Rationality

Lecture 7

Eric Pacuit

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

October 15, 2010

Rationality

What does it mean to be *rational* or *reasonable* as opposed to *irrational* or unreasonable?

Rationality

What does it mean to be *rational* or *reasonable* as opposed to *irrational* or unreasonable?

Rationality designates a capacity or set of capacities:

Rationality

What does it mean to be *rational* or *reasonable* as opposed to *irrational* or unreasonable?

Rationality designates a capacity or set of capacities: an agent is **rational** to the degree that he or she possesses and manifests the relevant range of capacities this involves.

Rationality

What does it mean to be *rational* or *reasonable* as opposed to *irrational* or unreasonable?

Rationality designates a capacity or set of capacities: an agent is **rational** to the degree that he or she possesses and manifests the relevant range of capacities this involves.

- the capacity to recognize or make correct judgements about reasons and other normative facts or truths

Rationality

What does it mean to be *rational* or *reasonable* as opposed to *irrational* or unreasonable?

Rationality designates a capacity or set of capacities: an agent is **rational** to the degree that he or she possesses and manifests the relevant range of capacities this involves.

- the capacity to recognize or make correct judgements about reasons and other normative facts or truths
- the capacity to *reason* well — to engage in valid forms of reasoning, to have one's reflections and deliberations proceed in ways that satisfy various formal constraints.

Key Issues

epistemic/theoretical vs. pragmatic/practical rationality

Key Issues

epistemic/theoretical vs. pragmatic/practical rationality

- what is rational for an agent to believe (be certain of, accept, know, etc.)
- what is rational for an agent to *do* (intend)?

Key Issues

epistemic/theoretical vs. pragmatic/practical rationality

- what is rational for an agent to believe (be certain of, accept, know, etc.)
- what is rational for an agent to *do* (intend)?

diachronic vs. synchronic rationality

Key Issues

epistemic/theoretical vs. pragmatic/practical rationality

- what is rational for an agent to believe (be certain of, accept, know, etc.)
- what is rational for an agent to *do* (intend)?

diachronic vs. synchronic rationality

- constraints on the way mental states change over time
- constraints on occurrent mental states

Key Issues

epistemic/theoretical vs. pragmatic/practical rationality

- what is rational for an agent to believe (be certain of, accept, know, etc.)
- what is rational for an agent to *do* (intend)?

diachronic vs. synchronic rationality

- constraints on the way mental states change over time
- constraints on occurrent mental states

normative vs. prescriptive vs. descriptive

Rational Beliefs

Beliefs can represent the world more or less accurately....the more accurate the better.

Rational Beliefs

Beliefs can represent the world more or less accurately....the more accurate the better.

But we can also judge some beliefs as being more *rational* than others.

Rational Beliefs

Beliefs can represent the world more or less accurately....the more accurate the better.

But we can also judge some beliefs as being more *rational* than others.

Accuracy and rationality are linked, they are not the same:

Rational Beliefs

Beliefs can represent the world more or less accurately....the more accurate the better.

But we can also judge some beliefs as being more *rational* than others.

Accuracy and rationality are linked, they are not the same: a fool may hold a belief irrationally — as a result of a lucky guess or wishful thinking — yet it might happen to be correct.

Rational Beliefs

Beliefs can represent the world more or less accurately....the more accurate the better.

But we can also judge some beliefs as being more *rational* than others.

Accuracy and rationality are linked, they are not the same: a fool may hold a belief irrationally — as a result of a lucky guess or wishful thinking — yet it might happen to be correct. Conversely, a detective might hold a belief on the basis of a careful and exhaustive examination of all the evidence and yet the evidence may be misleading, and the belief may turn out to be wrong.

Theoretical Reasoning

Rational beliefs are those that arise from **good thinking**, whether or not that thinking was successful in latching on to the truth.

But, what is **good thinking**?

Theoretical Reasoning

Rational beliefs are those that arise from **good thinking**, whether or not that thinking was successful in latching on to the truth.

But, what is **good thinking**?

- classical logic (modus ponens, modus tollens, etc.)
- non-monotonic/default logic
- closed-world reasoning
- induction (induction from examples)
- Bayesian inference
- case-reasoning/reasoning by analogy
- fast and frugal heuristics

Rational Actions

Belief/Desire Psychology: A practically rational agent will always act in ways that she estimates will best satisfy her desires

Rational Actions

Belief/Desire Psychology: A practically rational agent will always act in ways that she estimates will best satisfy her desires

But, actions are affected by emotions, habits, decision-making heuristics, and judgmental bias....

Rational Actions

Belief/Desire Psychology: A practically rational agent will always act in ways that she estimates will best satisfy her desires

But, actions are affected by emotions, habits, decision-making heuristics, and judgmental bias....

what *makes* an act rational is that it bears the **right relationship** to the actor's beliefs and desires.

Rational Actions

Belief/Desire Psychology: A practically rational agent will always act in ways that she estimates will best satisfy her desires

But, actions are affected by emotions, habits, decision-making heuristics, and judgmental bias....

what *makes* an act rational is that it bears the **right relationship** to the actor's beliefs and desires.

- Maximize expected utility

$$\sum_{o \in Out} [\text{how likely the act will lead to } o] \times [\text{how much the agent desires } o]$$

- Dominance reasoning
  a rational agent will not choose an action that *guarantees* a "sub-optimal outcome"

Instrumental Reasoning

Instrumental Reasoning

1. I ought to drink a beer
2. The necessary means for drinking a beer is going to a bar
3. I ought to go to the bar.

Instrumental Reasoning

1. I ought to drink a beer
2. The necessary means for drinking a beer is going to a bar
3. I ought to go to the bar.

1. I shall drink a bear
2. the necessary means to my drinking a beer is that I go to the bar
3. I shall go to the bar

Instrumental Reasoning

1. I ought to drink a beer
2. The necessary means for drinking a beer is going to a bar
3. I ought to go to the bar. *belief*

1. I shall drink a bear
2. the necessary means to my drinking a beer is that I go to the bar
3. I shall go to the bar *intention*

Intentions

Important distinctions:

1. (Present-directed) The intention with which someone acts
2. (Present-directed) Intentional action
3. (Future-directed) Intending to do some action

Intentions

Important distinctions:

1. (Present-directed) The intention with which someone acts
2. (Present-directed) Intentional action
3. (Future-directed) Intending to do some action

Some issues:

- Unifying account of *intentions*

  "Where we are tempted to speak of 'different senses' of a word which is clearly not equivocal, we may infer that we are pretty much in the dark about the character of the concept which it represents"
  - G.E.M. Anscombe, *Intention*, pg. 1

Intentions

Important distinctions:

1. (Present-directed) The intention with which someone acts
2. (Present-directed) Intentional action
3. (Future-directed) Intending to do some action

Some issues:

- Unifying account of *intentions*
- Intention as a *mental state*

  pro-attitude (vs. informational attitude), direction of fit, *conduct-controlling*

Intentions

Important distinctions:

1. (Present-directed) The intention with which someone acts
2. (Present-directed) Intentional action
3. (Future-directed) Intending to do some action

Some issues:

- Unifying account of *intentions*
- Intention as a *mental state*
- Intentions are (always) directed towards *actions*
  "Although we sometimes report intention as a propositional attitude — 'I intend that *p*' — such reports can always be recast as 'intending to ....' as when I intend to bring about that *p*. By contrast, it is difficult to rephrase such mundane expressions as 'I intend to walk home' in propositional terms"

Intentions

Important distinctions:

1. (Present-directed) The intention with which someone acts
2. (Present-directed) Intentional action
3. (Future-directed) Intending to do some action

Some issues:

- Unifying account of *intentions*
- Intention as a *mental state*
- Intentions are (always) directed towards *actions*

An extensive literature:

K. Setiya. *Intention*. Stanford Encyclopedia of Philosophy (2010).

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

“intention is a distinctive practical attitude marked by its pivotal role in planning for the future.

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

"intention is a distinctive practical attitude marked by its pivotal role in planning for the future. Intention involves desire, but even predominant desire is insufficient for intention, since it need not involve a commitment to act:

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

"intention is a distinctive practical attitude marked by its pivotal role in planning for the future. Intention involves desire, but even predominant desire is insufficient for intention, since it need not involve a commitment to act: intentions are conduct-controlling pro-attitudes, ones which we are disposed to retain without reconsideration, and which play a significant role as inputs to [means-end] reasoning" (pg. 20)

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

Committing to an action in advance is crucial for

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

Committing to an action in advance is crucial for

1. our capacity to make rational decisions (as a *bounded agent*)

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

Committing to an action in advance is crucial for

1. our capacity to make rational decisions (as a *bounded agent*)
2. our capacity to engage in complex, temporally extended projects

Functional Description of Intentions

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

Committing to an action in advance is crucial for

1. our capacity to make rational decisions (as a *bounded agent*)
2. our capacity to engage in complex, temporally extended projects
3. our capacity to coordinate with others

Stability of Plans

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

plans normally resist reconsideration:

Stability of Plans

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

plans normally resist reconsideration: "*an agent's habits and dispositions concerning the reconsideration or nonreconsideration of a prior intention or plan determine the stability of that intention or plan*".

Stability of Plans

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press (1987).

plans normally resist reconsideration: "*an agent's habits and dispositions concerning the reconsideration or nonreconsideration of a prior intention or plan determine the stability of that intention or plan*". .... "*The stability of [the agent's] plans will generally not be an isolated feature of those plans but will be linked to other features of [the agent's] psychology*" (pg. 65)

Intentions and beliefs are *entangled*

Intentions and beliefs are *entangled*

1. Intending to act just *is* a special kind of belief that one will;

Intentions and beliefs are *entangled*

1. Intending to act just *is* a special kind of belief that one will;

2. Intending to act *involves* a belief that one will so act;

Intentions and beliefs are *entangled*

1. Intending to act just *is* a special kind of belief that one will;
2. Intending to act *involves* a belief that one will so act;
3. Intending to act involves a belief that it is *possible* that one will so act.

Rationality constraints on intentions

Rationality constraints on intentions

1. *Consistency*: “one’s intentions, taken together with one’s beliefs fit together into a consistent model of one’s future”

Rationality constraints on intentions

1. *Consistency*: "one's intentions, taken together with one's beliefs fit together into a consistent model of one's future"

2. *Means-ends consistency*: "it is irrational that one intends E, , believes that E requires that one intend means M and yet not intend M"

Rationality constraints on intentions

1. *Consistency*: "one's intentions, taken together with one's beliefs fit together into a consistent model of one's future"

2. *Means-ends consistency*: "it is irrational that one intends E, believes that E requires that one intend means M and yet not intend M"

3. *Agglomeration*: "Intending A and Intending B implies Intending (A and B)"

M. Bratman. *Intention, Belief, Practical, Theoretical*. in *Spheres of Reason* (2009).

Reasoning

"Reasoning is not the conscious rehearsal of argument; it is a process in which antecedent beliefs and intentions are minimally modified, by addition and subtraction, in the interests of explanatory coherence and the satisfaction of intrinsic desires."
(G. Harman, pg. 56, "Practical Reasoning")

Philosophy of Normativity

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire:

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire: ought, should, must, duty, obligation, right, wrong;

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire: ought, should, must, duty, obligation, right, wrong; claims about what is justified, warranted, merited, reasonable, permissible;

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire: ought, should, must, duty, obligation, right, wrong; claims about what is justified, warranted, merited, reasonable, permissible; evaluative concepts: good, bad, better, and worse;

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire: ought, should, must, duty, obligation, right, wrong; claims about what is justified, warranted, merited, reasonable, permissible; evaluative concepts: good, bad, better, and worse; reasons

Philosophy of Normativity

How should we understand what it means to be **rationally required** to be in a relevant attitudinal state of mind?

Normative repertoire: ought, should, must, duty, obligation, right, wrong; claims about what is justified, warranted, merited, reasonable, permissible; evaluative concepts: good, bad, better, and worse; reasons

- How do we make sense of the fact that deliberative reflection can directly give rise to action?
- Which norms for the assessment of action are binding on us as agents? What about *moral norms*?
- Which normative attitude is "primary"? (ought, reason)

Philosophy of Normativity: Two Issues

Philosophy of Normativity: Two Issues

1. Internal vs. external reasons: there is a reason for A to φ:

   1.1 *Internal*: A has some motive which will be served furthered by his φ-ing. and if this turns out not to be so the sentence is false

   1.2 *External*: there is no such condition, and the reason-sentence will not be falsified by the absence of an appropriate motive.

Philosophy of Normativity: Two Issues

1. Internal vs. external reasons: there is a reason for A to φ:
   1.1 *Internal*: A has some motive which will be served furthered by his φ-ing. and if this turns out not to be so the sentence is false
   1.2 *External*: there is no such condition, and the reason-sentence will not be falsified by the absence of an appropriate motive.

2. The problem of *bootstrapping*

Rational Constraints on Beliefs

Rational Constraints on Beliefs

Conceptions of Beliefs

- **Binary**: "all-out" belief. For any statement p, the agent either does or does not believe p. It is natural to take an *unqualified* assertion as a statement of belief of the speaker.

- **Graded**: beliefs come in degrees. We are *more confident* in some of our beliefs than in others.

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

If an agent satisfies certain 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.

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^*.

Are the Axioms Requirements of Practical Rationality?

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

- The decision makers expected utility calculations should be sensitive to an agent's judgements about the probable causal consequences of the available options.

- Decision makers are sensitive to *risk* and *ambiguity* in ways that contradict standard expected utility calculations

- Decision makers are sensitive to *framing effects*

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: $P_A(w) = P(A \,\Box\!\!\rightarrow w)$ (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 $B_1 \mathrel{\Box\!\!\rightarrow} M$ (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

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

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 $B_1 \Box\!\!\rightarrow M$ (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

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

$V(B_2) = V(L)P(B_2 \Box\!\!\rightarrow L) + V(K)P(B_2 \Box\!\!\rightarrow K)$

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

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

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

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

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

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

Allais Paradox

M. Allais. *Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine*. Econometrica 21, 503-546, 1953.

Allais Paradox

Suppose there are three possible outcomes:

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

A lottery is a triple (p_1, p_2, p_3) meaning the player wins O_i with probability p_i.

Allais Paradox

Suppose there are three possible outcomes:

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

A lottery is a triple (p_1, p_2, p_3) meaning the player wins O_i with probability p_i.

Which lottery do you prefer?

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?

Allais Paradox

Suppose there are three possible outcomes:

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

A lottery is a triple (p_1, p_2, p_3) meaning the player wins O_i with probability p_i.

Which lottery do you prefer?

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?
2. $L_3 = (0.90, 0.00, 0.10)$ or $(0.89, 0.11, 0.00)$?

Allais Paradox

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?
2. $L_3 = (0.90, 0.00, 0.10)$ or $(0.89, 0.11, 0.00)$?

Allais Paradox

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?

2. $L_3 = (0.90, 0.00, 0.10)$ or $(0.89, 0.11, 0.00)$?

Many subjects report $L_1 \succ L_2$ and $L_3 \succ L_4$.

Allais Paradox

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?

2. $L_3 = (0.90, 0.00, 0.10)$ or $(0.89, 0.11, 0.00)$?

Many subjects report $L_1 \succ L_2$ and $L_3 \succ L_4$.

Why does this contradict standard expected utility calculations?

Allais Paradox

1. O_1 you receive \$0
2. O_2 you receive \$1M
3. O_3 you receive \$5M

1. $L_1 = (0.00, 1.00, 0.00)$ or $L_2 = (0.01, 0.89, 0.10)$?

2. $L_3 = (0.90, 0.00, 0.10)$ or $(0.89, 0.11, 0.00)$?

Many subjects report $L_1 \succ L_2$ and $L_3 \succ L_4$.

Why does this contradict standard expected utility calculations?
(Explanation on the next slide)

Allais Paradox

If $L_1 \succ L_2$ and the decision makers is maximizing expected utility, then we have $0.00 \cdot u_0 + 1.00 \cdot u_{1M} + 0.00 \cdot u_{5M} > 0.01 \cdot u_0 + 0.89 \cdot u_{1M} + 0.10 \cdot u_{5M}$. So, (after some algebraic manipulations)

$$0.11 \cdot u_{1M} > 0.01 \cdot u_0 + 0.10 u_{5M}$$

If $L_3 \succ L_4$ and the decision makers is maximizing expected utility, then we have $0.90 \cdot u_0 + 0.00 \cdot u_{1M} + 0.10 \cdot u_{5M} > 0.89 \cdot u_0 + 0.11 \cdot u_{1M} + 0.00 \cdot u_{5M}$. So, (after some algebraic manipulations)

$$0.01 \cdot u_0 + 0.10 \cdot u_{5M} > 0.11 \cdot u_{1M}$$

Putting these inequalities together, we have

$$0.11 \cdot u_{1M} > 0.01 \cdot u_0 + 0.10 u_{5M} > 0.11 \cdot u_{1M}$$

which implies $0.11 \cdot u_{1M} > 0.11 \cdot u_{1M}$, which is a contradiction.

Next Week: No Class (Break). **See the website for the midterm exam**.