Rationality

Lecture 9

Eric Pacuit

Center for Logic and Philosophy of Science
Tilburg University
ai.stanford.edu/∼epacuit
e.j.pacuit@uvt.nl

April 11, 2011

Instrumental Rationality

Is decision theory a formalization of instrumental rationality?

Instrumental Rationality

Is decision theory a formalization of instrumental rationality?

If "goals" and "preferences" are the same thing, then decision theory is simply a formal version of instrumental rationality.

Instrumental Rationality

Is decision theory a formalization of instrumental rationality?

If “goals” and “preferences” are the same thing, then decision theory is simply a formal version of instrumental rationality.

Decision theory gives the agent some way to determine what is the “best” option, but in general this need not be the option that leads to the highest satisfaction of one’s goals.

Ultimatum Game

There is a good (say an amount of money) to be divided between two players.

Ultimatum Game

There is a good (say an amount of money) to be divided between two players. In order for either player to get the money, both players must agree to the division.

Ultimatum Game

There is a good (say an amount of money) to be divided between two players. In order for either player to get the money, both players must agree to the division. One player is selected by the experimenter to go first and is given all the money (call her the "Proposer"): the Proposer gives and ultimatum of the form "I get x percent and you get y percent — take it or leave it!".

Ultimatum Game

There is a good (say an amount of money) to be divided between two players. In order for either player to get the money, both players must agree to the division. One player is selected by the experimenter to go first and is given all the money (call her the "Proposer"): the Proposer gives and ultimatum of the form "I get x percent and you get y percent — take it or leave it!". No negotiation is allowed ($x + y$ must not exceed 100%).

Ultimatum Game

There is a good (say an amount of money) to be divided between two players. In order for either player to get the money, both players must agree to the division. One player is selected by the experimenter to go first and is given all the money (call her the "Proposer"): the Proposer gives and ultimatum of the form "I get x percent and you get y percent — take it or leave it!". No negotiation is allowed ($x + y$ must not exceed 100%). The second player is the Disposer: she either accepts or rejects the offer. If the Disposer rejects, then both players get 0 otherwise they get the proposed division.

Ultimatum Game

There is a good (say an amount of money) to be divided between two players. In order for either player to get the money, both players must agree to the division. One player is selected by the experimenter to go first and is given all the money (call her the "Proposer"): the Proposer gives and ultimatum of the form "I get x percent and you get y percent — take it or leave it!". No negotiation is allowed ($x + y$ must not exceed 100%). The second player is the Disposer: she either accepts or rejects the offer. If the Disposer rejects, then both players get 0 otherwise they get the proposed division.

Suppose the players meet only once. It would seem that the Proposer should propose 99% for herself and 1% for the Disposer. And if the Disposer is instrumentally rational, then she should accept the offer.

Ultimatum Game

But this is not what happens in experiments: if the Disposer is offered 1%, 10% or even 20%, the Disposer very often rejects. Furthermore, the proposer tends demand only around 60%.

Ultimatum Game

But this is not what happens in experiments: if the Disposer is offered 1%, 10% or even 20%, the Disposer very often rejects. Furthermore, the proposer tends demand only around 60%.

A typical explanation is that the players' utility functions are not simply about getting funds to best advance their goals, but about acting according to some norms of fair play.

Ultimatum Game

But this is not what happens in experiments: if the Disposer is offered 1%, 10% or even 20%, the Disposer very often rejects. Furthermore, the proposer tends demand only around 60%.

A typical explanation is that the players' utility functions are not simply about getting funds to best advance their goals, but about acting according to some norms of fair play. But acting according to norms of fair play does not seem to be a goal: it is a principle to which a person wishes to conform.

Choice Processes and Outcomes

A. Sen. *Maximization and the Act of Choice*. Econometrica, Vol. 65, No. 4, 1997, 745 - 779.

"The formulation of maximizing behavior in economics has often parallels the modeling of maximization in physics an related disciplines.

Choice Processes and Outcomes

A. Sen. *Maximization and the Act of Choice*. Econometrica, Vol. 65, No. 4, 1997, 745 - 779.

"The formulation of maximizing behavior in economics has often parallels the modeling of maximization in physics an related disciplines. But maximizing *behavior* differs from nonvolitional maximization because of the fundamental relevance of the choice act, which has to be placed in a central position in analyzing maximizing behavior.

Choice Processes and Outcomes

A. Sen. *Maximization and the Act of Choice*. Econometrica, Vol. 65, No. 4, 1997, 745 - 779.

"The formulation of maximizing behavior in economics has often parallels the modeling of maximization in physics an related disciplines. But maximizing *behavior* differs from nonvolitional maximization because of the fundamental relevance of the choice act, which has to be placed in a central position in analyzing maximizing behavior. A person's preferences over *comprehensive* outcomes (including the choice process) have to be distinguished form the conditional preferences over *culmination* outcomes *given* the act of choice." (pg. 745)

Choice Functions

Suppose X is a set of options. And consider $B \subseteq X$ as a choice problem. A **choice function** is any function where $C(B) \subseteq B$. B is sometimes called a menu and $C(B)$ the set of "rational" or "desired" choices.

Choice Functions

Suppose X is a set of options. And consider $B \subseteq X$ as a choice problem. A **choice function** is any function where $C(B) \subseteq B$. B is sometimes called a menu and $C(B)$ the set of "rational" or "desired" choices.

A relation R on X **rationalizes a choice function** C if for all B $C(B) = \{x \in B \mid \text{for all } y \in B \quad xRy\}$. (i.e., the agent is chooses according to some preference ordering).

Choice Functions

Suppose X is a set of options. And consider $B \subseteq X$ as a choice problem. A **choice function** is any function where $C(B) \subseteq B$. B is sometimes called a menu and $C(B)$ the set of "rational" or "desired" choices.

A relation R on X **rationalizes a choice function** C if for all B $C(B) = \{x \in B \mid \text{for all } y \in B \quad xRy\}$. (i.e., the agent is chooses according to some preference ordering).

Sen's α: If $x \in C(A)$ and $B \subset A$ and $x \in B$ then $x \in C(B)$

Choice Functions

Suppose X is a set of options. And consider $B \subseteq X$ as a choice problem. A **choice function** is any function where $C(B) \subseteq B$. B is sometimes called a menu and $C(B)$ the set of "rational" or "desired" choices.

A relation R on X **rationalizes a choice function** C if for all B $C(B) = \{x \in B \mid \text{for all } y \in B \quad xRy\}$. (i.e., the agent is chooses according to some preference ordering).

Sen's α: If $x \in C(A)$ and $B \subset A$ and $x \in B$ then $x \in C(B)$

Sen's β: If $x, y \in C(A)$, $A \subset B$ and $y \in C(B)$ then $x \in C(B)$.

You arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you that chair. However, if the matter is left to your own choice, you may refuse to rush to it.

You arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you that chair. However, if the matter is left to your own choice, you may refuse to rush to it. You select a "less preferred" chair.

You arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you that chair. However, if the matter is left to your own choice, you may refuse to rush to it. You select a "less preferred" chair. Are you still a maximizer?

You arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you that chair. However, if the matter is left to your own choice, you may refuse to rush to it. You select a "less preferred" chair. Are you still a maximizer? Quite possibly you are, since your preference ranking for choice behavior may well be defined over "comprehensive outcomes", including choice processes (in particular, who does the choosing) as well as the outcomes at culmination (the distribution of chairs).

You arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you that chair. However, if the matter is left to your own choice, you may refuse to rush to it. You select a "less preferred" chair. Are you still a maximizer? Quite possibly you are, since your preference ranking for choice behavior may well be defined over "comprehensive outcomes", including choice processes (in particular, who does the choosing) as well as the outcomes at culmination (the distribution of chairs).

To take another example, you may prefer mangoes to apples, but refuse to pick the last mango from a fruit basket, and yet be very pleased if someone else were to "force" that last mango on you. "
(Sen, pg. 747)

Let $X = \{x, y, z\}$ and consider $B_1 = X$ and $B_2 = \{x, y\}$. Define

$$C(B_1) = C(\{x, y, z\}) = \{x\}$$
$$C(B_2) = C(\{x, y\}) = \{y\}$$

This choice function cannot be rationalized.

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

[Adapted from Tversky and Kahneman (1981)]

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

1. You must choose between two prevention programs, resulting in:
   - A: 200 participants will be saved for sure.
   - B: 33 % chance of saving all of them, otherwise no one will be saved.

[Adapted from Tversky and Kahneman (1981)]

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

1. You must choose between two prevention programs, resulting in:
   - A: 200 participants will be saved for sure.
   - B: 33 % chance of saving all of them, otherwise no one will be saved.

   72 % of the participants choose A over B.

[Adapted from Tversky and Kahneman (1981)]

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

2. You must choose between two prevention programs, resulting in:
   - A': 400 will not be saved, for sure.
   - B': 33 % chance of saving all of them, otherwise no one will be saved.

[Adapted from Tversky and Kahneman (1981)]

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

2. You must choose between two prevention programs, resulting in:
   - A': 400 will not be saved, for sure.
   - B': 33 % chance of saving all of them, otherwise no one will be saved.

   78 % of the participants choose B' over A'.

[Adapted from Tversky and Kahneman (1981)]

Framing effects

Logicophilia, a virulent virus, threatens 600 students at Tilburg University

1. You must choose between two prevention programs, resulting in:
   - A: 200 participants will be saved for sure.
   - B: 33 % chance of saving all of them, otherwise no one will be saved.

   72 % of the participants choose A over B.
2. You must choose between two prevention programs, resulting in:
   - A': 400 will not be saved, for sure.
   - B': 33 % chance of saving all of them, otherwise no one will be saved.

   78 % of the participants choose B' over A'.

[Adapted from Tversky and Kahneman (1981)]

Framing effects

The Experiment:	
A: 0 + 200 for sure.	**B:** (33% 600) + (66% 0).
⇒ 72 % of the participants choose A over B.	
A': 600 - 400 for sure.	**B':** (33% 600) + (66% 0).
⇒ 78 % of the participants choose B' over A'.	

- Standard decision theory is extensional
  - Choosing A and $A \leftrightarrow B$ implies Choosing B.

Framing effects

The Experiment:	
A: 0 + 200 for sure.	**B:** (33% 600) + (66% 0).
⇒ 72 % of the participants choose A over B.	
A': 600 - 400 for sure.	**B':** (33% 600) + (66% 0).
⇒ 78 % of the participants choose B' over A'.	

- Standard decision theory is extensional
  - Choosing A and $A \leftrightarrow B$ implies Choosing B.

  Also true of many formalisms of beliefs:
  - "Believing" A and $\vdash A \leftrightarrow B$ implies "Believing" B.

Conclusions, I

- *Instrumental rationality* is a fundamental account of "rationality", but it is not necessarily the "whole of rationality"

Conclusions, I

- *Instrumental rationality* is a fundamental account of "rationality", but it is not necessarily the "whole of rationality"

- Utility is not a sort of "value", but simply a representation of one's ordering of options based on one's underlying values, ends and principles.

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)

- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)

- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments

- How do we "measure" an agent's beliefs?

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)
- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments
- How do we "measure" an agent's beliefs?
  Ramsey/De Finnetti: derive probabilities from utilities;

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)
- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments
- How do we "measure" an agent's beliefs?
  Ramsey/De Finnetti: derive probabilities from utilities;
  Aumann-Anscombe: include "objective probabilities";

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)
- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments
- How do we "measure" an agent's beliefs?
  Ramsey/De Finnetti: derive probabilities from utilities;
  Aumann-Anscombe: include "objective probabilities";
  Savage: derive utilities & probabilities from *preferences*.

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)
- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments
- How do we "measure" an agent's beliefs?
  Ramsey/De Finnetti: derive probabilities from utilities;
  Aumann-Anscombe: include "objective probabilities";
  Savage: derive utilities & probabilities from *preferences*.
- How should an ideally rational agent *change her beliefs*?

Rational Constraints on Beliefs

- Deductive cogency: an ideal rational agent's beliefs should be *consistent* and *deductively closed*. (Preface Paradox, Lottery Paradox, Problems with Closure)
- An ideally rational agent's graded beliefs should satisfy the laws of probability: Dutch Book Arguments
- How do we "measure" an agent's beliefs?
  Ramsey/De Finnetti: derive probabilities from utilities;
  Aumann-Anscombe: include "objective probabilities";
  Savage: derive utilities & probabilities from *preferences*.
- How should an ideally rational agent *change her beliefs*?

Savage's Representation Theorem

A set of states S, a set of consequences O, **acts** are functions from S to O.

Savage's Representation Theorem

A set of states S, a set of consequences O, **acts** are functions from S to O.

1. each act/state pair produces a unique consequence that settles every issue the agent cares about
2. she is convinced that her behavior will make no causal difference to which state obtains.

Savage's Representation Theorem

A set of states S, a set of consequences O, **acts** are functions from S to O.

1. each act/state pair produces a unique consequence that settles every issue the agent cares about
2. she is convinced that her behavior will make no causal difference to which state obtains.

The agent is assumed to have preference ordering $\succeq$ over the set of acts.

Savage's Representation Theorem

A set of states S, a set of consequences O, **acts** are functions from S to O.

1. each act/state pair produces a unique consequence that settles every issue the agent cares about
2. she is convinced that her behavior will make no causal difference to which state obtains.

The agent is assumed to have preference ordering $\succeq$ over the set of acts.

Expected Utility:

$$Exp_{P,u}(\alpha) = \sum_{w \in W} P(w) \times u(\alpha, w)$$

Small Worlds

States: {the sixth egg is good, the sixth egg is rotten}
Consequences { 6-egg omelet, no omelet and five good eggs destroyed, 6-egg omelet and a saucer to wash....}
Acts: { break egg into bowl, break egg into saucer, throw egg away}

Small Worlds

States: $\{$the sixth egg is good, the sixth egg is rotten$\}$
Consequences $\{$ 6-egg omelet, no omelet and five good eggs destroyed, 6-egg omelet and a saucer to wash....$\}$
Acts: $\{$ break egg into bowl, break egg into saucer, throw egg away$\}$

	Good Egg	Rotten Egg
Break into bowl	6-egg omelet	No Omelet and five good eggs destroyed
Break into saucer	6-egg omelet and a saucer to wash	5-egg omelet and a saucer to wash
Throw away	5-egg omelet and one good egg destroyed	5-egg omelet

Representation

EU-coherence: There must be at least on probability P defined on states and one utility function for consequences that **represent** the agent's preferences in the sense that, for any acts α and β, she strictly (weakly) prefers α to β only if $Exp_{P,u}(\alpha)$ is greater (as great as) $Exp_{P,u}(\beta)$.

Axioms of Preference

For all acts α, β, γ and events X, Y

Axioms of Preference

For all acts α, β, γ and events X, Y

1. **Trichotomy** One of the following is true: $\alpha \succ \beta$ or $\beta \succ \alpha$ or $\alpha \approx \beta$,

Axioms of Preference

For all acts α, β, γ and events X, Y

1. **Trichotomy** One of the following is true: $\alpha \succ \beta$ or $\beta \succ \alpha$ or $\alpha \approx \beta$,
2. **Transitivity** $\succeq$ is transitive

Axioms of Preference

For all acts α, β, γ and events X, Y

1. **Trichotomy** One of the following is true: $\alpha \succ \beta$ or $\beta \succ \alpha$ or $\alpha \approx \beta$,
2. **Transitivity** $\succeq$ is transitive
3. **Sure-thing Principle** If α and β produce the same consequences in every state consistent with X, then the agent's preference between the two acts depends only on their consequences when X obtains.

Axioms of Preference

For all acts α, β, γ and events X, Y

1. **Trichotomy** One of the following is true: $\alpha \succ \beta$ or $\beta \succ \alpha$ or $\alpha \approx \beta$,
2. **Transitivity** $\succeq$ is transitive
3. **Sure-thing Principle** If α and β produce the same consequences in every state consistent with X, then the agent's preference between the two acts depends only on their consequences when X obtains.
4. **Wagers** For consequences O_1 and O_2 and any event X, there is an act [O_1 if X, O_2 else] that produces O_1 in any state that entails X and O_2 in any state that entails $\neg X$

Axioms of Preference

For all acts α, β, γ and events X, Y

1. **Trichotomy** One of the following is true: $\alpha \succ \beta$ or $\beta \succ \alpha$ or $\alpha \approx \beta$,
2. **Transitivity** $\succeq$ is transitive
3. **Sure-thing Principle** If α and β produce the same consequences in every state consistent with X, then the agent's preference between the two acts depends only on their consequences when X obtains.
4. **Wagers** For consequences O_1 and O_2 and any event X, there is an act [O_1 if X, O_2 else] that produces O_1 in any state that entails X and O_2 in any state that entails $\neg X$
5. **Savage's P4** If the agent prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] when O_1 is more desirable than O_2, then she will also prefer [O_1^* if X, O_2^* else] to [O_1^* if Y, O_2^* else] for any other outcomes such that O_1^* is more desirable than O_2^*.

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant.

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew *that the Democratic candidate were going to win, and decides that he would.*

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew *that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he* knew *a Republican candidate were going to win, and again he finds that he would.*

The Sure-Thing Principle

> *A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he* knew *that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he* knew *a Republican candidate were going to win, and again he finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say.* (Savage, 1954)

Representation Theorem

If an agent satisfies all of the above postulates (including some technical ones not discussed), then the agent acts *as if* she is maximizing an expected utility.

These axioms (along with a few others) guarantee the existence of a unique probability P and utility u, unique up to the arbitrary choice of a unit and zero-point, whose associated expectation represents the agent's preferences.

Defining Beliefs from Preferences

Savage's P4 If the agent prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] when O_1 is more desirable than O_2, then she will also prefer [O_1^* if X, O_2^* else] to [O_1^* if Y, O_2^* else] for any other outcomes such that O_1^* is more desirable than O_2^*.

Defining Beliefs from Preferences

Savage's P4 If the agent prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] when O_1 is more desirable than O_2, then she will also prefer [O_1^* if X, O_2^* else] to [O_1^* if Y, O_2^* else] for any other outcomes such that O_1^* is more desirable than O_2^*.

Definition A practically rational agent **believes** X **more strongly than she believes** Y if and only if she strictly prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] for some (hence any by P4) outcome with O_1 more desirable than O_2.

Defining Beliefs from Preferences

Savage's P4 If the agent prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] when O_1 is more desirable than O_2, then she will also prefer [O_1^* if X, O_2^* else] to [O_1^* if Y, O_2^* else] for any other outcomes such that O_1^* is more desirable than O_2^*.

Definition A practically rational agent **believes** X **more strongly than she believes** Y if and only if she strictly prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] for some (hence any by P4) outcome with O_1 more desirable than O_2.

If O_1 is preferred to O_2 then the agent *has a good reason* for preferring [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] exactly if she is more confident in X than in Y.

Are the Axioms Requirements of Practical Rationality?

I. Gilboa. *Questions in Decision Theory*. Annual Reviews in Economics, 2010.

Issues

- Decision makers are sensitive to *risk* and *ambiguity* in ways that contradict standard expected utility calculations (Allais, Ellsberg)

- Decision makers are sensitive to *framing effects*

Issues

- Decision makers are sensitive to *risk* and *ambiguity* in ways that contradict standard expected utility calculations (Allais, Ellsberg)
- Decision makers are sensitive to *framing effects*
- The decision makers expected utility calculations should be sensitive to an agent's judgements about the probable causal consequences of the available options.

Issues

- Decision makers are sensitive to *risk* and *ambiguity* in ways that contradict standard expected utility calculations (Allais, Ellsberg)
- Decision makers are sensitive to *framing effects*
- The decision makers expected utility calculations should be sensitive to an agent's judgements about the probable causal consequences of the available options. (Newcomb's Paradox)

Newcomb's Paradox

Two boxes in front of you, A and B.

Box A contains \$1,000 and box B contains either \$1,000,000 or nothing.

Newcomb's Paradox

Two boxes in front of you, A and B.

Box A contains \$1,000 and box B contains either \$1,000,000 or nothing.

Your choice: either open both boxes, or else just open B. (You can keep whatever is inside any box you open, but you may not keep what is inside a box you do not open).

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.

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).

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

Newcomb's Paradox

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

Dominance reasoning is appropriate only when probability of outcome is *independent of choice*. (A nasty nephew wants inheritance from his rich Aunt.

Newcomb's Paradox

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

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.

Newcomb's Paradox

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

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?

Newcomb's Paradox

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

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

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

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.*)

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

Newcomb's Problem: Causal Decision Theory

$V(A) = \sum_w V(w) \cdot P_A(w)$
(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)

Newcomb's Problem: Causal Decision Theory

$V(A) = \sum_w V(w) \cdot P_A(w)$
(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: $P_A(w) = P(A \mathrel{\Box\!\!\rightarrow} w)$ (Probability of *if A were chosen then w would be true*)

Newcomb's Problem: Causal Decision Theory

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

Newcomb's Problem: Causal Decision Theory

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

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

Newcomb's Problem: Causal Decision Theory

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

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

Newcomb's Problem: Causal Decision Theory

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

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

Newcomb's Problem: Causal Decision Theory

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

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

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

Newcomb's Problem: Causal Decision Theory

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

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

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

Newcomb's Problem: Causal Decision Theory

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

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

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

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \,\Box\!\!\rightarrow M$ (and $B_2 \,\Box\!\!\rightarrow L$) (both conditional 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

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \Box\!\!\rightarrow M$ (and $B_2 \Box\!\!\rightarrow L$) (both conditional 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

$$V(B_1) = V(M)P(B_1 \Box\!\!\rightarrow M) + V(N)P(B_1 \Box\!\!\rightarrow N)$$

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \Box\!\!\rightarrow M$ (and $B_2 \Box\!\!\rightarrow L$) (both conditional 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

$V(B_1) = V(M)P(B_1 \Box\!\!\rightarrow M) + V(N)P(B_1 \Box\!\!\rightarrow N) =$
$1000000 \cdot \mu + 0 \cdot 1 - \mu$

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \,\Box\!\!\rightarrow M$ (and $B_2 \,\Box\!\!\rightarrow L$) (both conditional 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

$V(B_1) = V(M)P(B_1 \,\Box\!\!\rightarrow M) + V(N)P(B_1 \,\Box\!\!\rightarrow N) =$
$1000000 \cdot \mu + 0 \cdot 1 - \mu = 1000000\mu$

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \boxright M$ (and $B_2 \boxright L$) (both conditional 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

$V(B_1) = V(M)P(B_1 \boxright M) + V(N)P(B_1 \boxright N) =$
$1000000 \cdot \mu + 0 \cdot 1 - \mu = 1000000\mu$

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

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \Box\!\!\rightarrow M$ (and $B_2 \Box\!\!\rightarrow L$) (both conditional 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

$V(B_1) = V(M)P(B_1 \Box\!\!\rightarrow M) + V(N)P(B_1 \Box\!\!\rightarrow N) = 1000000 \cdot \mu + 0 \cdot 1 - \mu = 1000000\mu$

$V(B_2) = V(L)P(B_2 \Box\!\!\rightarrow L) + V(K)P(B_2 \Box\!\!\rightarrow K) = 1001000 \cdot \mu + 1000 \cdot 1 - \mu$

Newcomb's Problem: Causal Decision Theory

Let μ be the assigned to the conditional $B_1 \mathrel{\Box\!\!\rightarrow} M$ (and $B_2 \mathrel{\Box\!\!\rightarrow} L$) (both conditional 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

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

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

Conclusions, II

- If people are *really awful* and calculating probabilities, then it certainly does not help to understand their actions in terms of maximizing expected utility

Conclusions, II

- If people are *really awful* and calculating probabilities, then it certainly does not help to understand their actions in terms of maximizing expected utility (BUT, when mistakes are pointed out people tend to adjust their probabilities, and if the cases are described in terms of *frequencies*, then people are much better)

Conclusions, II

- If people are *really awful* and calculating probabilities, then it certainly does not help to understand their actions in terms of maximizing expected utility (BUT, when mistakes are pointed out people tend to adjust their probabilities, and if the cases are described in terms of *frequencies*, then people are much better)
- We need an account of which distinctions are relevant and which are not...what justifies a preference.

Conclusions, II

- If people are *really awful* and calculating probabilities, then it certainly does not help to understand their actions in terms of maximizing expected utility (BUT, when mistakes are pointed out people tend to adjust their probabilities, and if the cases are described in terms of *frequencies*, then people are much better)

- We need an account of which distinctions are relevant and which are not...what justifies a preference.

- Utility theory is a way to formalize and model rational action, but it is not itself a complete theory of rational action.

J. Pollock. *Rational Choice and Action Omnipotence*. The Philosophical Review, Vol. 111, No. 1 (2002), pgs. 1 - 23.

Next Week: Game Theory