game Theory

Eric Pacuit
University of Maryland
pacuit.org

July 19, 2017

Plan

- ✓ Day 1: Rational Choice Theory, Decision Theory
- ✓ Day 2: Expected Utility Theory, Allais Paradox
- ▸ Day 3: Evidential and Causal Decision Theory, Introduction to Game Theory
- ▸ Day 4: Decisions over Time, Common Knowledge, Backward Induction and Epistemic Game Theory
- ▸ Day 5: Paradoxes of Interactive Epistemology, Framing in Games and Decisions

Allais Paradox

	Options	Red (1)	White (89)	Blue (10)
S_1	A	$1M$	$1M$	$1M$
	B	0	$1M$	$5M$

Allais Paradox

	Options	Red (1)	White (89)	Blue (10)
S_2	C	$1M$	0	$1M$
	D	0	0	$5M$

Allais Paradox

	Options	Red (1)	White (89)	Blue (10)
S_1	A	$1M$	$1M$	$1M$
	B	0	$1M$	$5M$
S_2	C	$1M$	0	$1M$
	D	0	0	$5M$

Allais Paradox

	Options	Red (1)	White (89)	Blue (10)
S_1	A	$1M$	$1M$	$1M$
	B	0	$1M$	$5M$
S_2	C	$1M$	0	$1M$
	D	0	0	$5M$

$$A \succeq B \text{ iff } C \succeq D$$

Comments on Expected Utility

Options	1/2	1/2
L_1	$1M$	$1M$
L_2	$3M$	$0M$

Comments on Expected Utility

Options	1/2	1/2
L_1	$1M$	$1M$
L_2	$3M$	$0M$

$EVM(L_1) = 1/2 \cdot 1 + 1/2 \cdot 1 = 1$
$EVM(L_1) = 1/2 \cdot 3 + 1/2 \cdot 0 = 1.5$

Comments on Expected Utility

Options	1/2	1/2
L_1	$1M$	$1M$
L_2	$3M$	$0M$

$EVM(L_1) = 1/2 \cdot 1 + 1/2 \cdot 1 = 1$
$EVM(L_1) = 1/2 \cdot 3 + 1/2 \cdot 0 = 1.5$

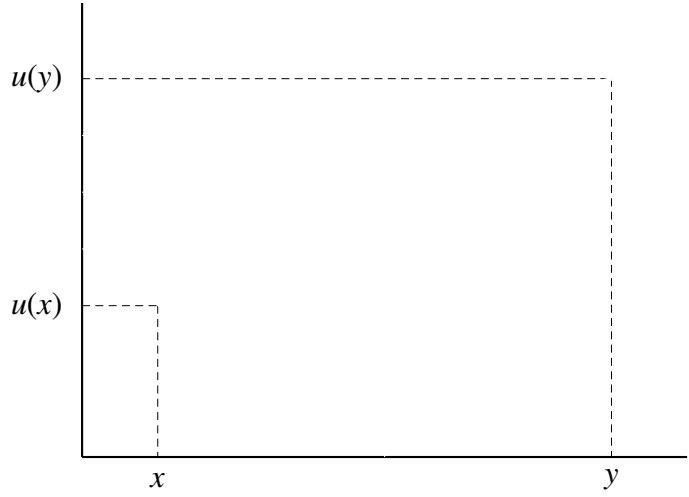
What numbers should we use in place of monetary value?

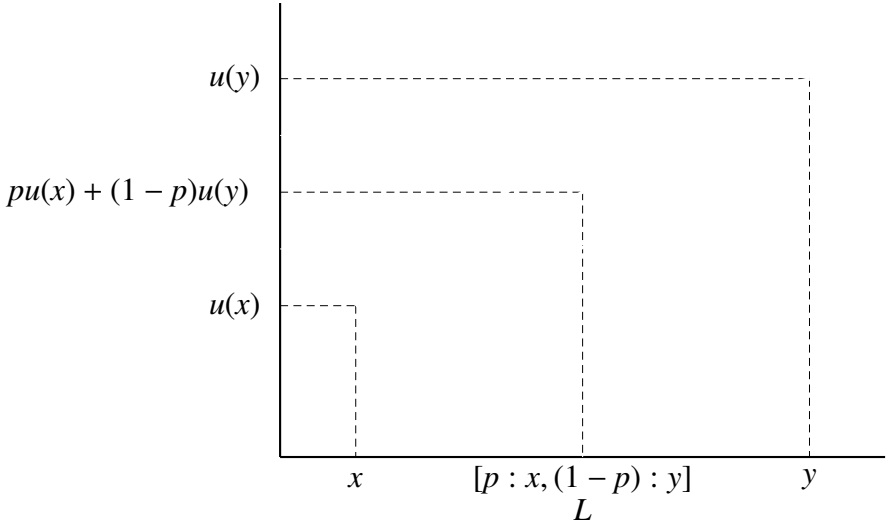
Comments on Expected Utility

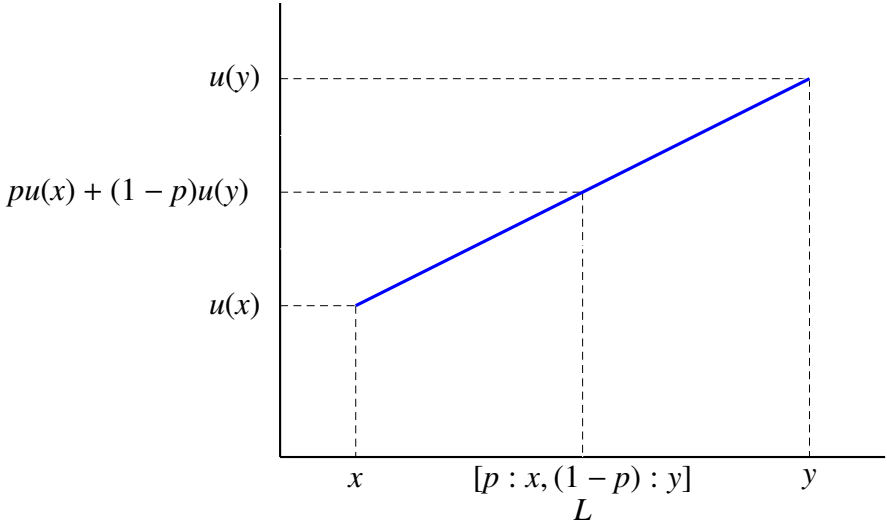
Options	1/2	1/2
L_1	$1M$	$1M$
L_2	$3M$	$0M$

$EVM(L_1) = 1/2 \cdot 1 + 1/2 \cdot 1 = 1$
$EVM(L_1) = 1/2 \cdot 3 + 1/2 \cdot 0 = 1.5$

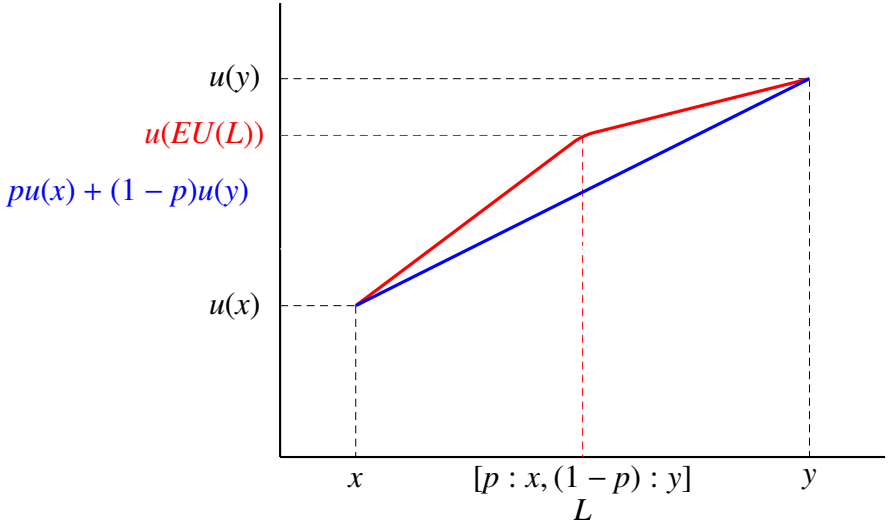
What numbers should we use in place of monetary value? (moral) value? personal utility?


$u(y)$
$u(x)$
x
y

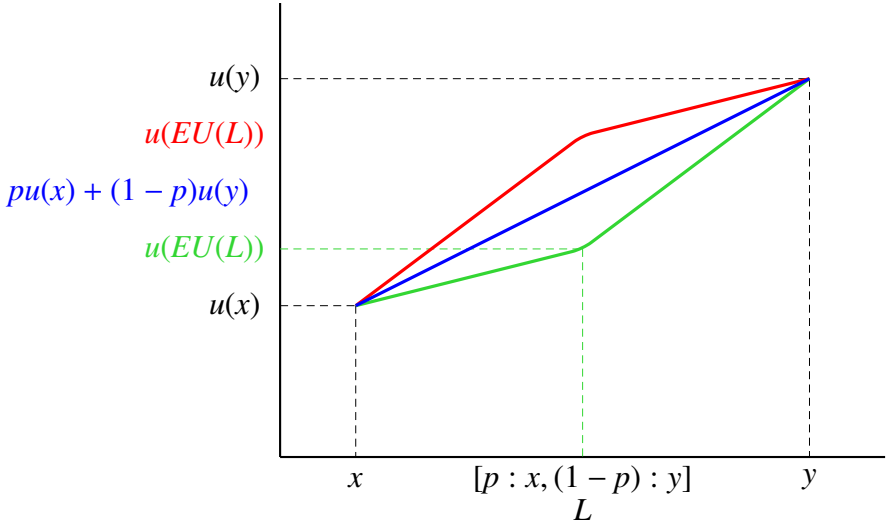

$u(y)$
$pu(x) + (1-p)u(y)$
$u(x)$
x
$[p : x, (1-p) : y]$
y
L



Risk neutral



Risk neutral Risk seeking



Risk neutral Risk seeking Risk averse

Allais Paradox

We should not conclude either

Allais Paradox

We should not conclude either

(a) The axioms of cardinal utility fail to adequately capture our understanding of rational choice, or

Allais Paradox

We should not conclude either

(a) The axioms of cardinal utility fail to adequately capture our understanding of rational choice, or

(b) those who choose A in S_1 and D is L_2 are irrational.

Allais Paradox

We should not conclude either

(a) The axioms of cardinal utility fail to adequately capture our understanding of rational choice, or

(b) those who choose A in S_1 and D is L_2 are irrational.

Rather, people's utility functions (*their rankings over outcomes*) are often far more complicated than the monetary bets would indicate....

Independence

Independence For all $L_1, L_2, L_3 \in \mathcal{L}$ and $a \in (0, 1]$,

$$L_1 \succ L_2 \text{ if, and only if, } [L_1 : a, L_3 : (1 - a)] \succ [L_2 : a, L_3 : (1 - a)].$$

Independence

Independence For all $L_1, L_2, L_3 \in \mathcal{L}$ and $a \in (0, 1]$,

$$L_1 \succ L_2 \text{ if, and only if, } [L_1 : a, L_3 : (1 - a)] \succ [L_2 : a, L_3 : (1 - a)].$$

$$L_1 \sim L_2 \text{ if, and only if, } [L_1 : a, L_3 : (1 - a)] \sim [L_2 : a, L_3 : (1 - a)].$$

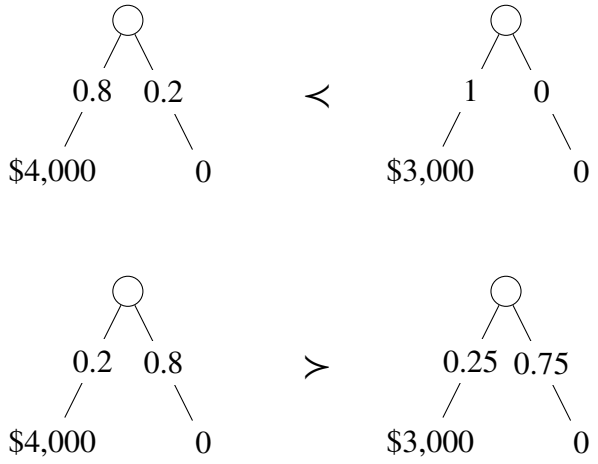
A: [\$4,000:0.80] B: [\$3,000:1]

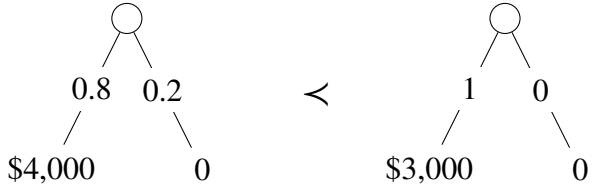
A: [\$4,000:0.80]

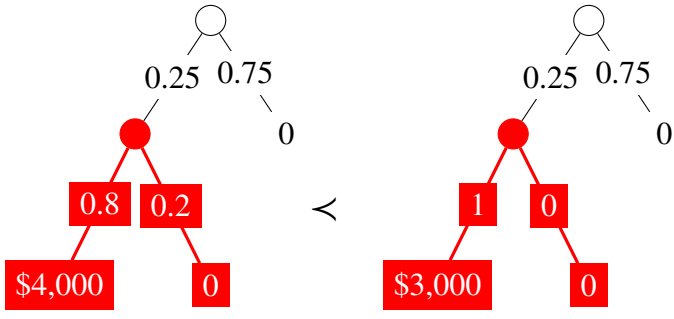
B: [\$3,000:1]

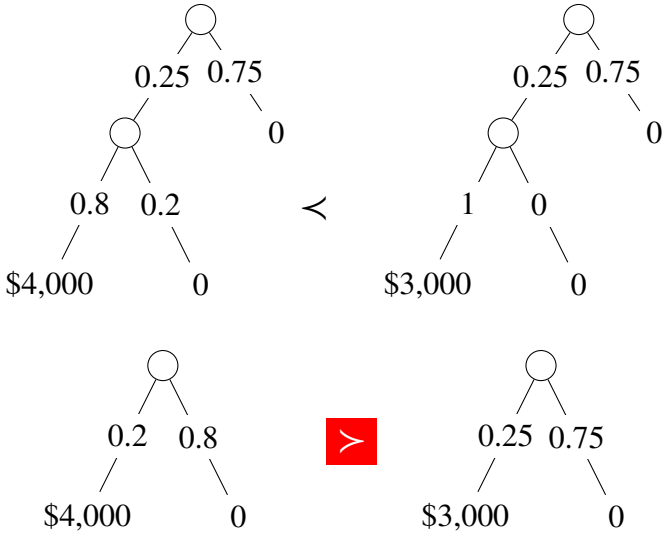
C: [\$4,000:0.20]

D: [\$3,000:0.25]

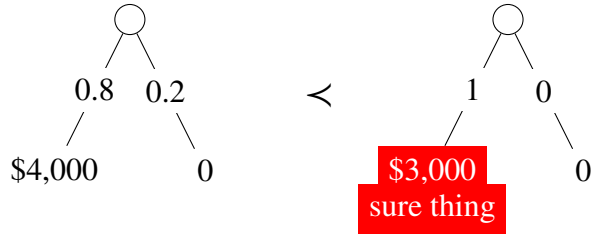

0.8
0.2
$4,000
0
≺
1
0
$3,000
0
0.2
0.8
$4,000
0
≻
0.25
0.75
$3,000
0

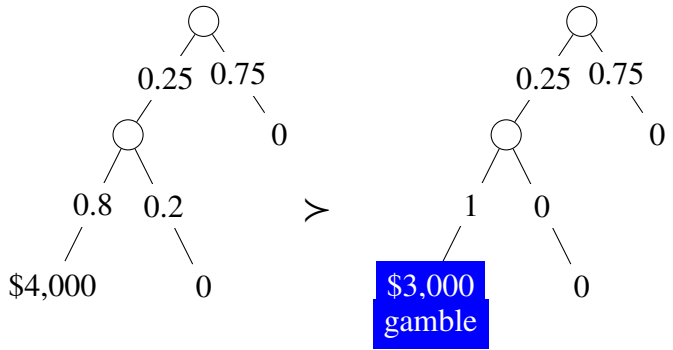

0.8
0.2
$4,000
0
≻
1
0
$3,000
0


0.25
0.75
0
0.8
0.2
$4,000
0
≻
0.25
0.75
0
1
0
$3,000
0



$\mathbf{0.25 * 0.8 = 0.2}$ $\mathbf{0.25 * 1 = 0.25}$


0.8
0.2
$4,000
0
≺
1
0
$3,000
sure thing
0


0.25
0.75
0
0.8
0.2
$4,000
0
≺
0.25
0.75
0
1
0
$3,000
gamble
0

Ellsberg Paradox

	30	60	
Lotteries	Blue	Yellow	Green
L_1	$1M$	0	0
L_2	0	$1M$	0

Ellsberg Paradox

	30	60	
Lotteries	Blue	Yellow	Green
L_3	$1M$	0	$1M$
L_4	0	$1M$	$1M$

Ellsberg Paradox

	30	60	
Lotteries	Blue	Yellow	Green
L_1	$1M$	0	0
L_2	0	$1M$	0
L_3	$1M$	0	$1M$
L_4	0	$1M$	$1M$

$$L_1 \succeq L_2 \text{ iff } L_3 \succeq L_4$$

Ambiguity Aversion

I. Gilboa and M. Marinacci. *Ambiguity and the Bayesian Paradigm*. Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress of the Econometric Society. D. Acemoglu, M. Arellano, and E. Dekel (Eds.). New York: Cambridge University Press, 2013.

Flipping a fair coin vs. flipping a coin of unknown bias: “The probability is 50-50”...

Flipping a fair coin vs. flipping a coin of unknown bias: “The probability is 50-50”...

- Imprecise probabilities
- Non-additive probabilities
- Qualitative probability

Evaluating Rational Choice Axioms

Any apparent violation of an axiom of the theory can always be interpreted in three different ways:

1. the subjects' preferences *genuinely* violate the axioms of the theory;
2. the subjects' preferences have changed during the course of the experiment;
3. the experimenter has overlooked a relevant feature of the context that affects the the subjects' preferences.

Aim of rational choice theory

- Explanation
- Prediction
- Recommendation

The Aim of Economics

> The main task of the social sciences is to explain social phenomena. It is not the only task, but it is the most important one, to which others are subordinated or on which they depend. (Elster, pg. 9)

J. Elster. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press, 2007.

Stability Individuals' preferences are stable over the period of the investigation.

Invariance Individuals' preferences are invariant to irrelevant changes in the context of making the decision.

The problem is that invariance is not a merely formal principle. If we left it to the agent to determine what counts as a “relevant” feature of the context, no choice would ever be irrational.

Principle of Individuation by Justifiers Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them.

A Dilemma

Either stick to the “formal axioms” of completeness, transitivity, Independence, etc. and refuse to assume the principles of stability and invariance.

A Dilemma

Either stick to the “formal axioms” of completeness, transitivity, Independence, etc. and refuse to assume the principles of stability and invariance. But then rational choice theory will be useless for all explanatory and predictive purposes because people could have fully rational preferences that constantly change or are immensely context-dependent.

A Dilemma

Either stick to the "formal axioms" of completeness, transitivity, Independence, etc. and refuse to assume the principles of stability and invariance. But then rational choice theory will be useless for all explanatory and predictive purposes because people could have fully rational preferences that constantly change or are immensely context-dependent. Alternatively, an economists can assume stability and invariance but only at the expense of making rational-choice theory a substantive theory, a theory laden not just with values but with *the economist's* values.

Dominance Reasoning and Act-State Dependence

	w_1	w_2
A	1	3
B	2	4

Dominance Reasoning and Act-State Dependence

	w_1	w_2
A	1	3
B	2	4

Dominance Reasoning and Act-State Dependence

Dominance reasoning is appropriate only when probability of outcome is *independent of choice*.

(A nasty nephew wants inheritance from his rich Aunt. The nephew wants the inheritance, but other things being equal, does not want to apologize. Does dominance give the nephew a reason to not apologize? *Whether or not the nephew is cut from the will may depend on whether or not he apologizes*.)

Newcomb's Paradox

A very powerful being, who has been invariably accurate in his predictions about your behavior in the past, has already acted in the following way:

1. If he has predicted that you will open just box B, he has in addition put \$1,000,000 in box B
2. If he has predicted you will open both boxes, he has put nothing in box B.

What should you do?

R. Nozick. *Newcomb's Problem and Two Principles of Choice*. 1969.

	$1 million in closed box	$0 in closed box
one-box	$1,000,000	$0
two-box	$1,001,000	$1,000

	\$1 million in closed box	\$0 in closed box
one-box	\$1,000,000	\$0
two-box	\$1,001,000	\$1,000

act-state dependence: $P(s) \neq P(s \mid A)$

Newcomb's Paradox

	B = 1M	B = 0
1 Box	1M	0
2 Boxes	1M + 1000	1000

Newcomb's Paradox

	B = 1M	B = 0
1 Box	1M	0
2 Boxes	1M + 1000	1000

	B = 1M	B = 0
1 Box	h	$1 - h$
2 Boxes	$1 - h$	h

Newcomb's Paradox

J. Collins. *Newcomb's Problem*. International Encyclopedia of Social and Behavorial Sciences, 1999.

Newcomb's Paradox

There is a conflict between maximizing your expected value (1-box choice) and dominance reasoning (2-box choice).

Newcomb's Paradox

There is a conflict between maximizing your expected value (1-box choice) and dominance reasoning (2-box choice).

What the Predictor did yesterday is *probabilistically dependent* on the choice today, but *causally independent* of today's choice.

$V(A) = \sum_w V(w) \cdot P_A(w)$
(the expected value of act A is a probability weighted average of the values of the ways w in which A might turn out to be true)

$V(A) = \sum_w V(w) \cdot P_A(w)$
(the expected value of act A is a probability weighted average of the values of the ways w in which A might turn out to be true)

Orthodox Bayesian Decision Theory: $P_A(w) := P(w \mid A)$ (Probability of w given A is chosen)

Causal Decision theory: $P_A(w) = P(A \mathrel{\square\!\!\rightarrow} w)$ (Probability of *if A were chosen then w would be true*)

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1)$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) = 1000000 \cdot 0.99 + 0 \cdot 0.01$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) = 1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) = 1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2)$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) = 1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2) = 1001000 \cdot 0.01 + 1000 \cdot 0.99$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) = 1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2) = 1001000 \cdot 0.01 + 1000 \cdot 0.99 = 11,000$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N)$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N) = 1000000 \cdot \mu + 0 \cdot (1 - \mu)$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N) = 1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N) = 1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \mathrel{\square\!\!\rightarrow} L) + V(K)P(B_2 \mathrel{\square\!\!\rightarrow} K)$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N) = 1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \mathrel{\square\!\!\rightarrow} L) + V(K)P(B_2 \mathrel{\square\!\!\rightarrow} K) = 1001000 \cdot \mu + 1000 \cdot (1 - \mu)$$

Let μ be the assigned to the conditional $B_1 \mathrel{\square\!\!\rightarrow} M$ (and $B_2 \mathrel{\square\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\square\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\square\!\!\rightarrow} N) = 1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \mathrel{\square\!\!\rightarrow} L) + V(K)P(B_2 \mathrel{\square\!\!\rightarrow} K) = 1001000 \cdot \mu + 1000 \cdot (1 - \mu) = 1000000\mu + 1000$$

Causal Decision Theory

A. Egan. *Some Counterexamples to Causal Decision Theory*. Philosophical Review, 116(1), pgs. 93 - 114, 2007.

Smoking Lesion: Susan is debating whether or not to smoke. She knows that smoking is strongly correlated with lung cancer, but only because there is a common cause a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and prefers smoking with cancer to not smoking with cancer. Should Susan smoke? Is seems clear that she should.

In The Smoking Lesion there is a strong correlation between smoking and getting cancer, despite the fact that smoking has no tendency to cause cancer, due to the fact that smoking and cancer have a common cause.

In The Smoking Lesion there is a strong correlation between smoking and getting cancer, despite the fact that smoking has no tendency to cause cancer, due to the fact that smoking and cancer have a common cause. Still, since Susan's $p(CANCER \mid SMOKE)$ is much higher than her $p(CANCER \mid NOT\ SMOKE)$, EDT assigns not smoking a higher value than smoking. And this seems wrong.

The Psychopath Button: Paul is debating whether to press the ‘kill all psychopaths’ button. It would, he thinks, be much better to live in a world with no psychopaths.

The Psychopath Button: Paul is debating whether to press the ‘kill all psychopaths’ button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button.

The Psychopath Button: Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

(Set aside your theoretical commitments and put yourself in Paul's situation. Would you press the button? Would you take yourself to be irrational for not doing so?)

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight.

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight. As the man knows, Death is a good predictor of his whereabouts. If he stays in Damascus, he thereby has evidence that Death will look for him in Damascus. However, if he goes to Aleppo he thereby has evidence that Death will look for him in Aleppo.

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight. As the man knows, Death is a good predictor of his whereabouts. If he stays in Damascus, he thereby has evidence that Death will look for him in Damascus. However, if he goes to Aleppo he thereby has evidence that Death will look for him in Aleppo. Wherever he decides to be at midnight, he has evidence that he would be better off at the other place. No decision is stable.

A. Gibbard and W. Harper. *Counterfactuals and Two Kinds of Expected Utility*. In Ifs: Conditionals, Belief, Decision, Chance, and Time, pp. 153 - 190, 1978.

- The crucial distinction is between an act and a decision to perform the act.

- Before performing an act, an agent may assess the act in light of a decision to perform it. Information the decision carries may affect the act's expected utility and its ranking with respect to other acts.

- Decision makers should make self-ratifying, or ratifiable, decisions.

Two Forms of Ratificationism

- As an *elimination rule*: ratificationism requires you to reject all unratifiable acts, and to then choose among the ratifiable alternatives.
- As an *equilibrium rule*: ratificationism requires you to choose an act that is ratifiable relative to the beliefs and desires you will have when your deliberations cease (“reflective equilibrium”).

H. Gaifman. *Self-reference and the acyclicity of rational choice*. Annals of Pure and Applied Loigc, 96, pgs. 117 - 140, 1999.

The Irrational Choice

Mr. Z offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20. Before making his decision, Adam is informed by Mr. Z that if he acts irrationally, Mr. Z will give him a bonus of \$100. (...to eliminate noise factors, assume that Adam believes that Mr. Z is serious, has the relevant knowledge, is a perfect reasoner and is completely trustworthy.)

"...the bonus condition in Z's statement has truth-conditions, and once Adam has chosen it can be evaluated...It is only from the perspective of *Adam* qua *deliberating rational agent* that the bonus condition must be excluded as meaningless."

“He could have chosen by whim, because of a feeling, a mood, or for no reason. The question how irrational choice is possible, what constitutes such a whim, impulse, temporary incoherence, weakness of will, or what have you, does not concern me here. I take it for granted that there will be cases which we shall characterize in this way (else ‘rational’ becomes a vacuous constraint).

“He could have chosen by whim, because of a feeling, a mood, or for no reason. The question how irrational choice is possible, what constitutes such a whim, impulse, temporary incoherence, weakness of will, or what have you, does not concern me here. I take it for granted that there will be cases which we shall characterize in this way (else ‘rational’ becomes a vacuous constraint). And if Adam chooses in this way he qualifies for the bonus, and will probably be surprised when he gets it. It is only from the perspective of *Adam* qua *deliberating rational agent* that the bonus condition must be excluded as meaningless.” (Gaifman, pg. 123)

The Rational Choice

Mr. Z offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20. Before making his decision, Adam is informed by Mr. Z that if he acts **rationally**, Mr. Z will give him a bonus of \$100. (...to eliminate noise factors, assume that Adam believes that Z. is serious, has the relevant knowledge, is a perfect reasoner and is completely trustworthy.)

Cassandra, a prophet of doom, used to warn people against disastrous actions, but her warnings went unheeded. She was doomed to be disbelieved by the same god who had given her the gift of foresight. And she knew it.

Imagine that, upon being asked for advice by some person, she warns the person against a certain action; but she also predicts that the person will not heed the warning. She makes thereby two predictions: that a certain action will have bad results, and that the person will take this action.

Eve's choice changes Cassandra's reliability as an expert. I cannot use an expert's advice as guide to my choosing, and at the same time use my choosing as evidence for the expert's reliability. That is, Cassandra's second prediction has no place in Eve's deliberations.

In its full generality the thesis means that, whatever information one uses in one's deliberations, one cannot use any non-trivial information about the likeliness of what one will choose.

- What about taking the advice of someone who calculates faster?

- What about taking the advice of someone who calculates faster? My choosing was already done: I chose the option determined by a certain mathematical condition. Then I chose to shortcut the implementation by “using” C as a computing device. The same would apply had the choosing been a consequence of logical deduction — in as much as the deduction comes under “computation”.

- What about taking the advice of someone who calculates faster? My choosing was already done: I chose the option determined by a certain mathematical condition. Then I chose to shortcut the implementation by "using" C as a computing device. The same would apply had the choosing been a consequence of logical deduction — in as much as the deduction comes under "computation".

- What about choosing by "gut-feeling"? The no-nonsense Eve decides in certain cases to go by her feelings: *that* is her choice. She implements it when she acts according to what she feels.

- What about basing a choice on past decisions?

- What about basing a choice on past decisions? Known or believed past performance can enter into the deliberation (“I know from experience that I tend to judge right in these situations”). To be sure, very often the line between deliberation and unthinking intuition is hopelessly blurred. Someone who estimates the probability of his own pending decision, can be construed as one who has chosen to delegate the deciding authority to a partner that acts by feel, inclination, the pull of certain forces, and the like. Yet, choosing-on-impulse can be shifted to the implementing stage and considered as “external” to the deliberation, in as much as the agent can reason *about* it.

Signaling through choice

The act of choosing may itself carry some rewards, say, a feeling of being in control. But this presupposes that there is also a less “active” (do-nothing) option, and the more “active” A, is preferred because it involves doing. But then one chooses A, for the “doing” that goes with it, not for the sake of choosing A.

You can choose in bizarre ways, in order to be original.

On can also choose A, in order to impress someone else.

It is understood that if one chooses A then one actually makes A true. But we should clearly distinguish between making A true and *choosing* to make A true.

(AC) The reason for choosing A can refer to each of the available options, but they cannot refer in an essential way to the *choosing* from these options, except through considerations of signaling.

(AC*) One should not use conditional probabilities (or likeliness estimates) of choices, which are obtained by conditionalizing on some event (or parameter) upon which the choice, in the agent's judgement, has no bearing.

The choice has no bearing means that it is considered irrelevant to the event in question. Such events can be subject to probabilistic estimates outside the choice context.

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

(c) is ruled out by (AC)

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

(c) is ruled out by (AC)

If Mr. Z is not assumed to be a perfect reasoner, Adam may rationally try to outsmart Z. (c) can be rephrased as a legitimate case of signaling: Adam signals (deceptively) to Mr. Z that choosing $S1$ he is behaving irrationally. Deceptive signaling is, of course, useless if you deal with a omniscient reasoner.

Newcomb's Paradox

(N1) Take one box for the reason: Given the evidence, if I take one box (make $B1$ true), I am likely to find there a very large sum; but if I take two I am likely to find the first empty, and the payoff from the second is comparatively paltry. The reasoning can be case in terms of expected utilities, where $P(E \mid B1)$ and $P(not - E \mid B2)$ are sufficiently high.

(N2) Take two boxes for the reason: Given the evidence, my doing does not influence in any way what the box already contains. Whatever is there, I do better by choosing $B2$.

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some "mental gymnastics" (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some "mental gymnastics" (1-box)
- "Tickle"-defense (2-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some "mental gymnastics" (1-box)
- "Tickle"-defense (2-box)
- Evidential Decision Theory: decisions to act provides evidence for the consequences (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some "mental gymnastics" (1-box)
- "Tickle"-defense (2-box)
- Evidential Decision Theory: decisions to act provides evidence for the consequences (1-box)
- Ratifiability: decision makers must assess the act in light of the decision to perform it and only choose acts that are self-ratifiable (1-box)