Conditionals in Game Theory

Ilaria Canavotto, University of Maryland
Eric Pacuit, University of Maryland

Lecture 2

ESSLLI 2022

Yesterday: introduction to game theory

1. Normal form games
2. Pure strategies, mixed strategies, expected utilities
3. Best response and Nash equilibrium
4. Correlated equilibrium
5. Game models
6. Bayesian rationality
7. Aumann's 1987 theorem

   (Given CPA, correlated equilibria can be viewed as resulting from Bayesian rationality)

Plan for today

1. Bayesian rationality and counterfactual rationality
2. Stalnaker-Lewis semantics for counterfactuals
3. Bayesian rationality $\neq$ counterfactual rationality
4. Bayesian rationality $=$ counterfactual rationality given independence
5. Counterfactual rationality and ratifiability
6. Shin's notion of counterfactual rationality

Bayesian rationality and counterfactual rationality

O. Board. *The Equivalence of Bayes and Causal Rationality*. Theory and Decision 61, pp. 1-19, 2006.

Game model with prior and posterior beliefs

Given a strategic-form game $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, a model of G is a tuple

$$\langle W, (I_i, p_i)_{i \in N}, \sigma \rangle$$

where

- W is a set of *possible worlds* (possible outcomes of the game)
- σ is a function $\sigma : W \to \Pi_{i \in N} S_i$ (recall notation: $\sigma_i(w)$ and $\sigma_{-i}(w)$)
- For each $i \in N$, $I_i : W \to \wp(W)$ is player i's information correspondence.
    - Truth: For all $w \in W$, $w \in I_i(w)$
    - Consistency: For all $w \in W$, $I_i(w) \neq \varnothing$
    - Fully introspective: For all $w, v \in W$, if $v \in I_i(w)$, then $I_i(w) = I_i(v)$
    - Own-choice knowledge: For all $w \in W$, $I_i(w) \subseteq [\sigma_i(w)]$
- For each $i \in N$, $p_i \in \Delta(W)$ is a probability measure on W

i's posterior beliefs at w

$$p_{i,w}([\varphi]) = p_i([\varphi] \mid I_i(w)) = \frac{p_i([\varphi] \cap I_i(w))}{p_i(I_i(w))}$$

- *Remark.* We assume that $p_i(w) > 0$ for all $w \in W$.

Bayesian rationality

- Player i is **Bayes rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(a, s_{-i})$$

Bayesian rationality

- Player i is **Bayes rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(a, s_{-i})$$

[E]ach player forms a subjective probability assessment over her opponents' strategy profiles by updating her prior with respect to her private information, which includes information about which strategy choice she will carry out. She then evaluates alternative strategy choices according to this probability assessment. (p.8)

Bayesian rationality

- Player i is **Bayesian rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w)) u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w)) u_i(a, s_{-i})$$

> *[E]ach player forms a subjective probability assessment over her opponents' strategy profiles by updating her prior with respect to her private information, which includes information about which strategy choice she will carry out. She then evaluates alternative strategy choices according to this probability assessment. (p.8)*

From Bayesian rationality to counterfactual rationality

[T]he various actions of each player might be inter-connected: my opponents' choices given that I play s_i might not be the same as they would have been had I chosen to play s_i'. Each player must consider what her opponents will do given her actual choice, and also what they would do if she were to choose something else. (p.8)

A causal expected utility calculus, then, depends on counterfactual sentences such as "if it were the case that player i chose strategy s_i, then it would be the case that her opponents chose strategy profile s_{-i}. (p.8)

Counterfactual rationality

- Player i is **Bayes rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) u_i(a, s_{-i})$$

- Player i is **counterfactually rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \,\Box\!\!\rightarrow [s_{-i}]) u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \,\Box\!\!\rightarrow [s_{-i}]) u_i(a, s_{-i})$$

Counterfactual rationality

- Player i is **Bayes rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w))\, u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w))\, u_i(a, s_{-i})$$

- Player i is **counterfactually rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(a, s_{-i})$$

Counterfactual rationality

- Player i is **Bayes rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w)) u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_i([s_{-i}] \mid I_i(w)) u_i(a, s_{-i})$$

- Player i is **counterfactually rational** at w if, for all $a \in S_i$:

???

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}]) u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]) u_i(a, s_{-i})$$

Stalnaker-Lewis semantics for counterfactuals

Basic ideas

> *This is how to evaluate a conditional: First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true...*

R. Stalnaker. *A theory of conditionals*. in Studies in Logical Theory, pp. 98–112, 1968.

Basic ideas

> *This is how to evaluate a conditional: First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true...*
>
> *... the problem is to make the transition from belief conditions to truth conditions... a possible world is the ontological analogue of a stock of hypothetical beliefs. The following set of truth conditions, using this notion, is a first approximation to the account I shall propose:*

R. Stalnaker. *A theory of conditionals*. in Studies in Logical Theory, pp. 98–112, 1968.

Basic ideas

> *This is how to evaluate a conditional: First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true...*
>
> *... the problem is to make the transition from belief conditions to truth conditions... a possible world is the ontological analogue of a stock of hypothetical beliefs. The following set of truth conditions, using this notion, is a first approximation to the account I shall propose:*
>
> *Consider a possible world in which A is true, and which otherwise differs minimally from the actual world. "If A, then B" is true (false) just in case B is true (false) in that possible world. (p.102).*

R. Stalnaker. *A theory of conditionals*. in Studies in Logical Theory, pp. 98–112, 1968.

Basic ideas

> *'If kangaroos had no tails, they would topple over' seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. I shall give a general analysis of counterfactual conditionals along these lines. (p.1)*

D. Lewis. *Counterfactuals*. Blackwell, 1973.

Semantics: selection functions

Let A and C be two events. Then the event $A \boxright C$ occurs at w if and only if...

- **Stalnaker 1968:** C occurs at **the** closest A-world to w.
  - For each $w \in W$, $f_w : \mathit{events} \rightarrow W$ selects, for each $E \in \mathit{events}$, the closest E-world to w.
  - $A \boxright C$ occurs at w iff $f_w(A) \in C$.

Semantics: selection functions

Let A and C be two events. Then the event $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w if and only if...

- **Stalnaker 1968:** C occurs at **the** closest A-world to w.
  - For each $w \in W$, $f_w : \mathit{events} \to W$ selects, for each $E \in \mathit{events}$, the closest E-world to w.
  - $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w iff $f_w(A) \in C$.
  - f_w satisfies:

    (1) $f_w(E) \in E$

    (2) $f_w(E) = w$ if $w \in E$

    (3) $f_w(E) = \lambda$ if $E = \varnothing$

    (4) if $f_w(E) \in E'$ and $f_w(E') \in E$, then $f_w(E) = f_w(E')$

Semantics: selection functions

Let A and C be two events. Then the event $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w if and only if...

- **Lewis 1973:** C occurs at **every** closest A-world to w.
  - For each $w \in W$, $f_w : \mathit{events} \rightarrow 2^W$ selects, for each $E \in \mathit{events}$, the set of closest E-worlds to w.
  - $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w iff $f_w(A) \subseteq C$.

Semantics: selection functions

Let A and C be two events. Then the event $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w if and only if...

- **Lewis 1973:** C occurs at **every** closest A-world to w.
  - For each $w \in W$, $f_w : \mathit{events} \rightarrow 2^W$ selects, for each $E \in \mathit{events}$, the set of closest E-worlds to w.
  - $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w iff $f_w(A) \subseteq C$.
  - f_w satisfies:

    (1) $f_w(E) \subseteq E$

    (2) $f_w(E) = \{w\}$ if $w \in E$

    (3) if $E \neq \varnothing$, then $f_w(E) \neq \varnothing$

    (4) if $E \subseteq E'$ and $E \cap f_w(E') \neq \varnothing$, then $f_w(E) = E \cap f_w(E')$

Semantics: relative closeness

Let A and C be two events. Then the event $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w if and only if...

- **Lewis 1973:** some $A \cap C$-world is closer to w than any $A \cap \overline{C}$-world, if there are any A-worlds.
  - For each $w \in W$, $\preceq_w \subseteq W \times W$ is a relation of relative closeness—read $v \preceq_w u$ as "v is at least as close to w as u."
  - $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w iff either $A = \varnothing$
    or there is $v \in A \cap C$ s.t., for all $u \in A \cap \overline{C}$, $u \preceq_w v$

Semantics: relative closeness

Let A and C be two events. Then the event $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w if and only if...

- **Lewis 1973:** some $A \cap C$-world is closer to w than any $A \cap \overline{C}$-world, if there are any A-worlds.
  - For each $w \in W$, $\preceq_w \subseteq W \times W$ is a relation of relative closeness—read $v \preceq_w u$ as "v is at least as close to w as u."
  - $A \mathrel{\Box\!\!\rightarrow} C$ occurs at w iff either $A = \varnothing$
    or there is $v \in A \cap C$ s.t., for all $u \in A \cap \overline{C}$, $u \preceq_w v$
  - $\preceq_w$ satisfies:

    (1) Centering: if $v \preceq_w w$, then $v = w$

    (2) Transitivity: if $v_1 \preceq_w v_2$ and $v_2 \preceq_w v_3$, then $v_1 \preceq_w v_3$

    (3) Linearity: either $v_1 \preceq_w v_2$ or $v_2 \preceq_w v_1$

Semantics: relative closeness

Two important (though controversial) additional properties that $\preceq_w$ could satisfy:

(4) Limit assumption: for each $E \in \mathit{events}$, if $E \neq \varnothing$, then $\mathit{min}_w(E) \neq \varnothing$, where

$$\mathit{min}_w(E) = \{v \in E \mid \text{for all } u \in E, v \preceq_w u\}$$

(5) Antisymmetry: if $v_1 \preceq_w v_2$ and $v_2 \preceq_w v_1$, then $v_1 = v_2$

Relative closeness and selection functions

Key fact. If $\preceq_w$ satisfies 1-4, then working with relations of relative similarity is the same as working with Lewis-style selection functions. If $\preceq_w$ satisfies 1-5 then working with relations of relative similarity is the same as working with Stalnaker-style selection functions.

D. Lewis. *Counterfactuals (Chapter 2)*. Blackwell, 1973.

G. Grahne. *Updates and counterfactuals*. J. Logic Computat., 8 (1), pp. 97-117, 1973.

Relative closeness and selection functions

Key fact. If $\preceq_w$ satisfies 1-4, then working with relations of relative similarity is the same as working with Lewis-style selection functions. If $\preceq_w$ satisfies 1-5 then working with relations of relative similarity is the same as working with Stalnaker-style selection functions.

Today we work with relations of relative closeness that satisfy 1-5.

D. Lewis. *Counterfactuals (Chapter 2)*. Blackwell, 1973.

G. Grahne. *Updates and counterfactuals*. J. Logic Computat., 8 (1), pp. 97-117, 1973.

Relative closeness and selection functions

Key fact. If $\preceq_w$ satisfies 1-4, then working with relations of relative similarity is the same as working with Lewis-style selection functions. If $\preceq_w$ satisfies 1-5 then working with relations of relative similarity is the same as working with Stalnaker-style selection functions.

Today we work with relations of relative closeness that satisfy 1-5.

- $min_w(E)$ is a singleton

D. Lewis. *Counterfactuals (Chapter 2)*. Blackwell, 1973.

G. Grahne. *Updates and counterfactuals*. J. Logic Computat., 8 (1), pp. 97-117, 1973.

Relative closeness and selection functions

Key fact. If $\preceq_w$ satisfies 1-4, then working with relations of relative similarity is the same as working with Lewis-style selection functions. If $\preceq_w$ satisfies 1-5 then working with relations of relative similarity is the same as working with Stalnaker-style selection functions.

Today we work with relations of relative closeness that satisfy 1-5.

- $min_w(E)$ is a singleton
- $A \mathrel{\Box\!\!\rightarrow} C = \{w \in W \mid min_w(A) \subseteq C\}$

D. Lewis. *Counterfactuals (Chapter 2)*. Blackwell, 1973.

G. Grahne. *Updates and counterfactuals*. J. Logic Computat., 8 (1), pp. 97-117, 1973.

Game model with relations of relative closeness

Given a strategic-form game $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, a model of G is a tuple

$$\langle W, (I_i)_{i \in N}, (p_i)_{i \in N}, \sigma, (\preceq_w)_{w \in W} \rangle$$

- W is a set of *possible worlds* (possible outcomes of the game)
- σ is a function $\sigma : W \to \Pi_{i \in N} S_i$ (recall notation: $\sigma_i(w)$ and $\sigma_{-i}(w)$)
  - Sufficiency: For each $i \in N$ and $a \in S_i$, there is $w \in W$ s.t. $\sigma_i(w) = a$.
- For each $i \in N$, $I_i : W \to \wp(W)$ is player i's information correspondence satisfying Truth, Consistency, Full introspection, Own-choice knowledge.
- For each $i \in N$, $p_i \in \Delta(W)$ is a probability measure on W
- For each $w \in W$, $\preceq_w \subseteq W \times W$ satisfies conditions 1-5.

Back to counterfactual rationality now!

What is the probability of $A \mathrel{\Box\!\!\rightarrow} C$?

Player i is **counterfactually rational** at w if, for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(\sigma_i(w), s_{-i}) \geqslant$$

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(a, s_{-i})$$

What is the probability of $A \mathrel{\Box\!\!\rightarrow} C$?

- We know that $A \mathrel{\Box\!\!\rightarrow} C = \{w \in W \mid min_w(A) \subseteq C\}$

- We can assume that agents form subjective beliefs for the event $A \mathrel{\Box\!\!\rightarrow} C$ as they form subjective beliefs for any other event:

$$p_{i,w}(A \mathrel{\Box\!\!\rightarrow} C) = p_i(A \mathrel{\Box\!\!\rightarrow} C \mid I_i(w)) = \frac{p_i(A \mathrel{\Box\!\!\rightarrow} C \cap I_i(w))}{p_i(I_i(w))}$$

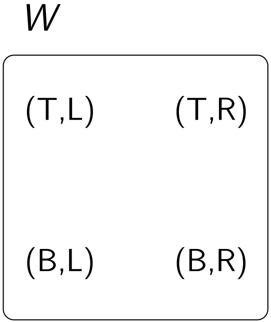
Bayesian rationality $\neq$ counterfactual rationality

		Bob	
		L	*R*
Ann	*T*	1,1	0,0
	B	0,0	2,2

Game *G*

		Bob	
		L	*R*
Ann	*T*	1,1	0,0
	B	0,0	2,2

Game *G*



W

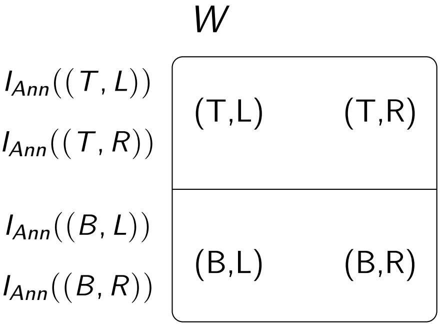
(T,L) (T,R)

(B,L) (B,R)

Model of *G*

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2

Game G



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

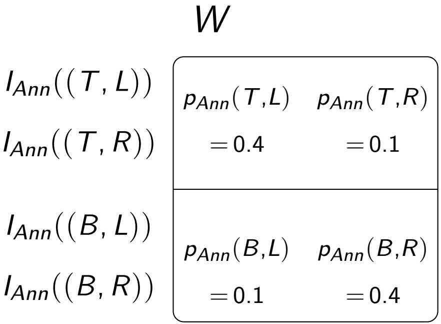
(T,L) (T,R)

(B,L) (B,R)

Model of G

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2

Game G



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

$I_{Ann}((B, L))$

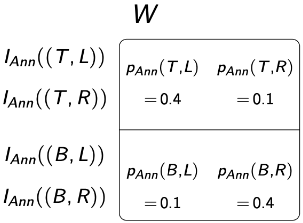
$I_{Ann}((B, R))$

$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

Model of G

Ann is Bayes rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(T,L) = 0.4$

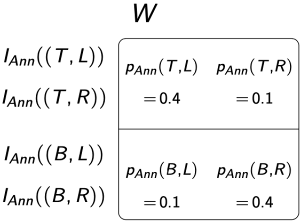
$p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$

$p_{Ann}(B,R) = 0.4$

Ann is Bayes rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

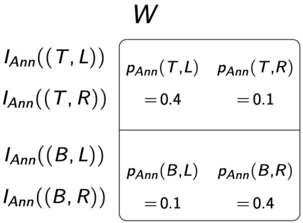
$p_{Ann}(T,L) = 0.4$ $\quad p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$ $\quad p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(T, L) + p_{Ann,(T,L)}([R]) \cdot u_{Ann}(T, R) \geqslant$$
$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(B, L) + p_{Ann,(T,L)}([R]) \cdot u_{Ann}(B, R)$$

Ann is Bayes rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

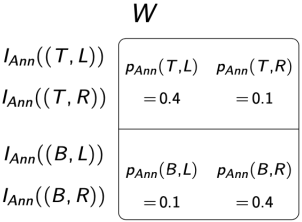
$p_{Ann}(T,L) = 0.4$ $\quad p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$ $\quad p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(T,L) + p_{Ann,(T,L)}([R]) \cdot u_{Ann}(T,R) \geqslant$$
$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(B,L) + p_{Ann,(T,L)}([R]) \cdot u_{Ann}(B,R)$$

Ann is Bayes rational at world (T, L)

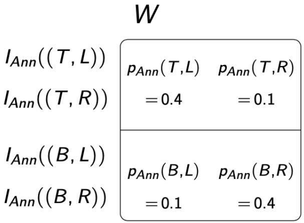
		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(T,L) + p_{Ann,(T,L)}([R]) \cdot 0 \geqslant$$
$$p_{Ann,(T,L)}([L]) \cdot 0 + p_{Ann,(T,L)}([R]) \cdot u_{Ann}(B,R)$$

Ann is Bayes rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(T,L) = 0.4$

$p_{Ann}(T,R) = 0.1$

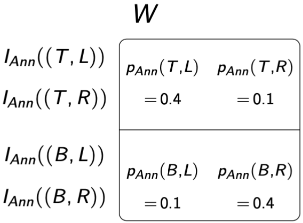
$p_{Ann}(B,L) = 0.1$

$p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([L]) \cdot u_{Ann}(T,L) \geqslant p_{Ann,(T,L)}([R]) \cdot u_{Ann}(B,R)$$

Ann is Bayes rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$p_{Ann}(T, L) = 0.4$

$p_{Ann}(T, R) = 0.1$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

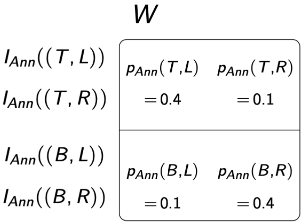
$p_{Ann}(B, L) = 0.1$

$p_{Ann}(B, R) = 0.4$

$$\frac{p_{Ann}([L] \cap I_{Ann}((T, L)))}{p_{Ann}(I_{Ann}((T, L)))} \cdot u_{Ann}(T, L) \geqslant \frac{p_{Ann}([R] \cap I_{Ann}((T, L)))}{p_{Ann}(I_{Ann}((T, L)))} \cdot u_{Ann}(B, R)$$

Ann is Bayes rational at world (T, L)

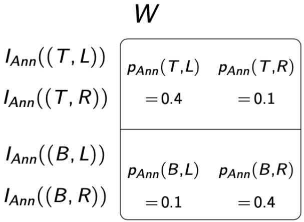
		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



$$0.8 \cdot 1 \geqslant 0.2 \cdot 2$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$p_{Ann}(T,L) = 0.4$ $\quad$ $p_{Ann}(T,R) = 0.1$

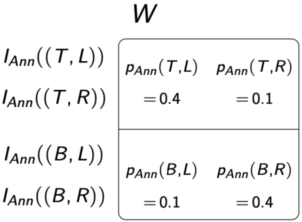
$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(B,L) = 0.1$ $\quad$ $p_{Ann}(B,R) = 0.4$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

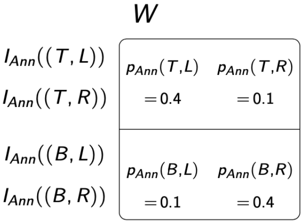
$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\square\!\!\rightarrow} [L]) \cdot u_{Ann}(T, L) + p_{Ann,(T,L)}([T] \mathrel{\square\!\!\rightarrow} [R]) \cdot u_{Ann}(T, R) \not\geqslant$$

$$p_{Ann,(T,L)}([B] \mathrel{\square\!\!\rightarrow} [L]) \cdot u_{Ann}(B, L) + p_{Ann,(T,L)}([B] \mathrel{\square\!\!\rightarrow} [R]) \cdot u_{Ann}(B, R)$$

Ann is **not** counterfactually rational at world (T, L)

Ann \ Bob	L	R
T	1,1	0,0
B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

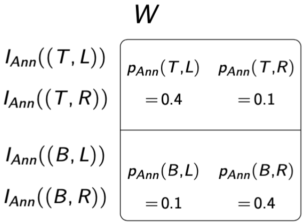
$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(T, L) + p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(T, R) \not\geq$$

$$p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(B, L) + p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(B, R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(T,L) = 0.4$

$p_{Ann}(T,R) = 0.1$

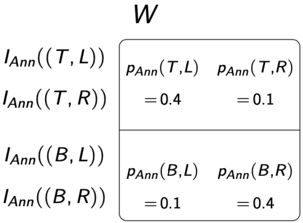
$p_{Ann}(B,L) = 0.1$

$p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(T, L) + p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [R]) \cdot 0 \not\geqslant$$
$$p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [L]) \cdot 0 + p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(B, R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(T,L) = 0.4$

$p_{Ann}(T,R) = 0.1$

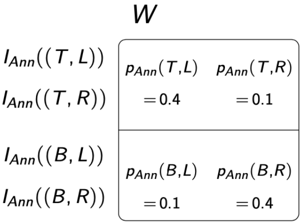
$p_{Ann}(B,L) = 0.1$

$p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(T,L) \not\geqslant$$
$$p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(B,R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(T,L) = 0.4$ $\quad p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$ $\quad p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(T,L) \not\geq$$
$$p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(B,R)$$

Ann is **not** counterfactually rational at world (T, L)

We know that, for any $a \in S_{Ann}$ and $b \in S_{Bob}$:

$$p_{Ann,(T,L)}([a] \mathrel{\Box\!\!\rightarrow} [b]) = \frac{p_{Ann}([a] \mathrel{\Box\!\!\rightarrow} [b] \cap I_{Ann}((T,L)))}{p_i(I_{Ann}((T,L)))}$$

$$= \frac{p_{Ann}([a] \mathrel{\Box\!\!\rightarrow} [b] \cap \{(T,L),(T,R)\})}{p_i(\{(T,L),(T,R)\})}$$

Ann is **not** counterfactually rational at world (T, L)

We know that, for any $a \in S_{Ann}$ and $b \in S_{Bob}$:

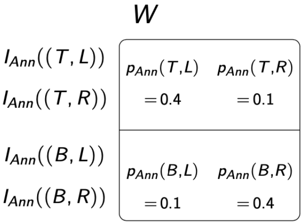
$$p_{Ann,(T,L)}([a] \,\Box\!\!\rightarrow [b]) = \frac{p_{Ann}([a] \,\Box\!\!\rightarrow [b] \cap I_{Ann}((T,L)))}{p_i(I_{Ann}((T,L)))}$$

$$= \frac{p_{Ann}([a] \,\Box\!\!\rightarrow [b] \cap \{(T,L),(T,R)\})}{p_i(\{(T,L),(T,R)\})}$$

So, in order to calculate $p_{Ann,(T,L)}([a] \,\Box\!\!\rightarrow [b])$, we only need to know whether $(T,L) \in [a] \,\Box\!\!\rightarrow [b]$ or $(T,R) \in [a] \,\Box\!\!\rightarrow [b]$; the other worlds can be disregarded.

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(T,L) = 0.4$ $\quad p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$ $\quad p_{Ann}(B,R) = 0.4$

Assume that:

$$(T, L) \preceq_{(T,L)} (B, R) \preceq_{(T,L)} (T, R) \preceq_{(T,L)} (B, L)$$

$$(T, R) \preceq_{(T,R)} (B, L) \preceq_{(T,R)} (T, L) \preceq_{(T,R)} (B, R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2

W

$I_{Ann}((T,L))$ $I_{Ann}((T,R))$	$p_{Ann}(T,L)$ $=0.4$	$p_{Ann}(T,R)$ $=0.1$
$I_{Ann}((B,L))$ $I_{Ann}((B,R))$	$p_{Ann}(B,L)$ $=0.1$	$p_{Ann}(B,R)$ $=0.4$

Assume that:

$$(T,L) \preceq_{(T,L)} (B,R) \preceq_{(T,L)} (T,R) \preceq_{(T,L)} (B,L)$$

$$(T,R) \preceq_{(T,R)} (B,L) \preceq_{(T,R)} (T,L) \preceq_{(T,R)} (B,R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2

W

$I_{Ann}((T,L))$ $I_{Ann}((T,R))$	$p_{Ann}(T,L)$ $=0.4$	$p_{Ann}(T,R)$ $=0.1$
$I_{Ann}((B,L))$ $I_{Ann}((B,R))$	$p_{Ann}(B,L)$ $=0.1$	$p_{Ann}(B,R)$ $=0.4$

Assume that:

$$(T, L) \preceq_{(T,L)} (B, R) \preceq_{(T,L)} (T, R) \preceq_{(T,L)} (B, L)$$

$$(T, R) \preceq_{(T,R)} (B, L) \preceq_{(T,R)} (T, L) \preceq_{(T,R)} (B, R)$$

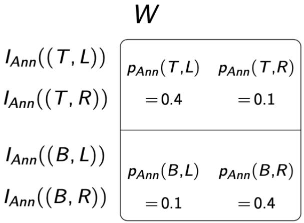
So:

$$[T] \mathrel{\Box\!\!\rightarrow} [L] \cap I_{Ann}((T, L)) = \{(T, L)\}$$

$$[B] \mathrel{\Box\!\!\rightarrow} [R] \cap I_{Ann}((T, L)) = \{(T, L)\}$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

Assume that:

$$(T,L) \preceq_{(T,L)} (B,R) \preceq_{(T,L)} (T,R) \preceq_{(T,L)} (B,L)$$
$$(T,R) \preceq_{(T,R)} (B,L) \preceq_{(T,R)} (T,L) \preceq_{(T,R)} (B,R)$$

So:

$[T] \mathrel{\Box\!\!\rightarrow} [L] \cap I_{Ann}((T,L)) = \{(T,L)\}$

$[B] \mathrel{\Box\!\!\rightarrow} [R] \cap I_{Ann}((T,L)) = \{(T,L)\}$

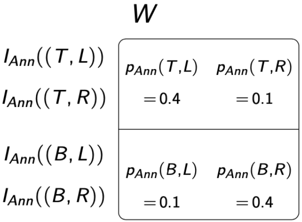
Before we had:

$[L] \cap I_{Ann}((T,L)) = \{(T,L)\}$

$[R] \cap I_{Ann}((T,L)) = \{(T,R)\}$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

$I_{Ann}((B, L))$

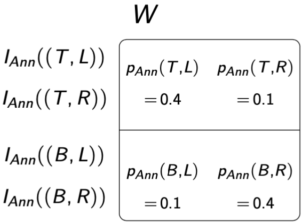
$I_{Ann}((B, R))$

$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

$$p_{Ann,(T,L)}([T] \mathrel{\Box\!\!\rightarrow} [L]) \cdot u_{Ann}(T, L) \not\geq p_{Ann,(T,L)}([B] \mathrel{\Box\!\!\rightarrow} [R]) \cdot u_{Ann}(B, R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T,L))$

$I_{Ann}((T,R))$

$I_{Ann}((B,L))$

$I_{Ann}((B,R))$

$p_{Ann}(T,L) = 0.4$

$p_{Ann}(T,R) = 0.1$

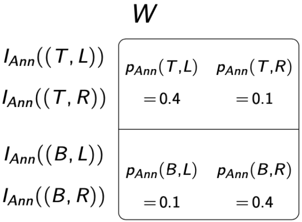
$p_{Ann}(B,L) = 0.1$

$p_{Ann}(B,R) = 0.4$

$$\frac{p_{Ann}(\{(T,L)\})}{p_{Ann}(\{(T,L),(T,R)\})} \cdot u_{Ann}(T,L) \ngeq \frac{p_{Ann}(\{(T,L)\})}{p_{Ann}(\{(T,L),(T,R)\})} \cdot u_{Ann}(B,R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

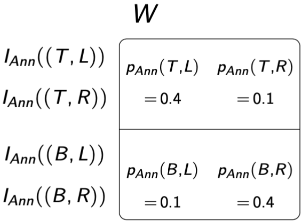
$p_{Ann}(T,L) = 0.4$ $p_{Ann}(T,R) = 0.1$

$p_{Ann}(B,L) = 0.1$ $p_{Ann}(B,R) = 0.4$

$$0.8 \cdot u_{Ann}(T, L) \not\succcurlyeq 0.8 \cdot u_{Ann}(B, R)$$

Ann is **not** counterfactually rational at world (T, L)

		Bob	
		L	R
Ann	T	1,1	0,0
	B	0,0	2,2



W

$I_{Ann}((T, L))$

$I_{Ann}((T, R))$

$p_{Ann}(T,L) = 0.4$ $\quad p_{Ann}(T,R) = 0.1$

$I_{Ann}((B, L))$

$I_{Ann}((B, R))$

$p_{Ann}(B,L) = 0.1$ $\quad p_{Ann}(B,R) = 0.4$

$$0.8 \cdot 1 \not\geqslant 0.8 \cdot 2$$

So, is Ann rational or not??

So, is Ann rational or not??

... the objects of uncertainty faced by the player (in this case her opponent's strategy) are not independent of the various acts available to her and Bayes rationality gives us the "wrong" result (that is, it does not coincide with causal [i.e. counterfactual] rationality).

$$(T,L) \preceq_{(T,L)} (B,R) \preceq_{(T,L)} (T,R) \preceq_{(T,L)} (B,L)$$

$$(T,R) \preceq_{(T,R)} (B,L) \preceq_{(T,R)} (T,L) \preceq_{(T,R)} (B,R)$$

So, is Ann rational or not??

... the objects of uncertainty faced by the player (in this case her opponent's strategy) are not independent of the various acts available to her and Bayes rationality gives us the "wrong" result (that is, it does not coincide with causal [i.e. counterfactual] rationality).
But *there is something odd about the causal structure of the game. If the players are moving simultaneously, or at least in ignorance of each other's choice ... then their strategy choice should be independent of each other. (p. 11)*

$$(T,L) \preceq_{(T,L)} (B,R) \preceq_{(T,L)} (T,R) \preceq_{(T,L)} (B,L)$$

$$(T,R) \preceq_{(T,R)} (B,L) \preceq_{(T,R)} (T,L) \preceq_{(T,R)} (B,R)$$

Bayesian rationality = counterfactual rationality
given Independence

Formal definition of independence

Let $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic-form game.

Let $M = \langle W, (I_i)_{i \in N}, \sigma, (\preceq_w)_{w \in W} \rangle$ be a model of G.

Formal definition of independence

Let $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic-form game.
Let $M = \langle W, (I_i)_{i \in N}, \sigma, (\preceq_w)_{w \in W} \rangle$ be a model of G.

Independence: for all $w, v \in W$, $i \in N$, and $a \in S_i$

if $v \in min_w([a])$, then $\sigma_{-i}(w) = \sigma_{-i}(v)$.

Formal definition of independence

Let $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic-form game.
Let $M = \langle W, (I_i)_{i \in N}, \sigma, (\preceq_w)_{w \in W} \rangle$ be a model of G.

Independence: for all $w, v \in W$, $i \in N$, and $a \in S_i$

if $v \in min_w([a])$, then $\sigma_{-i}(w) = \sigma_{-i}(v)$.

$$(T,L) \preceq_{(T,L)} (B,R) \preceq_{(T,L)} (T,R) \preceq_{(T,L)} (B,L)$$

$$(T,R) \preceq_{(T,R)} (B,L) \preceq_{(T,R)} (T,L) \preceq_{(T,R)} (B,R)$$

Formal definition of independence

Let $G = \langle N, (S_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a strategic-form game.
Let $M = \langle W, (I_i)_{i\in N}, \sigma, (\preceq_w)_{w\in W}\rangle$ be a model of G.

Independence: for all $w, v \in W$, $i \in N$, and $a \in S_i$

if $v \in min_w([a])$, then $\sigma_{-i}(w) = \sigma_{-i}(v)$.

$$(T, L) \preceq_{(T,L)} (B, R) \preceq_{(T,L)} (T, R) \preceq_{(T,L)} (B, L)$$
$$(T, R) \preceq_{(T,R)} (B, L) \preceq_{(T,R)} (T, L) \preceq_{(T,R)} (B, R)$$

Independence ensures that the strategies of the players are causally independent of one another AND that there is common belief in causal independence.

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all w and for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])u_i(a, s_{-i})$$

if and only if

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}])u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}])u_i(a, s_{-i})$$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all w and for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])\, u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}])\, u_i(a, s_{-i})$$

if and only if

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}])\, u_i(a, s_{-i})$$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all w and for all $a \in S_i$:

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) u_i(a, s_{-i})$$

if and only if

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([\sigma_i(w)] \mathrel{\Box\!\!\rightarrow} [s_{-i}]) u_i(\sigma_i(w), s_{-i}) \geqslant \sum_{s_{-i} \in S_{-i}} p_{i,w}([a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]) u_i(a, s_{-i})$$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$

- $[s_{-i}] \subseteq [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$ (immediate)

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$

- $[s_{-i}] \subseteq [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$ (immediate)
- $[a] \mathrel{\Box\!\!\rightarrow} [s_{-i}] \subseteq [s_{-i}]$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \Box\!\!\rightarrow [s_{-i}]$

- $[s_{-i}] \subseteq [a] \Box\!\!\rightarrow [s_{-i}]$ (immediate)
- $[a] \Box\!\!\rightarrow [s_{-i}] \subseteq [s_{-i}]$
  1. Suppose $w \in [a] \Box\!\!\rightarrow [s_{-i}]$, i.e. $min_w([a]) \subseteq [s_{-i}]$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \boxright [s_{-i}]$

- $[s_{-i}] \subseteq [a] \boxright [s_{-i}]$ (immediate)
- $[a] \boxright [s_{-i}] \subseteq [s_{-i}]$
  1. Suppose $w \in [a] \boxright [s_{-i}]$, i.e. $min_w([a]) \subseteq [s_{-i}]$
  2. By Sufficiency there is $v \in [a]$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$

- $[s_{-i}] \subseteq [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$ (immediate)
- $[a] \mathrel{\Box\!\!\rightarrow} [s_{-i}] \subseteq [s_{-i}]$

  1. Suppose $w \in [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$, i.e. $min_w([a]) \subseteq [s_{-i}]$
  2. By Sufficiency there is $v \in [a]$
  3. Hence, by the Limit Assumption, there is $v' \in min_w([a])$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \boxright [s_{-i}]$

- $[s_{-i}] \subseteq [a] \boxright [s_{-i}]$ (immediate)
- $[a] \boxright [s_{-i}] \subseteq [s_{-i}]$
  1. Suppose $w \in [a] \boxright [s_{-i}]$, i.e. $min_w([a]) \subseteq [s_{-i}]$
  2. By Sufficiency there is $v \in [a]$
  3. Hence, by the Limit Assumption, there is $v' \in min_w([a])$
  4. Hence, by Independence, $\sigma_{-i}(w) = \sigma_{-i}(v')$

Main theorem

Theorem. In any model satisfying Independence, player i is Bayes rational at w iff player i is counterfactually rational at w.

TBS: for all $a \in S_i$ and $s_{-i} \in S_{-i}$: $[s_{-i}] = [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$

- $[s_{-i}] \subseteq [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$ (immediate)
- $[a] \mathrel{\Box\!\!\rightarrow} [s_{-i}] \subseteq [s_{-i}]$

1. Suppose $w \in [a] \mathrel{\Box\!\!\rightarrow} [s_{-i}]$, i.e. $min_w([a]) \subseteq [s_{-i}]$
2. By Sufficiency there is $v \in [a]$
3. Hence, by the Limit Assumption, there is $v' \in min_w([a])$
4. Hence, by Independence, $\sigma_{-i}(w) = \sigma_{-i}(v')$
5. By 1, 3, and 4, $\sigma_{-i}(w) = \sigma_{-i}(v') = s_{-i}$

Counterfactual rationality and ratifiability

Another way to understand counterfactual rationality

> *A player should never find herself at a possible world at which ... her payoff would be higher if she were to deviate from the strategy she has chosen. This is the principle which motivates our rationality criterion. (p. 29)*

H.S. Shin. *A reconstruction of Jeffrey's notion of ratifiability in terms of counterfactual belief*. Theory and Decision 31, pp. 21-47, 1991.

Another way to understand counterfactual rationality

> *A player should never find herself at a possible world at which ... her payoff would be higher if she were to deviate from the strategy she has chosen. This is the principle which motivates our rationality criterion. (p. 29)*

- This phrasing recalls the idea of ratifiability

H.S. Shin. *A reconstruction of Jeffrey's notion of ratifiability in terms of counterfactual belief*. Theory and Decision 31, pp. 21-47, 1991.

The idea behind ratifiability

Ratifiability is a type of stability of decision:

> *The notion of ratifiability is applicable only where, during deliberation, the agent finds it conceivable that he will not manage to perform the act he finally decides to perform, but will find himself performing one of the other available acts instead...*

R.C. Jeffrey. *The Logic of Decision*. McGraw-Hill, New York, 1965..

The idea behind ratifiability

Ratifiability is a type of stability of decision:

> *The notion of ratifiability is applicable only where, during deliberation, the agent finds it conceivable that he will not manage to perform the act he finally decides to perform, but will find himself performing one of the other available acts instead...*
> *... The option in question is ratifiable or not depending on whether or not the expected desirability of actually carrying it out (having chosen it) is at least as great as the expected desirability of actually carrying out each of the alternatives (in spite of having chosen to carry out a different option, as hypothesized). (pp. 18-20)*

R.C. Jeffrey. *The Logic of Decision*. McGraw-Hill, New York, 1965..

A famous example: Newcomb's paradox

There are two boxes in front of us:

- box A, which contains \$1,000;
- box B, which contains either \$1,000,000 or nothing.

A famous example: Newcomb's paradox

There are two boxes in front of us:

- box A, which contains \$1,000;
- box B, which contains either \$1,000,000 or nothing.

We have two choices:

- we open only box B.
- we open both box A and box B;

A famous example: Newcomb's paradox

There are two boxes in front of us:

- box A, which contains $1,000;
- box B, which contains either $1,000,000 or nothing.

We have two choices:

- we open only box B.
- we open both box A and box B;

We can keep whatever is inside any box we open, but we may not keep what is inside a box that we do not open.

A famous example: Newcomb's paradox

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

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

What should we do?

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

A famous example: Newcomb's paradox

- If we decide to open just box B, it is highly probable that the being has put \$1,000,000 in box B. But then our decision to open just box B makes it more desirable to actually choose to open both boxes.

A famous example: Newcomb's paradox

- If we decide to open just box B, it is highly probable that the being has put \$1,000,000 in box B. But then our decision to open just box B makes it more desirable to actually choose to open both boxes.
    - Opening just box B is **not** ratifiable.

A famous example: Newcomb's paradox

- If we decide to open just box B, it is highly probable that the being has put \$1,000,000 in box B. But then our decision to open just box B makes it more desirable to actually choose to open both boxes.
  - Opening just box B is **not** ratifiable.

- If we decide to open both boxes, it is highly probable that the being has put nothing in box B. Hence, our decision to open both boxes does make it more desirable to actually open both boxes.

A famous example: Newcomb's paradox

- If we decide to open just box B, it is highly probable that the being has put \$1,000,000 in box B. But then our decision to open just box B makes it more desirable to actually choose to open both boxes.
  - Opening just box B is **not** ratifiable.

- If we decide to open both boxes, it is highly probable that the being has put nothing in box B. Hence, our decision to open both boxes does make it more desirable to actually open both boxes.
  - Opening both boxes is ratifiable. (It is also the dominant strategy.)

A famous example: Newcomb's paradox

- If we decide to open just box B, it is highly probable that the being has put \$1,000,000 in box B. But then our decision to open just box B makes it more desirable to actually choose to open both boxes.
    - Opening just box B is **not** ratifiable.
- If we decide to open both boxes, it is highly probable that the being has put nothing in box B. Hence, our decision to open both boxes does make it more desirable to actually open both boxes.
    - Opening both boxes is ratifiable. (It is also the dominant strategy.)

For a different analysis, see:

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

Ratifiability formalized

Let $G = \langle \{1,2\}, S_1, S_2, u_1, u_2 \rangle$ be a normal form game. We define a model of G:

Let At be the following set of atomic propositions: for all $i \in \{1,2\}$ and $a \in S_i$,

- $dec_i(a)$ means "player i decides to play strategy a"
- $per_i(a)$ means "player i performs strategy a"

Ratifiability formalized

Let $G = \langle \{1,2\}, S_1, S_2, u_1, u_2 \rangle$ be a normal form game. We define a model of G:

Let At be the following set of atomic propositions: for all $i \in \{1,2\}$ and $a \in S_i$,

- $dec_i(a)$ means "player i decides to play strategy a"
- $per_i(a)$ means "player i performs strategy a"

The set of states W is the set of functions $w : At \to \{0,1\}$ such that, for all $i \in \{1,2\}$ and $a, b \in S_i$ with $a \neq b$,

$$w(dec_i(a)) = 1 \text{ iff } w(dec_i(b)) = 0 \qquad w(per_i(a)) = 1 \text{ iff } w(per_i(b)) = 0$$

Ratifiability formalized

Let $G = \langle \{1,2\}, S_1, S_2, u_1, u_2 \rangle$ be a normal form game. We define a model of G:

Let At be the following set of atomic propositions: for all $i \in \{1,2\}$ and $a \in S_i$,

- $dec_i(a)$ means "player i decides to play strategy a"
- $per_i(a)$ means "player i performs strategy a"

The set of states W is the set of functions $w : At \to \{0,1\}$ such that, for all $i \in \{1,2\}$ and $a, b \in S_i$ with $a \neq b$,

$$w(dec_i(a)) = 1 \text{ iff } w(dec_i(b)) = 0 \qquad w(per_i(a)) = 1 \text{ iff } w(per_i(b)) = 0$$

We can define $\sigma : W \to S_1 \times S_2$ by setting, for all $(a,b) \in S_1 \times S_2$:

$$\sigma(w) = (a,b) \text{ iff } per_1(a) = 1 \text{ and } per_2(b) = 1$$

Ratifiability formalized

Let us now define, for each $i \in \{1, 2\}$ and $a \in S_i$:

$$\delta^i_a = \{w \in W \mid w(dec_i(a)) = 1\}$$

The event that i decides to play a

$$\pi^i_a = \{w \in W \mid w(per_i(a)) = 1\}$$

The event that i performs a

Ratifiability formalized

Let us now define, for each $i \in \{1, 2\}$ and $a \in S_i$:

$$\delta^i_a = \{w \in W \mid w(dec_i(a)) = 1\}$$

The event that i decides to play a

$$\pi^i_a = \{w \in W \mid w(per_i(a)) = 1\}$$

The event that i performs a

By construction, these events do not coincide, and we leave open as a logical possibility the divergence between decisions and performances. (p. 23)

Ratifiability formalized

Let us now define, for each $i \in \{1, 2\}$ and $a \in S_i$:

$$\delta^i_a = \{w \in W \mid w(dec_i(a)) = 1\}$$

The event that i decides to play a

- We can define $I_i : W \to \wp(W)$ as:

$$v \in I_i(w) \text{ iff } w, v \in \delta^i_a.$$

$$\pi^i_a = \{w \in W \mid w(per_i(a)) = 1\}$$

The event that i performs a

By construction, these events do not coincide, and we leave open as a logical possibility the divergence between decisions and performances. (p. 23)

Ratifiability formalized

Finally, for each $i \in \{1, 2\}$, we have a prior probability $p_i \in \Delta(W)$.

Ratifiability formalized

Finally, for each $i \in \{1, 2\}$, we have a prior probability $p_i \in \Delta(W)$.

$EU_i(a \mid b)$ is the expected utility of player i when she performs a even if she has decided to play b.

$$EU_i(a \mid b) \;:=\; \sum_{t \in S_{-i}} p_i(\pi_t^{-i} \mid \pi_a^i \cap \delta_b^i) \cdot u_i(a, t)$$

- $\pi_t^{-i} = \{w \in W \mid w(per_{-i}(t)) = 1\}$ is the event that $-i$ performs t
- $\pi_a^i = \{w \in W \mid w(per_i(a)) = 1\}$ is the event that i performs a
- $\delta_b^i = \{w \in W \mid w(dec_i(a)) = 1\}$ is the event that i decides to play b

Remark. $EU_i(a \mid b)$ is only defined when $\pi_a^i \cap \delta_b^i$ is non-null under p_i.

Ratifiability formalized

Let $\epsilon > 0$ be a given very small number.

The probability p_i is modestly ϵ-ratifiable if it satisfies:

1. **Existence of small trembles**

   $p_i(\pi^i_a \mid \delta^i_b) = \epsilon$ for all $a, b \in S_i$ s.t. $a \neq b$ (when defined)

2. **Trembles are independent of the opponent's decisions**

   $p_i(\pi^i_a \mid \delta^i_b \cap \delta^{-i}_t) = p_i(\pi^i_a \mid \delta^i_b)$ for all $a, b \in S_i$ and $t \in S_{-i}$ (when defined).

3. **Stability of deliberation**

   $EU^i(b \mid b) \geqslant EU^i(a \mid b)$ for all $a, b \in S_i$ whenever defined.

4. **Modesty**

   $p_i(\pi^{-i}_t) = p(\delta^{-i}_t)$ for all $t \in S_{-i}$

Ratifiability formalized

The precise magnitude *of the trembles should play no part in the analysis. Rather, what matters is that such trembles exist, and that they be "small"... (p. 25)*

Ratifiability formalized

> *The precise* magnitude *of the trembles should play no part in the analysis. Rather, what matters is that such trembles exist, and that they be "small"... (p. 25)*

... p_i is modestly ratifiable if there are sequences $(p_1, p_2, \ldots, p_n, \ldots)$ and $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \ldots)$ such that, for all n,

1. p_n is modestly ϵ_n-ratifiable AND
2. $p_n \rightarrow p_i$ as $\epsilon_n \rightarrow 0$.

Main theorems

Theorem 1. p_i is modestly ratifiable iff p_i is a correlated equilibrium.

Main theorems

Theorem 1. p_i is modestly ratifiable iff p_i is a correlated equilibrium.

- Assuming that the players have common prior probabilities (i.e. $p_i = p_{-i}$), correlated equilibria can be viewed as the result of Bayesian rationality.

R. Aumann. *Correlated equilibrium as an expression of Bayesian rationality*. Econometrica 55, pp. 1-18, 1987.

Main theorems

Theorem 1. p_i is modestly ratifiable iff p_i is a correlated equilibrium.

- Assuming that the players have common prior probabilities (i.e. $p_i = p_{-i}$), correlated equilibria can be viewed as the result of Bayesian rationality.

R. Aumann. *Correlated equilibrium as an expression of Bayesian rationality*. Econometrica 55, pp. 1-18, 1987.

Theorem 2. Player i is counterfactually rational at w iff i is Bayes rational at w.

Main theorems

Theorem 1. p_i is modestly ratifiable iff p_i is a correlated equilibrium.

- Assuming that the players have common prior probabilities (i.e. $p_i = p_{-i}$), correlated equilibria can be viewed as the result of Bayesian rationality.

R. Aumann. *Correlated equilibrium as an expression of Bayesian rationality*. Econometrica 55, pp. 1-18, 1987.

Theorem 2. Player i is counterfactually rational at w iff i is Bayes rational at w.

No need of independence??

Shin’s notion of counterfactual rationality

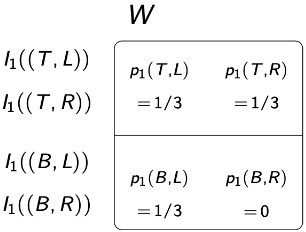
Another way to understand counterfactual rationality

> *A player should never find himself at a possible world at which ... her payoff would be higher if she were to deviate from the strategy she has chosen. This is the principle which motivates our rationality criterion. (p. 29)*

Shin H.S.. *A reconstruction of Jeffrey's notion of ratifiability in terms of counterfactual belief*. Theory and Decision 31, pp. 21-47, 1991.

Another way to define a model of a game

		Player 2	
		L	R
Player 1	T	6,6	2,7
	B	7,2	0,0



W

$I_1((T,L))$

$I_1((T,R))$

$I_1((B,L))$

$I_1((B,R))$

$p_1(T,L) = 1/3$

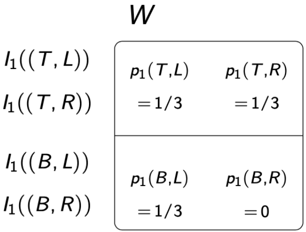
$p_1(T,R) = 1/3$

$p_1(B,L) = 1/3$

$p_1(B,R) = 0$

Another way to define a model of a game

		Player 2	
		L	R
Player 1	T	6,6	2,7
	B	7,2	0,0



W

$I_1((T,L))$

$I_1((T,R))$

$I_1((B,L))$

$I_1((B,R))$

$p_1(T,L) = 1/3$

$p_1(T,R) = 1/3$

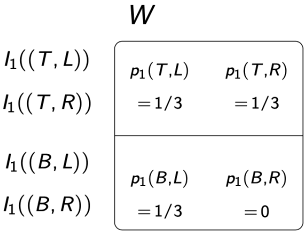
$p_1(B,L) = 1/3$

$p_1(B,R) = 0$

At world (T,L), player 1 believes that she is at a world where she plays T with probability 1 and player 2 plays L (R) with probability 0.5

Another way to define a model of a game

		Player 2	
		L	R
Player 1	T	6,6	2,7
	B	7,2	0,0

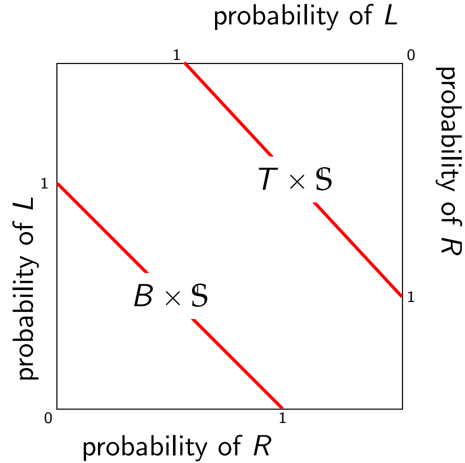


At world (T, L), player 1 believes that she is at a world where she plays T with probability 1 and player 2 plays L (R) with probability 0.5

- Define $\beta^1 : W \to S_1 \times \mathbb{S}$, where $\mathbb{S}$ is the one dimensional unit simplex representing the set of all probability distributions over $\{L, R\}$

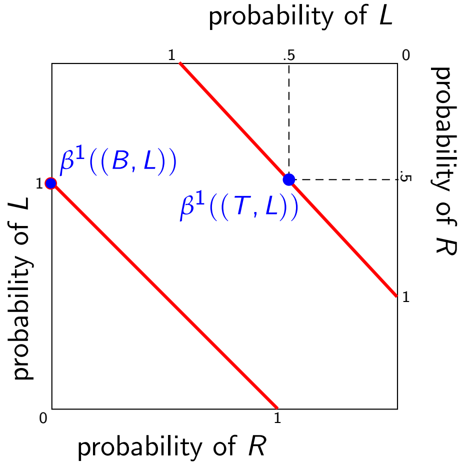
Belief space

The belief space of player 1 is $\{T, B\} \times \mathbb{S}$:



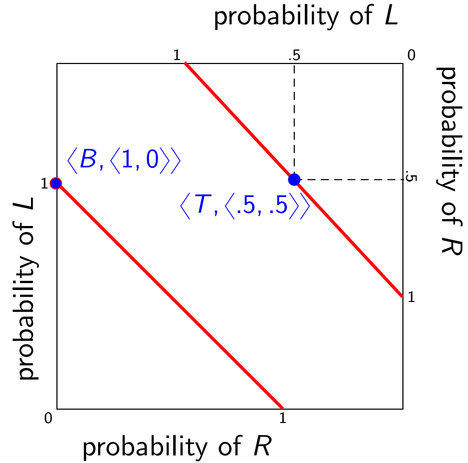
Belief space

The belief space of player 1 is $\{T, B\} \times \mathbb{S}$:



Belief space

The belief space of player 1 is $\{T, B\} \times \mathbb{S}$:


probability of L
1
.5
0
probability of R
$\langle B, \langle 1, 0 \rangle \rangle$
$\langle T, \langle .5, .5 \rangle \rangle$
.5
1
probability of L
1
0
1
probability of R

“Library stack metric”

We now define a distance measure, λ, to measure the distance (or closeness) between states in player 1’s belief space.

“Library stack metric”

We now define a distance measure, λ, to measure the distance (or closeness) between states in player 1’s belief space.

Let $\langle a, y\rangle$ and $\langle a', y'\rangle$ be two worlds in i’s belief space:

- $a, a' \in \{T, B\}$
- $y = \langle y_1, y_2\rangle \in \mathbb{R}^2$ with $y_1 + y_2 = 1$
- $y' = \langle y_1', y_2'\rangle \in \mathbb{R}^2$ with $y_1 + y_2 = 1$

“Library stack metric”

We now define a distance measure, λ, to measure the distance (or closeness) between states in player 1’s belief space.

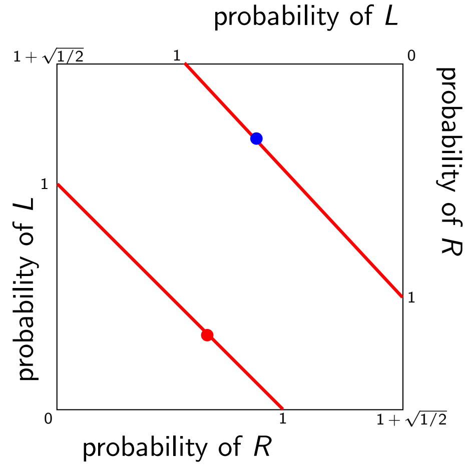
Let $\langle a, y\rangle$ and $\langle a', y'\rangle$ be two worlds in i’s belief space:

- $a, a' \in \{T, B\}$
- $y = \langle y_1, y_2\rangle \in \mathbb{R}^2$ with $y_1 + y_2 = 1$
- $y' = \langle y_1', y_2'\rangle \in \mathbb{R}^2$ with $y_1 + y_2 = 1$

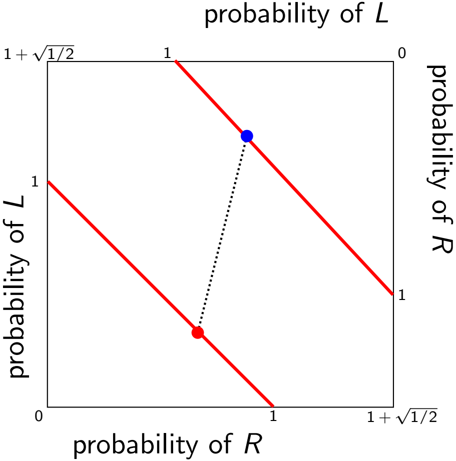
Then:

$$\lambda(\langle a, y\rangle, \langle a', y'\rangle) = \begin{cases} \sqrt{|y_1 - y_1'|^2 + |y_2 - y_2'|^2} & \text{if } a = a' \\ \sqrt{|y_1 - y_1'|^2 + |y_2 - y_2'|^2} + 1 & \text{if } a \neq a' \end{cases}$$

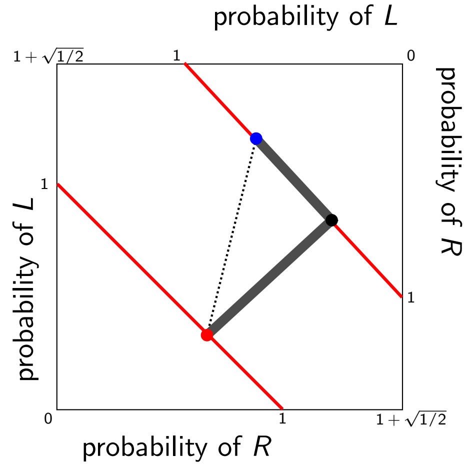
“Library stack metric”



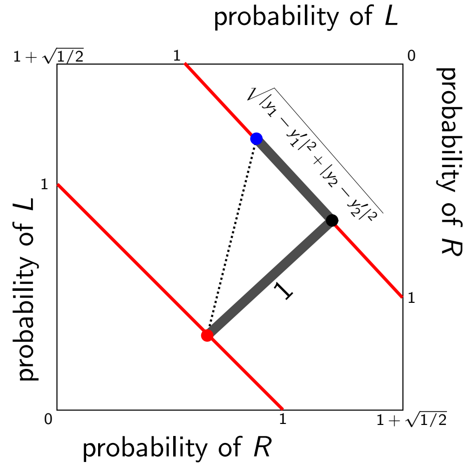
“Library stack metric”



“Library stack metric”

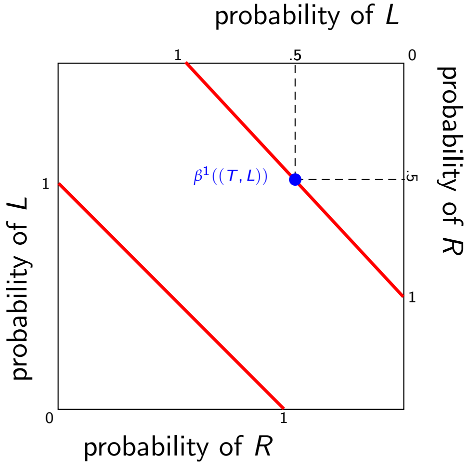


“Library stack metric”



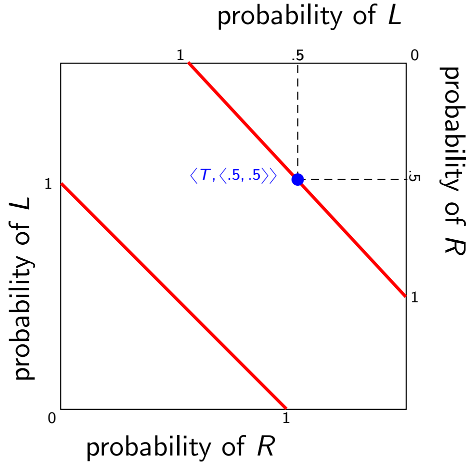
Counterfactual rationality

Player 1 is λ-rational at (T, L) if she believes that she is at a world at which, according to the metric λ, her payoff would not be higher if she were to play B.



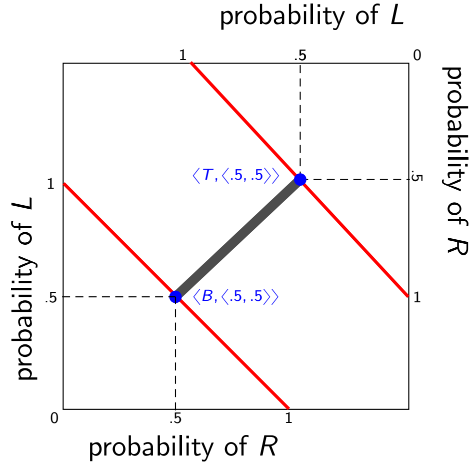
Counterfactual rationality

Player 1 is λ-rational at (T, L) if she believes that she is at a world at which, according to the metric λ, her payoff would not be higher if she were to play B.



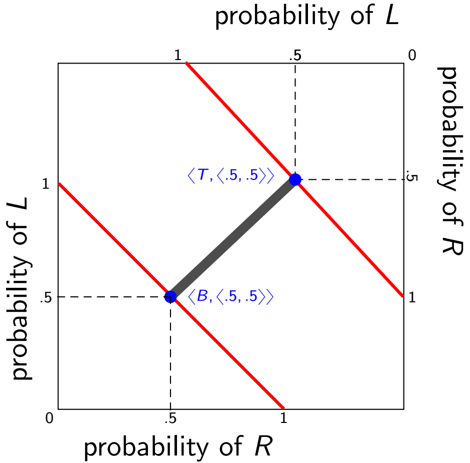
Counterfactual rationality

Player 1 is λ-rational at (T, L) if she believes that she is at a world at which, according to the metric λ, her payoff would not be higher if she were to play B.



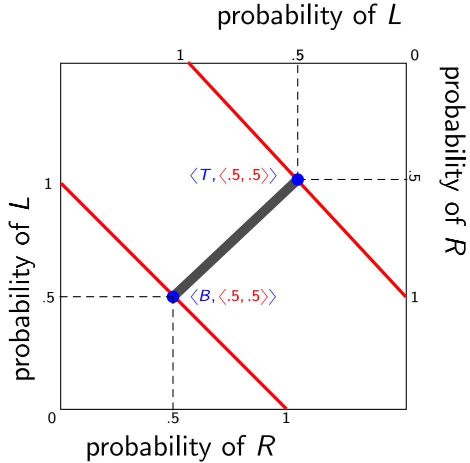
Counterfactual rationality

Player 1 is λ-rational at (T, L) if she believes that she is at a world at which, according to the metric λ, her payoff would not be higher if she were to play B. (Expected payoff at $\langle T, \langle .5, .5\rangle\rangle = 4$; expected payoff at $\langle B, \langle .5, .5\rangle\rangle = 3.5$.)



Counterfactual rationality

Player 1 is λ-rational at (T, L) if she believes that she is at a world at which, according to the metric λ, her payoff would not be higher if she were to play B. (Expected payoff at $\langle T, \langle .5, .5\rangle\rangle = 4$; expected payoff at $\langle B, \langle .5, .5\rangle\rangle = 3.5$.)



Take home messages

- The notion of Bayesian rationality is based on the (hidden) assumption that the players' choices are independent of one another—and that there is common belief that this is the case.

Take home messages

- The notion of Bayesian rationality is based on the (hidden) assumption that the players' choices are independent of one another—and that there is common belief that this is the case.
- If we allow (beliefs in) dependencies between the players' choices, then we can distinguish two notions of rationality: Bayesian rationality and counterfactual rationality.

  Keep in mind: some relations of relative closeness (like those defined by Shin) build in the assumption of independence of choices.

Take home messages

- The notion of Bayesian rationality is based on the (hidden) assumption that the players' choices are independent of one another—and that there is common belief that this is the case.
- If we allow (beliefs in) dependencies between the players' choices, then we can distinguish two notions of rationality: Bayesian rationality and counterfactual rationality.

  Keep in mind: some relations of relative closeness (like those defined by Shin) build in the assumption of independence of choices.
- Besides (common belief in) independence of choices, the notion of Bayesian rationality encodes the idea that the players' choices are rational when they are ratifiable (i.e., stable or non self-defeating).

A key question

Is it plausible to assume that the principle of independence of choices characterizes rational beliefs?

A key question

Is it plausible to assume that the principle of independence of choices characterizes rational beliefs?

Answer 1: YES
especially when we consider games in normal form, where the players are typically assumed to move simultaneously and be ignorant of each other's strategies.

A key question

> *[A] causal independence assumption is part of the idealization built into the normal form.*

W.L. Harper. *Causal decision theory and game theory: A classic argument for equilibrium solutions, a defense of weak equilibria, and a new problem for the normal form representation*. Causation in Decision, Belief Change and Statistics II, 1988.

> *[I]n a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players.*

R. Stalnaker. *Knowledge, belief and counterfactual reasoning in games*. Economics and Philosophy 12, pp. 133-163, 1996.

A key question

Is it plausible to assume that the principle of independence of choices characterizes rational beliefs?

Answer 1: YES
especially when we consider games in normal form, where the players are typically assumed to move simultaneously and be ignorant of each other's strategies.

Answer 2: NO
if we do not exclude that the players can communicate or be "translucent" to one another or when we consider games where the players move sequentially.

Plan

Tomorrow:

S.J. Brams. *Newcomb's problem and the Prisoner's Dilemma*. Journal of Conflict Resolution 19(4), pp. 596-612, 1975.

J. Halpern and R. Pass. *Game theory with translucent players*. Int J Game Theory 47, pp. 949-976, 2018.

Wednesday:

S.M. Hutteger & G.J. Rothfus. *Bradley conditionals and dynamic choice*. Synthese 199, pp. 6585-99, 2021.

J. Halpern. *Substantive rationality and backward induction*. Games and Economic Behavior 37, pp. 425-435, 2001.