Royal Netherlands Academy of Arts and Sciences (KNAW)

Master Class

Amsterdam, February 8th, 2007

Epistemic Foundations of Game Theory

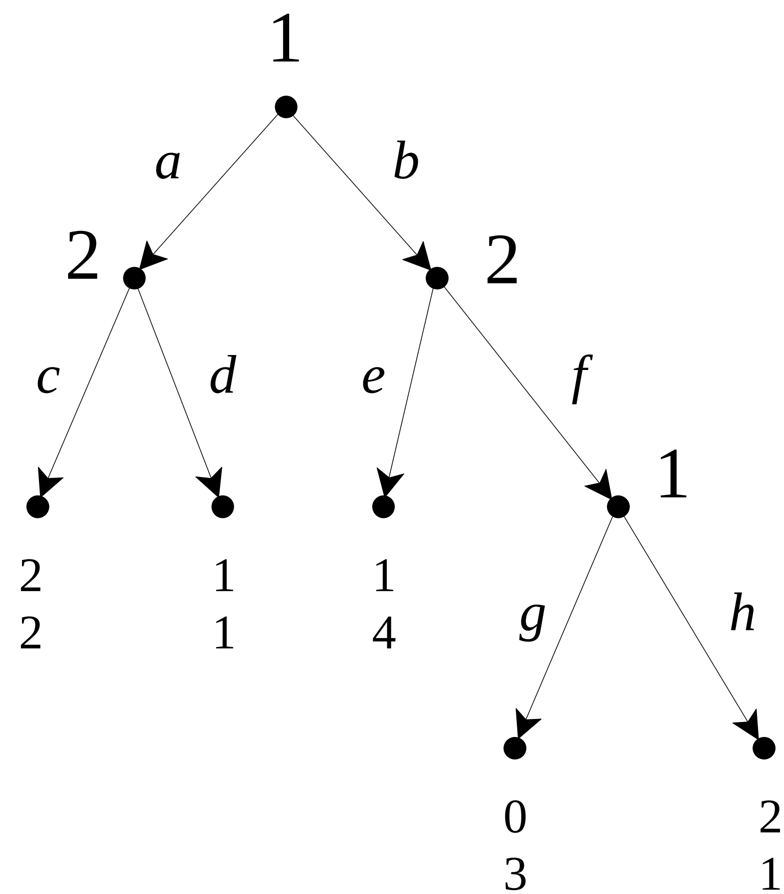
Lecture 2

Giacomo Bonanno

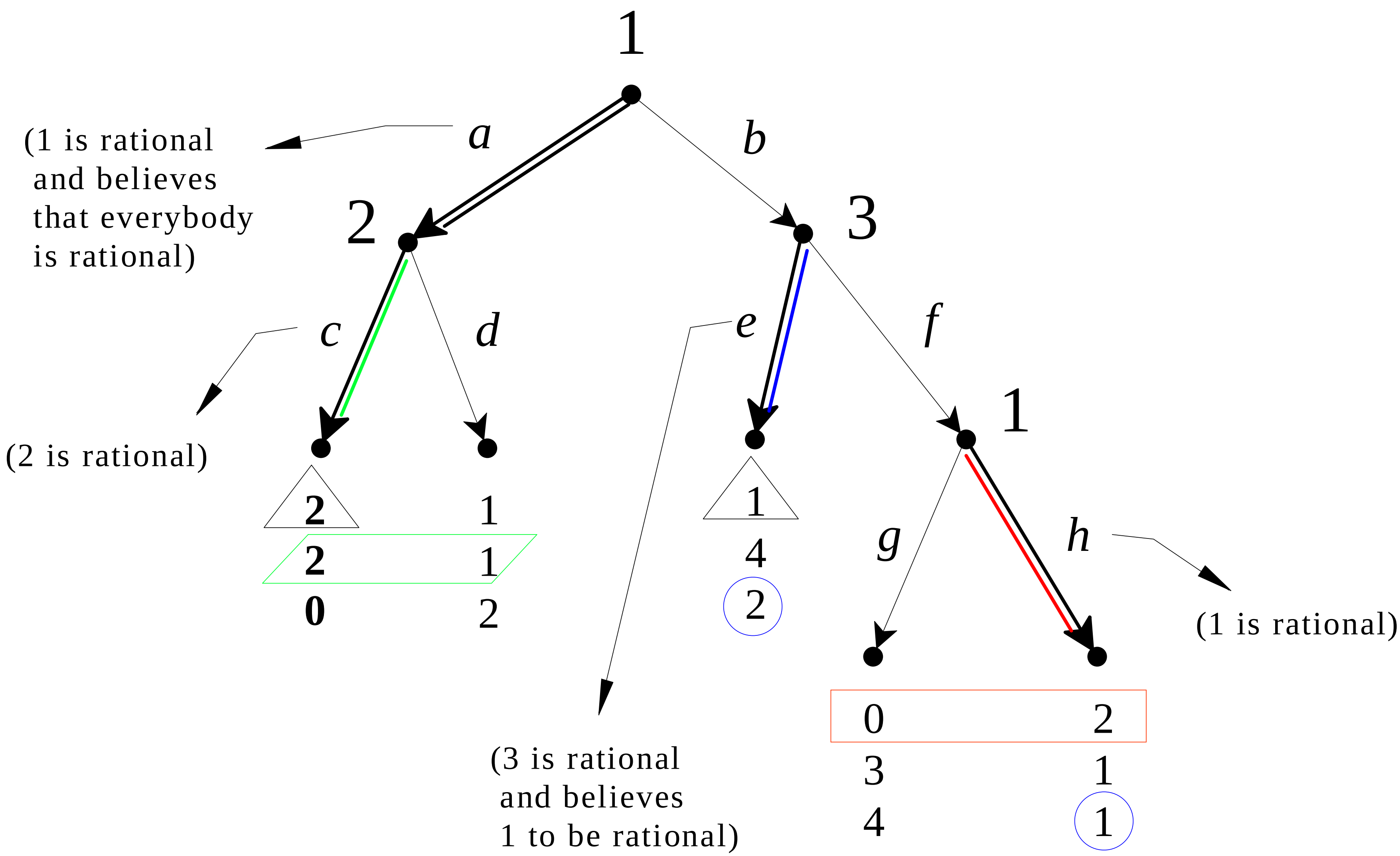
(http://www.econ.ucdavis.edu/faculty/bonanno/)

EXTENSIVE GAMES WITH PERFECT INFORMATION

- tree
- *n* players
- assignment of one player to every non-terminal node
- assignment of an ***ordinal*** payoff to every player at every terminal node



BACKWARD-INDUCTION SOLUTION



STRATEGIES IN PERFECT-INFORMATION GAMES

Non-terminal nodes are called *decision nodes*

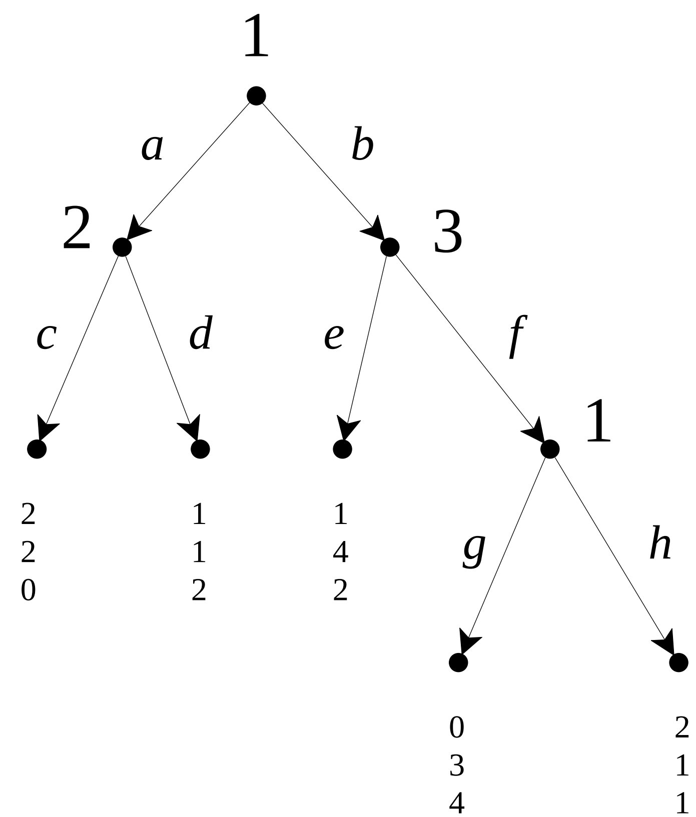
X : set of decision nodes

X_i : set of decision nodes assigned to player i

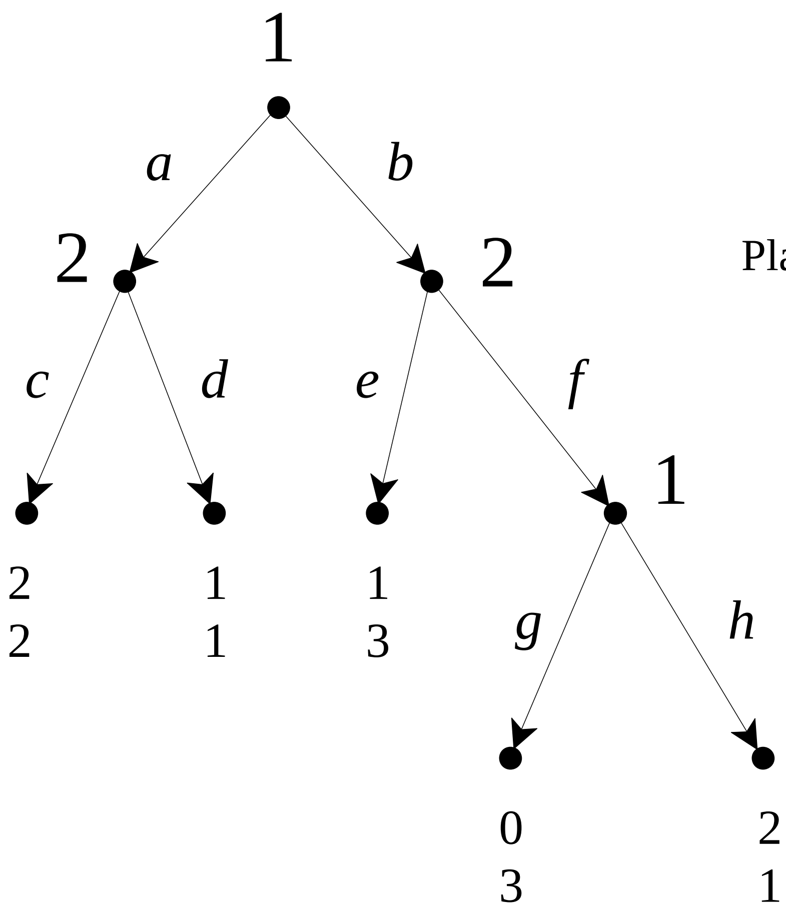
Definition. A strategy of player i is a function that assigns to every $x \in X_i$ a choice at x

Player 1's strategies:

(a,g), (a,h), (b,g) and (b,h)



THE STRATEGIC FORM OF A PERFECT-INFORMATION GAME



Player 2

Player 1	*ce*	*cf*	*de*	*df*
ag	2 , 2	2 , 2	1 , 1	1 , 1
ah	2 , 2	2 , 2	1 , 1	1 , 1
bg	1 , 3	0 , 3	1 , 3	0 , 3
bh	1 , 3	2 , 1	1 , 3	2 , 1

EPISTEMIC MODEL OF A PERFECT-INFORMATION GAME

(Knowledge based)

- Set of states Ω
- Equivalence relation $\mathcal{K}_i$ on Ω for every player *i*
- For every player *i* a function $\sigma_i : \Omega \to S_i$ satisfying

$$\text{if } \omega' \in \mathcal{K}_i(\omega) \text{ then } \sigma_i(\omega') = \sigma_i(\omega)$$

Thus a standard epistemic model for the associated strategic form

Recall from Lecture 1:

Let s_i and t_i be two strategies of player i: $s_i, t_i \in S_i$

$s_i \succ_i t_i$ is interpreted as "strategy s_i is better for player i than strategy t_i"

$s_i \succ_i t_i$ is true at state ω if $u_i(s_i, \sigma_{-i}(\omega)) > u_i(t_i, \sigma_{-i}(\omega))$

that is, s_i is better than t_i against $\sigma_{-i}(\omega)$

profile of strategies chosen by the players other than i

Let $\|s_i \succ_i t_i\| = \{\omega \in \Omega : u_i(s_i, \sigma_{-i}(\omega)) > u_i(t_i, \sigma_{-i}(\omega))\}$ event that s_i is better than t_i

If $s_i \in S_i$, let $\|s_i\| = \{\omega \in \Omega : \sigma_i(\omega) = s_i\}$ event that player i chooses s_i

Let $\boldsymbol{R_i}^{EA}$ be the event representing the proposition "player *i* is *ex ante* rational"

$$\|s_i\| \cap K_i \|t_i \succ_i s_i\| \subseteq \neg \boldsymbol{R_i^{EA}}$$

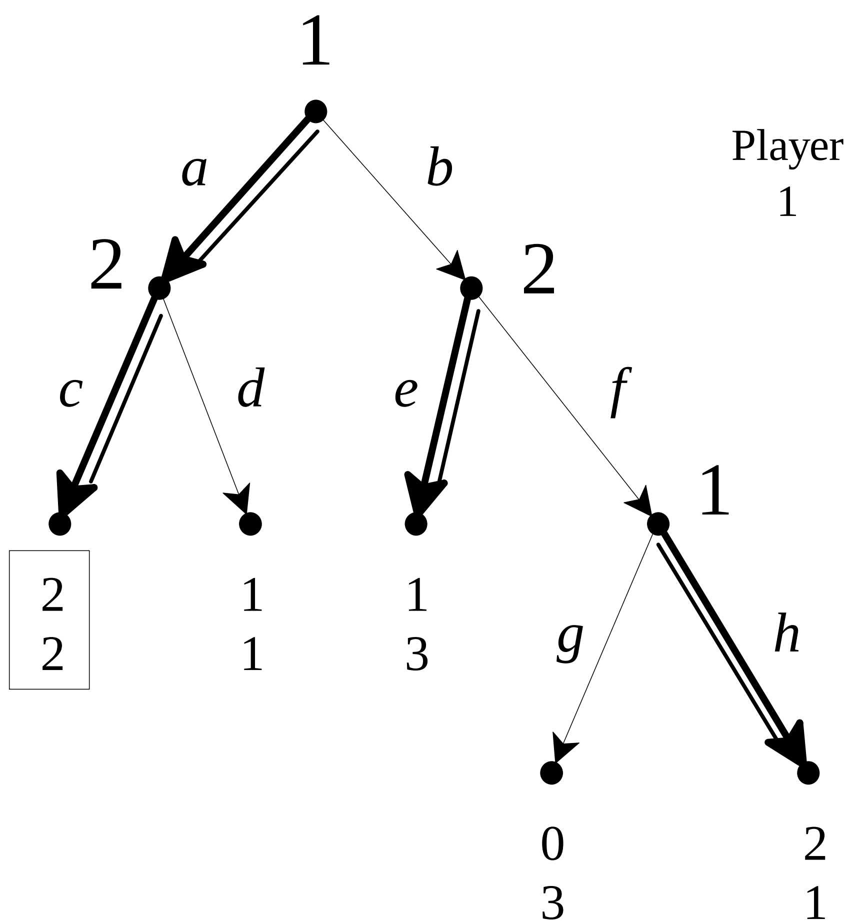
$$\neg \boldsymbol{R_i^{EA}} = \bigcup_{s_i \in S_i} \bigcup_{t_i \in S_i} \left(\|s_i\| \cap K_i \|t_i \succ_i s_i\| \right)$$

$$\boldsymbol{R^{EA}} = \boldsymbol{R_1^{EA}} \cap \ldots \cap \boldsymbol{R_n^{EA}}$$ all players are rational

Recall from Lecture 1:

PROPOSITION: if at a state there is common knowledge of *ex ante* rationality then the strategy profile chosen at that state belongs to the game obtained by applying the iterated deletion of strictly dominated strategies; conversely, for every such strategy profile there is a model and a state where (1) the strategy profile is chosen and (2) there is common knowledge of *ex ante* rationality.

This notion of rationality is not sufficient to yield backward induction



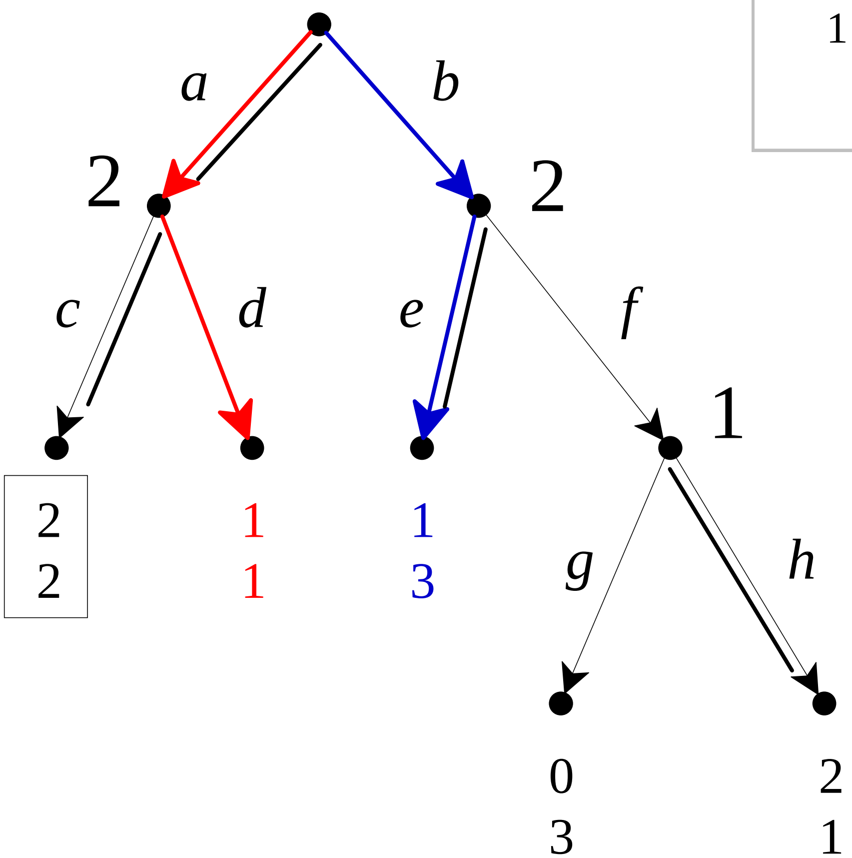
Player 2

Player 1	*ce*	*cf*	*de*	*df*
ag	2 , 2	2 , 2	1 , 1	1 , 1
ah	2 , 2	2 , 2	1 , 1	1 , 1
bg	1 , 3	0 , 3	1 , 3	0 , 3
bh	1 , 3	2 , 1	1 , 3	2 , 1

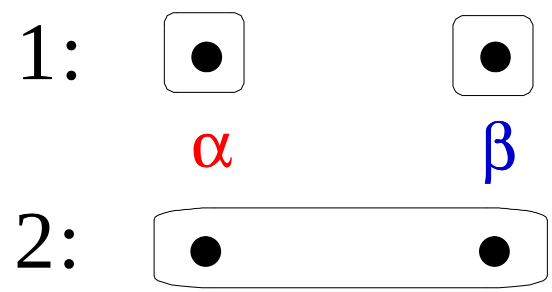
Here there are no strictly dominated strategies

Thus every strategy profile is consistent with common belief/knowledge of *ex ante* rationality

For example:



		Player 2			
		ce	*cf*	*de*	*df*
Player 1	*ag*	2 , 2	2 , 2	1 , 1	1 , 1
	ah	2 , 2	2 , 2	1 , 1	1 , 1
	bg	1 , 3	0 , 3	1 , 3	0 , 3
	bh	1 , 3	2 , 1	1 , 3	2 , 1



1's strategy: ***ah*** ***bh***

2's strategy: ***de*** ***de***

(For 2 *ce* better than *de* at α but not at β, thus at α she does not know that *ce* is better.)

Here: *ex ante* rationality and common knowledge of *ex ante* rationality at both states.

Let $\boldsymbol{R_i}^{\text{EA/S}}$ be the event representing the proposition "player *i* is *ex ante* rational in a strong sense"

$$\|s_i\| \cap K_i \|t_i \succsim_i s_i\| \cap \neg K_i \neg \|t_i \succ_i s_i\| \subseteq \neg \boldsymbol{R_i^{EA/S}}$$

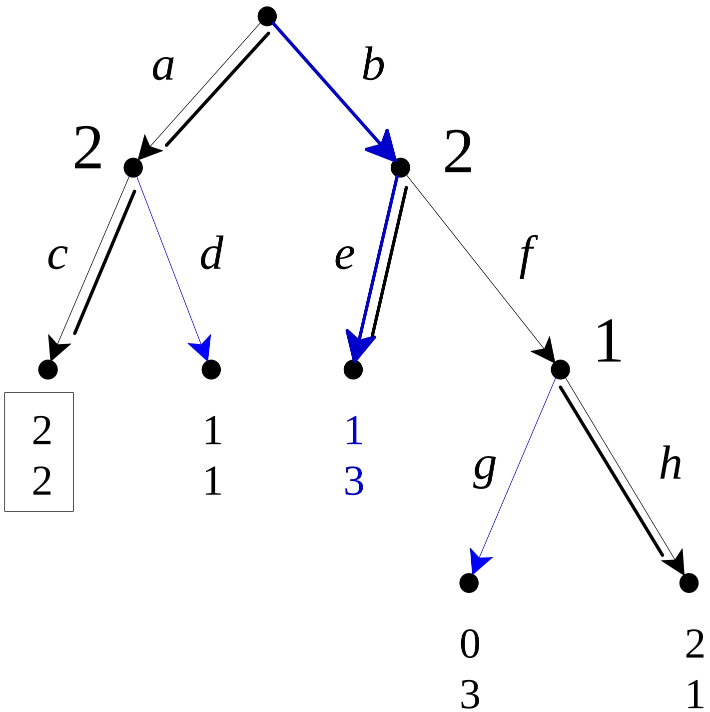
$$\neg \boldsymbol{R_i^{EA/S}} = \bigcup_{s_i \in S_i} \bigcup_{t_i \in S_i} \left(\|s_i\| \cap K_i \|t_i \succsim_i s_i\| \cap \neg K_i \neg \|t_i \succ_i s_i\| \right)$$

$$\boldsymbol{R^{EA/S}} = \boldsymbol{R_1^{EA/S}} \cap \ldots \cap \boldsymbol{R_n^{EA/S}}$$

all players are rational in a strong sense

Recall from Lecture 1:

PROPOSITION: if at a state there is common knowledge of *ex ante* rationality in a strong sense then the strategy profile chosen at that state belongs to the set T^{∞} of strategy profiles that survive the iterated deletion of inferior profiles; conversely, for every such strategy profile there is a model and a state where (1) the strategy profile is chosen and (2) there is common knowledge of *ex ante* rationality in a strong sense.



	ce	*cf*	*de*	*df*
ag	2 , 2	2 , 2	1 , 1	1 , 1
ah	2 , 2	2 , 2	1 , 1	1 , 1
bg	1 , 3	0 , 3	1 , 3	0 , 3
bh	1 , 3	2 , 1	1 , 3	2 , 1

player 1 using *ah*

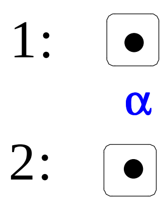
player 2 using *ce*

	ce	*cf*	*de*
ag	2 , 2	2 , 2	1 , 1
ah	2 , 2	2 , 2	1 , 1
bg			1 , 3
bh	1 , 3		1 , 3

player 2 using *cf*

player 1 using *ah*

	ce	*cf*	*de*
ag	2 , 2	2 , 2	
ah	2 , 2	2 , 2	
bg			1 , 3
bh			1 , 3



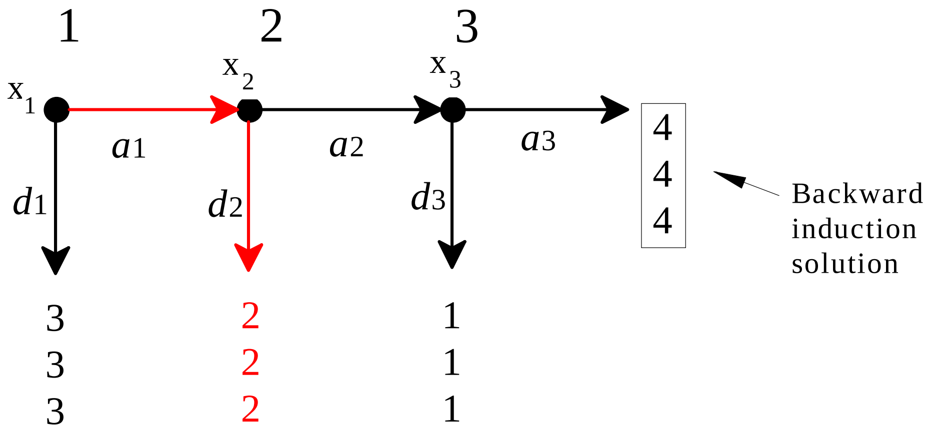
1's strategy: ***bg***

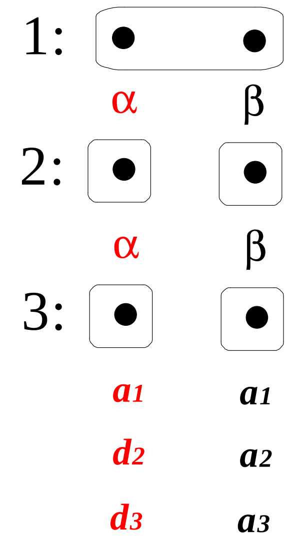
2's strategy: ***de***

Thus even common knowledge of *ex ante* rationality in a strong sense is not sufficient to yield backward induction

In this example all the strategy profiles in T^∞ are Nash equilibria. Is it the case that common knowledge of ex ante rationality in the strong sense gives Nash equilibrium **play** in perfect information games?

The answer is NO!





There is no Nash equilibrium that yields the play $a_1 d_2$ (the Nash equilibria are marked in blue)

$a3$:

	a2	d2
a1	**4,4,4**	2,2,2
d1	3,3,3	**3,3,3**

$d3$:

	a2	d2
a1	1,1,1	2,2,2
d1	**3,3,3**	**3,3,3**

$a3$:

	a2	d2
a1	4,4,4	2,2,2
d1	3,3,3	3,3,3

$d3$:

	a2	d2
a1		2,2,2
d1	3,3,3	3,3,3

First round: eliminate (a_1, a_2, d_3) through player 3 and a_3

second round: eliminate (a_1, d_2, a_3) through player 2 and a_2

$a3$:

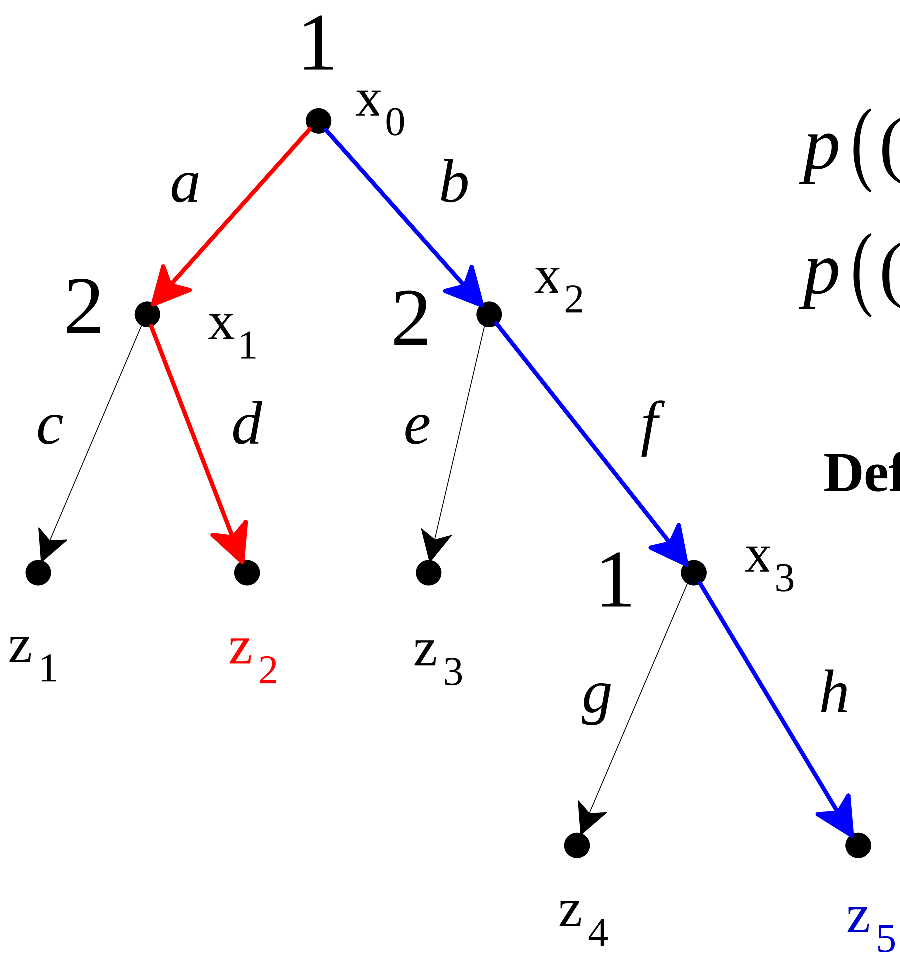
	a2	d2
a1	4,4,4	
d1	3,3,3	3,3,3

$d3$:

	a2	d2
a1		2,2,2
d1	3,3,3	3,3,3

Going beyond *ex ante* rationality

Given a strategy profile s, let $p(s)$ be the associated play



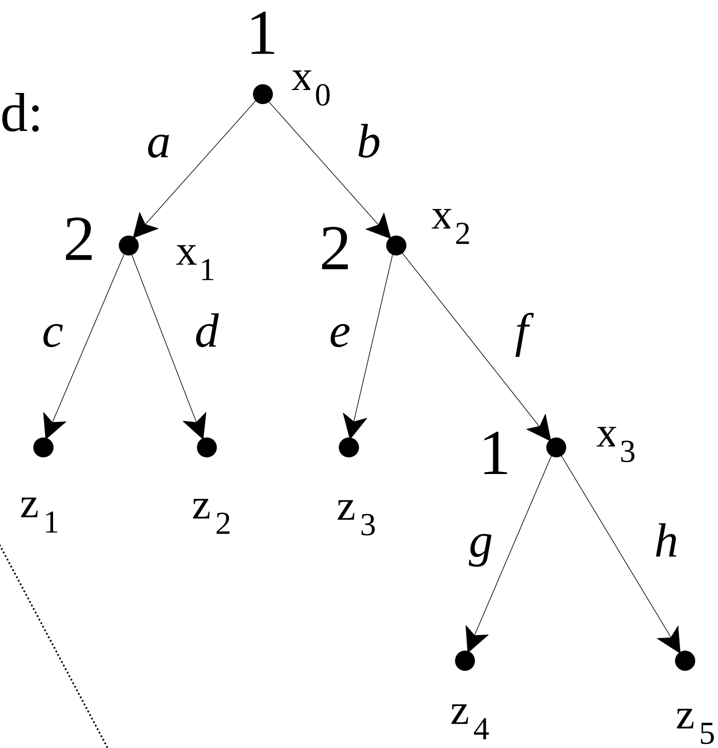
$$p\big((ag, df)\big) = x_0 x_1 z_2$$

$$p\big((bh, df)\big) = x_0 x_2 x_3 z_5$$

Definition. At state ω node x is *reached* if and only if $x \in p\big(\sigma(\omega)\big)$.

Definition. Given an epistemic model, for every node x, let $\|x\|$ be the event that node x is reached:

$$\|x\| = \{\omega \in \Omega : x \in p(\sigma(\omega))\}$$



	α	β	γ	δ	ε
	●	●	●	●	●
1's strategy:	*ag*	*bh*	*bg*	*bh*	*bg*
2's strategy:	*df*	*df*	*ce*	*de*	*cf*

$$\|x_1\| = \{\alpha\}, \quad \|x_2\| = \{\beta, \gamma, \delta, \varepsilon\}$$

$$\|x_3\| = \{\beta, \varepsilon\}, \quad \|z_1\| = \varnothing, \quad \|z_2\| = \{\alpha\}, \quad \text{etc.}$$

Let $E, F \subseteq \Omega$ be two events.

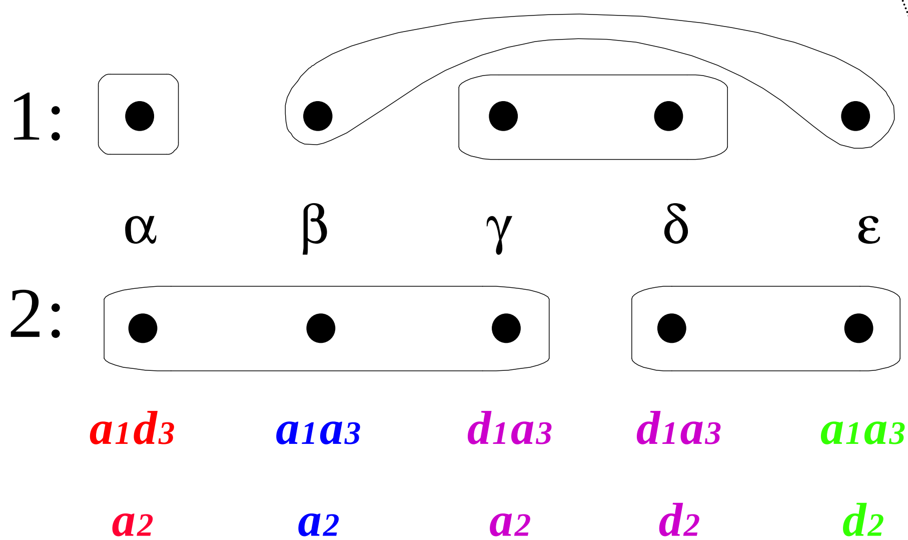
Denote by $E \to F$ the event $\neg E \cup F$ (if E then F)

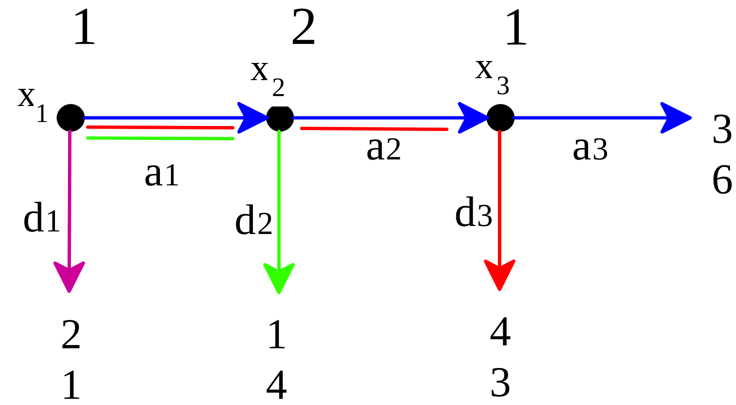
Let $\boldsymbol{R_i}^{\text{RN}}$ be the event representing the proposition "player i is rational *at reached nodes*"

if $x \in X_i$ $$\|x\| \cap \|s_i\| \cap K_i\left(\|x\| \to \|t_i \succ_i s_i\|\right) \subseteq \neg \boldsymbol{R_i^{RN}}$$

$$\neg \boldsymbol{R_i^{RN}} = \bigcup_{x \in X_i} \bigcup_{s_i \in S_i} \bigcup_{t_i \in S_i} \left(\|s_i\| \cap K_i\left(\|x\| \to \|t_i \succ_i s_i\|\right) \cap \|x\| \right)$$

$$\boldsymbol{R^{RN}} = \boldsymbol{R_1^{RN}} \cap \ldots \cap \boldsymbol{R_n^{RN}}$$ all players are rational at reached nodes





$\|d_2 \succ_2 a_2\| = \{\alpha\}$ $\|x_2\| = \{\alpha, \beta, \varepsilon\}$ $\neg\|x_2\| \cup \|d_2 \succ_2 a_2\| = \{\alpha, \gamma, \delta\}$

$K_2\left(\|x_2\| \rightarrow \|d_2 \succ_2 a_2\|\right) = \varnothing$ Thus player 2 is rational at nodes α and β and trivially at γ.

$\|a_2 \succ_2 d_2\| = \{\beta, \varepsilon\}$ $\|x_2\| = \{\alpha, \beta, \varepsilon\}$ $\neg\|x_2\| \cup \|a_2 \succ_2 d_2\| = \{\beta, \gamma, \delta, \varepsilon\}$

$K_2\left(\|x_2\| \rightarrow \|a_2 \succ_2 d_2\|\right) = \{\delta, \varepsilon\}$ $\|x_2\| \cap \|d_2\| \cap K_2\left(\|x_2\| \rightarrow \|a_2 \succ_2 d_2\|\right) = \{\varepsilon\}$

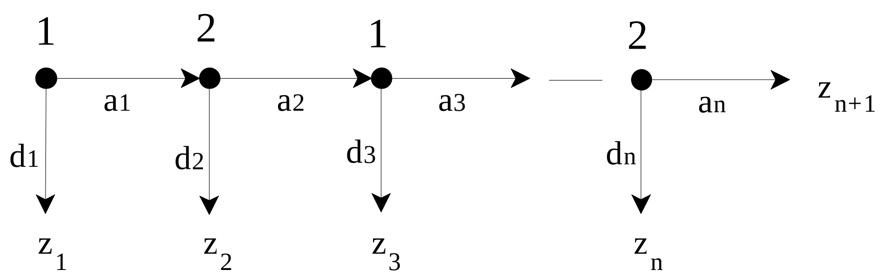
Thus **player 2 is** trivially rational at state δ, and **irrational at ε**. $K_* \boldsymbol{R} = \varnothing$

Backward Induction terminating games

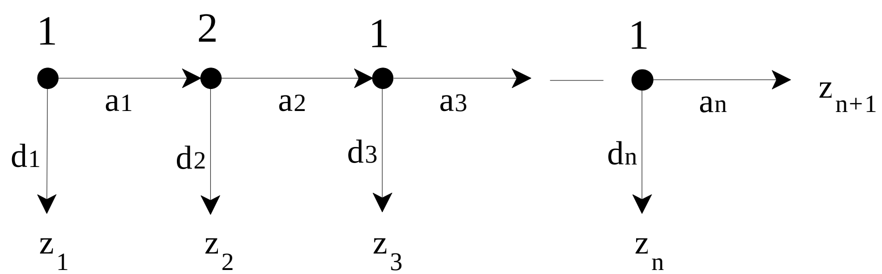
Definition. *A BI terminating game* is a perfect information game where (1) at each decision node there is a choice the terminates the game (it leads to a terminal node) and (2) the backward-induction solution prescribes a terminating choice at every decision node.

The best-known example is the **centipede game** (n is the number of decision nodes)

if n is even



if n is odd



$$u_1(z_1) = 2$$
$$u_2(z_1) = 1$$

for $1 < k \le n$

$$u_1(z_k) = u_2(z_{k-1})$$
$$u_2(z_k) = u_1(z_{k-1}) + 2$$

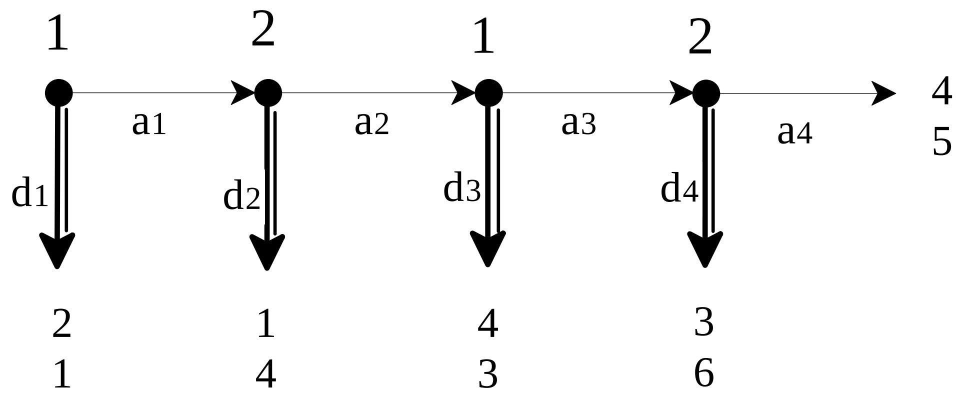
If n is even

$$u_1(z_{n+1}) = u_1(z_n) + 1$$
$$u_2(z_{n+1}) = u_2(z_n) - 1$$

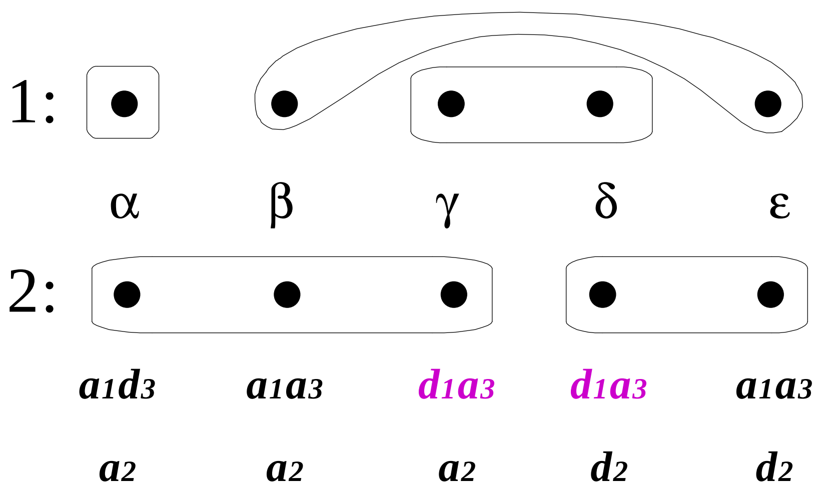
If n is odd

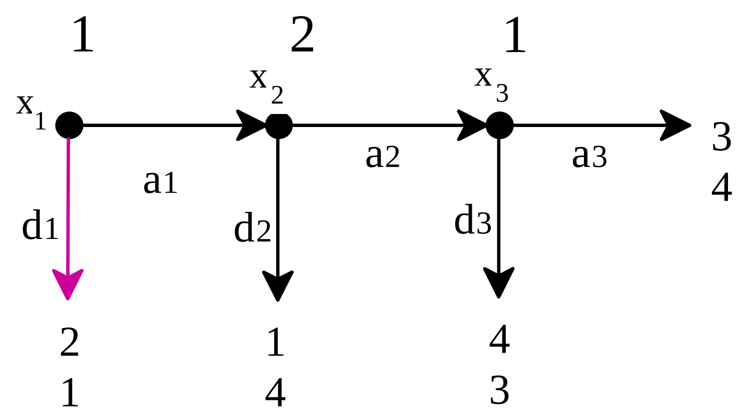
$$u_1(z_{n+1}) = u_1(z_n) - 1$$
$$u_2(z_{n+1}) = u_2(z_n) + 1$$

$n = 4$



Definition. Given an epistemic model of a **BI** terminating game, let BI be the event that the backward-induction **play** obtains, that is, $\boldsymbol{BI} = \{\omega \in \Omega : p(\sigma(\omega)) = x_1 z_1\}$





BI = {γ,δ}

PROPOSITION 1. In every BI terminating game, $K_* \boldsymbol{R}^{RN} \subseteq \boldsymbol{BI}$

PROPOSITION 2. For every BI terminating game, there is a model of it where $K_* \boldsymbol{R}^{RN} \neq \varnothing$

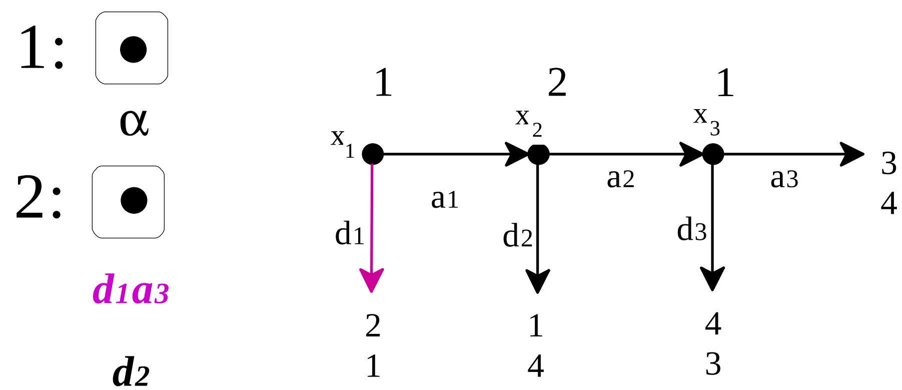
Aumann, R., A note on the centipede game, *Games and Economic Behavior*, 1998, 23: 97-105.

Broome, J. and W. Rabinowicz, Bacwards induction in the centipede game, *Analysis*, 1999, 59:237-242.

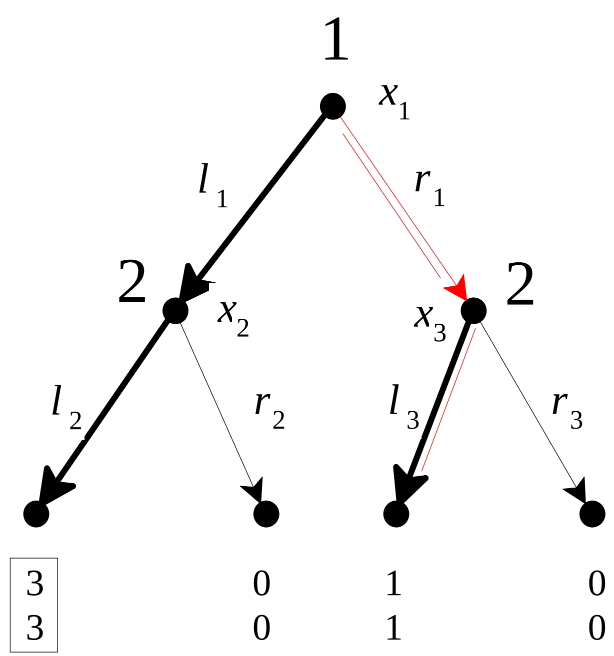
Rabinowicz, W., Grappling with the centipede, *Economics and Philosophy*, 1998, 14: 95-126.

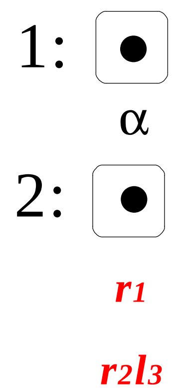
Sugden, R., Rational choice: a survey of contributions from economics and philosophy, *Economic Journal*, 1991, 101:751-785.

Note: it is not necessarily the case that if $\omega \in \Omega$ is such that at ω there is common knowledge of rationality then $\sigma(\omega)$ coincides with the backward-induction **strategy profile**. What is true is that player 1's strategy assigns the terminating choice to the root.



In general perfect-information games common knowledge of Rationality at Reached Nodes does **not** yield the backward-induction play.

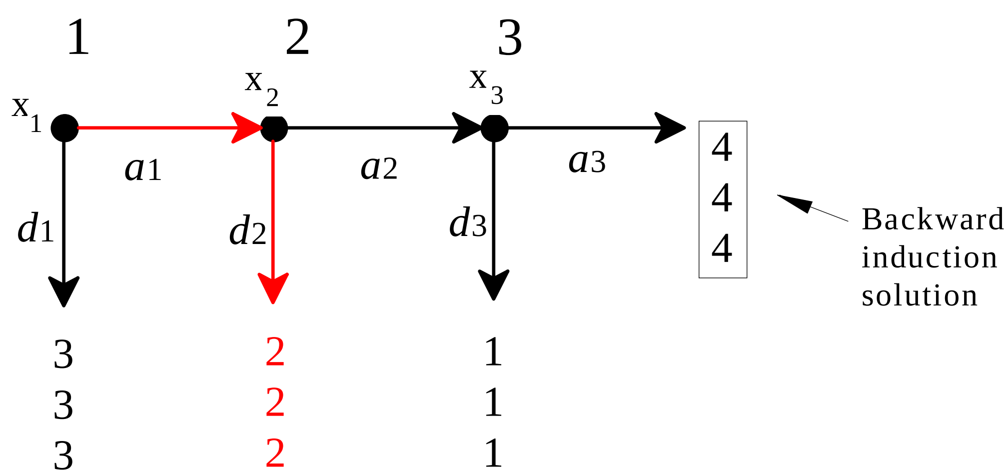


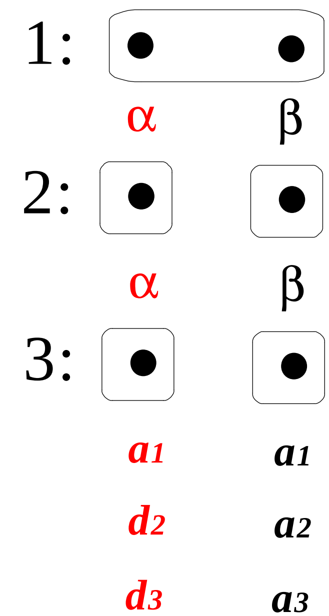


The backward induction play is $l_1 l_2$ while in this model we get $r_1 l_3$

(r_1,$r_2 l_3$) is a Nash equilibrium. Does common knowledge of Rationality at Reached Nodes at least yield a play that can be sustained by a Nash equilibrium?

NO! In general, common knowledge of Rationality at Reached Nodes does not yield Nash equilibrium play





a3

	a2	*d2*
a1	**4,4,4**	2,2,2
d1	3,3,3	**3,3,3**

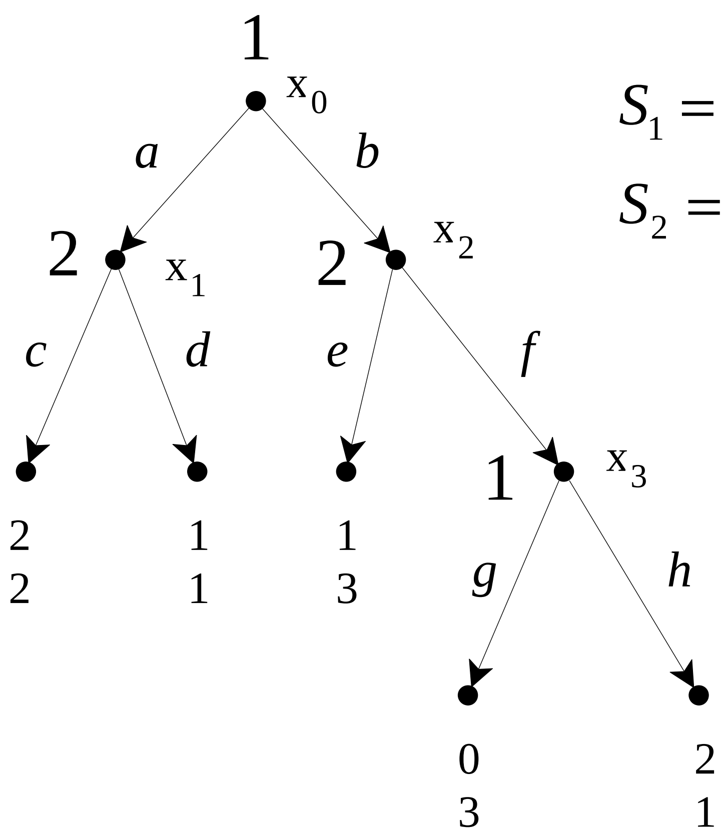
d3

	a2	*d2*
a1	1,1,1	2,2,2
d1	**3,3,3**	**3,3,3**

The Nash equilibria are marked in blue

Dealing with general perfect-information games

Let $x \in X_i$ be a decision node of player *i*. Denote by S_i^x the set of player *i*'s strategies in the subgame that starts at node *x*.



$$S_1 = \{ag, ah, bg, bh\},\ S_1^{x_3} = \{g, h\}$$

$$S_2 = \{cd, cf, de, df\},\ S_2^{x_1} = \{c, d\},\ S_2^{x_2} = \{e, f\}$$

Let x be a decision node of player i and let $s_i^x, t_i^x \in S_i^x$ be two strategies of player i in the subgame that starts at node x

$s_i^x \succ_i t_i^x$ is interpreted as "for player i, strategy s_i^x is better than strategy t_i^x in the subgame that starts at node x"

$s_i^x \succ_i t_i^x$ is true at state ω if, **starting from node *x***, s_i^x gives a higher payoff to player i than t_i^x against $\sigma_{-i}(\omega)$

Let $\left\| s_i^x \succ_i t_i^x \right\|$ be the event that $s_i^x \succ_i t_i^x$ is true.

If x is a node of player i, let $\sigma_i(\omega)|_x$ denote the restriction of $\sigma_i(\omega)$ to the subgame that starts at x

If $s_i^x \in S_i^x$, let $\left\| s_i^x \right\| = \left\{ \omega \in \Omega : s_i^x = \sigma_i(\omega)|_x \right\}$

SUSBSTANTIVE RATIONALITY (Aumann, GEB 1995)

Recall that if $E, F \subseteq \Omega$, $E \to F$ is the event $\neg E \cup F$ (if E then F)

Let $\boldsymbol{R_i}^{\mathrm{SR}}$ be the event representing the proposition "player *i* is substantively rational"

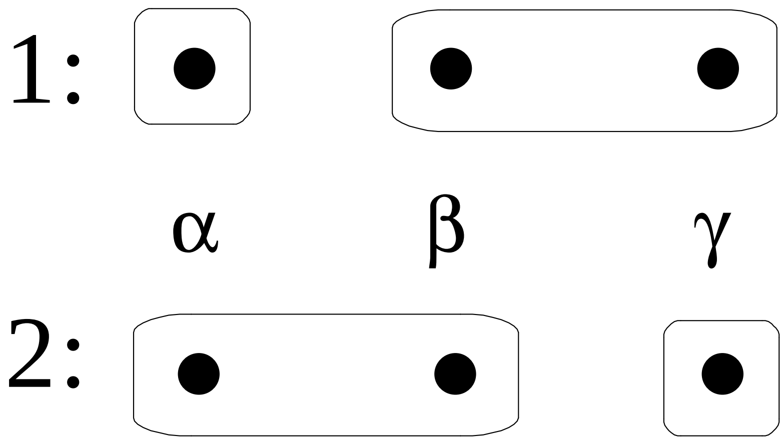
if $x \in X_i$

$$\left\| s_i^x \right\| \cap K_i\left(\left\| t_i^x \succ_i s_i^x \right\|\right) \subseteq \neg \boldsymbol{R_i^{SR}}$$

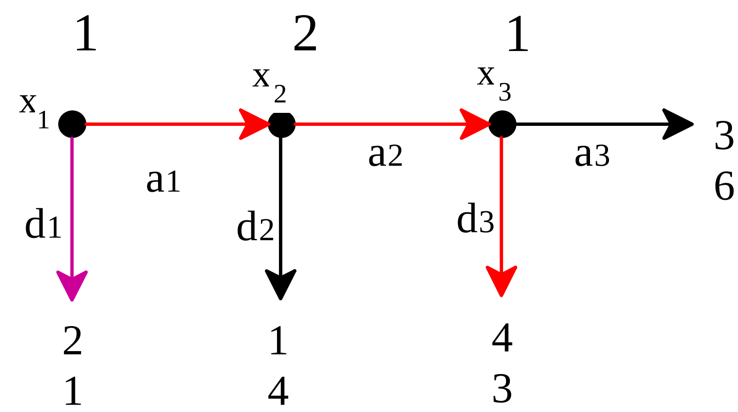
$$\neg \boldsymbol{R_i^{SR}} = \bigcup_{x \in X_i} \bigcup_{s_i \in S_i^x} \bigcup_{t_i \in S_i^x} \left(\left\| s_i^x \right\| \cap K_i\left(\left\| t_i^x \succ_i s_i^x \right\|\right)\right)$$

$$\boldsymbol{R^{SR}} = \boldsymbol{R_1^{SR}} \cap \ldots \cap \boldsymbol{R_n^{SR}}$$

all players are substantively rational



a_1d_3	d_1d_3	d_1d_3
a_2	a_2	d_2



$$\boldsymbol{R}_2^{EA} = \{\alpha, \beta, \gamma\} \text{ (\textit{ex ante} rationality)}$$

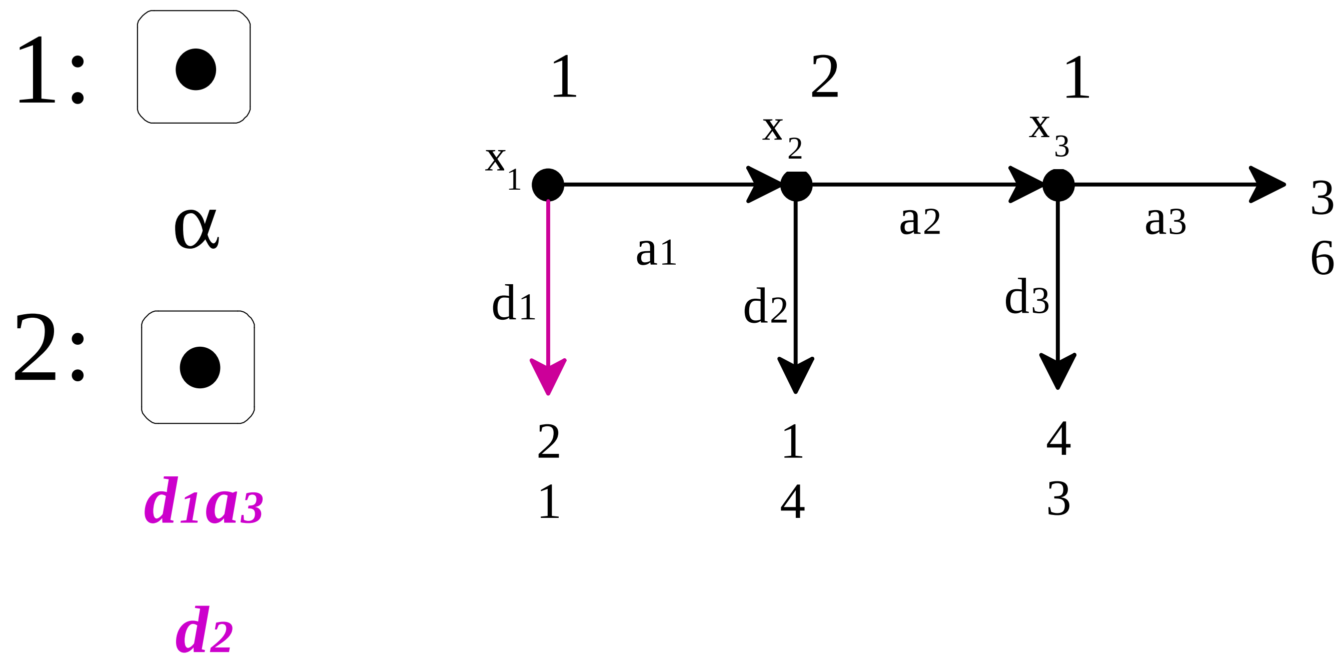
$$\boldsymbol{R}_2^{RN} = \{\beta, \gamma\} \text{ (rationality at reached nodes)}$$

$$\boldsymbol{R}_2^{SR} = \{\gamma\} \text{ (substantive rationality)}$$

PROPOSITION 3. In every perfect information game, $K_* \boldsymbol{R}^{\boldsymbol{SR}} \subseteq \boldsymbol{BI}$

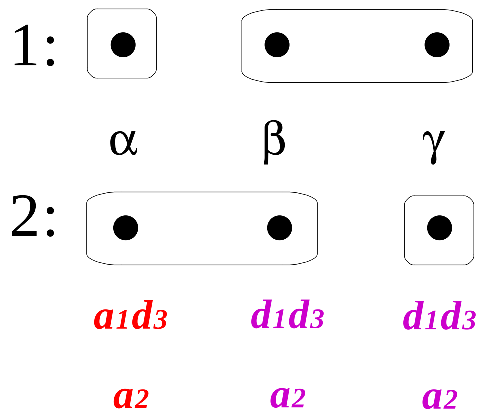
PROPOSITION 4. For every perfect information game, there is a model of it where $K_* \boldsymbol{R}^{\boldsymbol{SR}} \neq \varnothing$

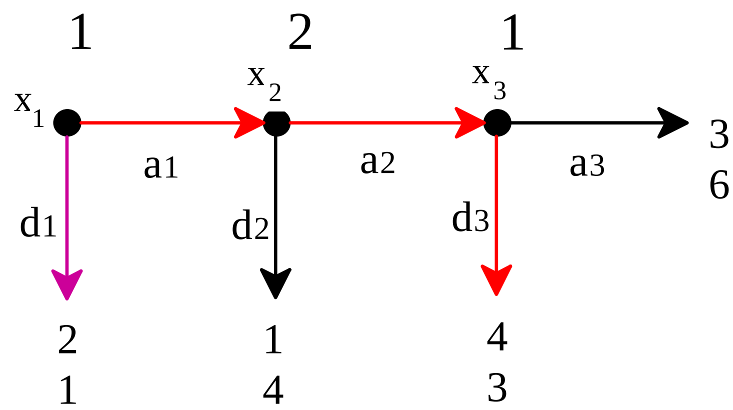
Aumann, R., Backward induction and common knowledge of rationality, *Games and Economic Behavior*, 1995, 8: 6-19.



Why is player 2 substantively irrational at state α? What is true at state α that makes player 2 substantively irrational?

At state α player 2 is not taking any actions, because her node x_2 is not reached. In fact, at state α player 2 ***knows*** that her node is not reached. So what makes her irrational (according to the notion of substantive rationality) must be her ***plan*** to choose d_2 ***if her decision node were to be reached***. This is a ***counterfactual*** statement.





The association of a strategy profile with every state gives rise to *two types of counterfactuals*:

(1) An objective statement about what the relevant player would do at a node that is not reached.

(2) (With the help of the partitions) a subjective statement about what a player believes would happen if he were to take a different action from the one he is actually taking.

(1) Thus at state γ it is true that **player 2** would take action a_2 if her node x_2 were to be reached (although it is not in fact reached and she knows that it is not reached)

(2) At states β and γ **player 1** knows that if he were to take action a_1 instead of d_1 at the root (he knows that he is taking d_1) then his payoff would be 4 (the payoff associated with $a_1a_2d_3$)

Modeling counterfactuals indirectly through strategies is not satisfactory. We have abandoned the modular approach suggested in Lecture 1, since there exists a module that deals with counterfactuals.

Modeling Counterfactuals

For every $\omega \in \Omega$, let $\mathcal{P}_\omega$ be a relation on Ω satisfying, $\forall \alpha, \beta \in \Omega$,

(1) either $\alpha \in \mathcal{P}_\omega(\beta)$ or $\beta \in \mathcal{P}_\omega(\alpha)$ (completeness)

(2) if $\beta \in \mathcal{P}_\omega(\alpha)$ then $\mathcal{P}_\omega(\beta) \subseteq \mathcal{P}_\omega(\alpha)$ (transitivity)

(3) if $\alpha \in \mathcal{P}_\omega(\beta)$ and $\beta \in \mathcal{P}_\omega(\alpha)$ then $\alpha = \beta$ (antisymmetry)

(4) $\omega' \in \mathcal{P}_\omega(\omega)$, for all $\omega' \in \Omega$ (centeredness)

The interpretation of $\beta \in \mathcal{P}_\omega(\alpha)$ or $\alpha \mathcal{P}_\omega \beta$ is that state α is at least as close to to state ω as state β is. Thus, for every state ω, the closeness relation $\mathcal{P}_\omega$ determines a strict ordering of the set of states based on closeness to ω, with ω itself being the closest state.

$\mathcal{P}_\omega(\alpha)$ = set of states that are not closer to ω than α is.

REPRESENTATION

$$\Omega = \{\alpha, \beta, \gamma, \delta\}$$

	α	β	γ	δ	
	●	●	●	●	
closest	α	β	γ	δ	ordering
	γ	α	δ	β	
	β	δ	α	γ	
farthest	δ	γ	β	α	

Given a state ω and an event E, denote by min(ω,E) the closest state to ω that belongs to event E. Thus if $\omega \in E$, then min(ω,E) = ω.

In the above example, if $E = \{\beta, \delta\}$ then min(α, E) = β

Recall that, if $E, F \subseteq \Omega$ are two events, $E \to F$ denotes the event $\neg E \cup F$ (if E then F). Thus $\omega \in E \to F$ if either $\omega \notin E$ or $\omega \in E \cap F$.

$\to$ represents the material conditional, which is true whenever the antecedent is false

We use the symbol $\hookrightarrow$ to denote the counterfactual conditional. Thus $E \hookrightarrow F$ is interpreted as "if E were the case then F would be the case"

Definition. $E \hookrightarrow F = \left\{ \omega \in \Omega : \min(\omega, E) \in F \right\}$

	α	β	γ	δ
	●	●	●	●
closest	α	β	γ	δ
	γ	α	δ	β
	β	δ	α	γ
farthest	δ	γ	β	α

If $E = \{\beta, \delta\}$ and $F = \{\alpha, \gamma, \delta\}$

then $E \hookrightarrow F = \{\gamma, \delta\}$

while $E \to F = \{\alpha, \gamma, \delta\}$

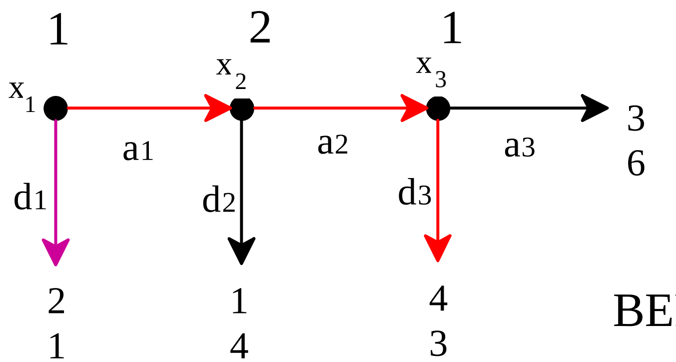
Note that, for all $E, F \subseteq \Omega$, $\quad E \hookrightarrow F \subseteq E \to F$

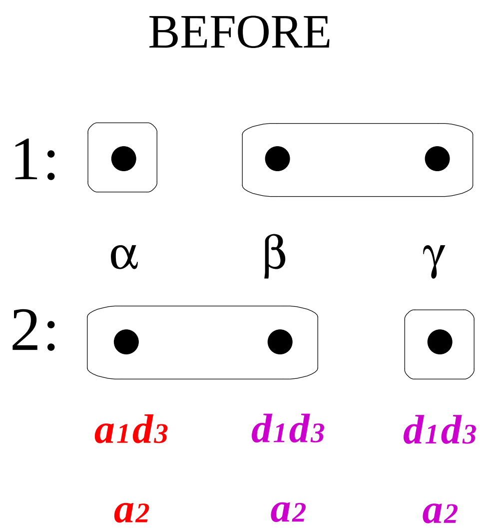
MODELING STRATEGIES WITH COUNTERFACTUALS

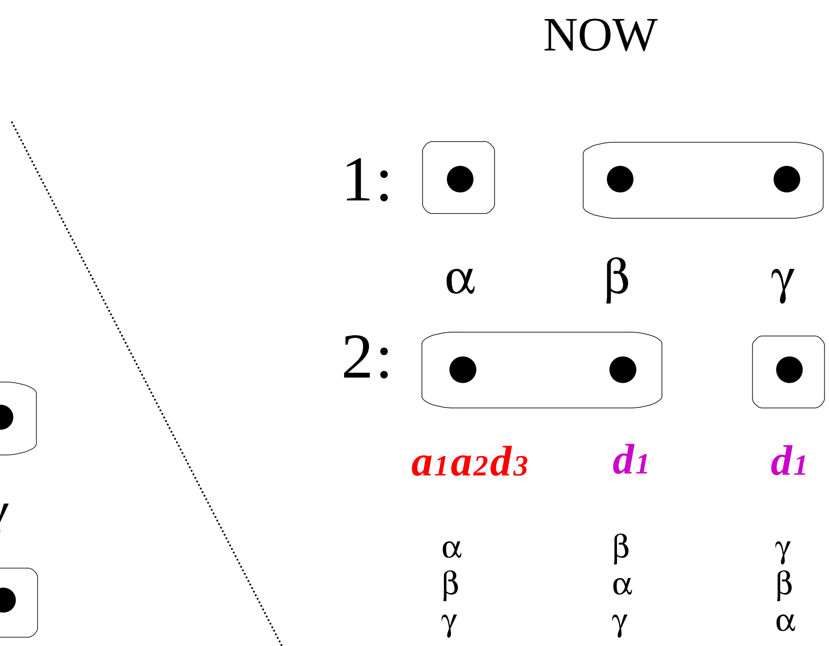
Given a perfect information game define an epistemic model of it as before, but with the following changes:

(1) replace the n functions $\sigma_i : \Omega \to S$ with a single function $d : \Omega \to P$ where P is the set of plays of the game written in terms of actions taken,

(2) add a set of closeness relations $\{\mathcal{P}_\omega\}_{\omega \in \Omega}$



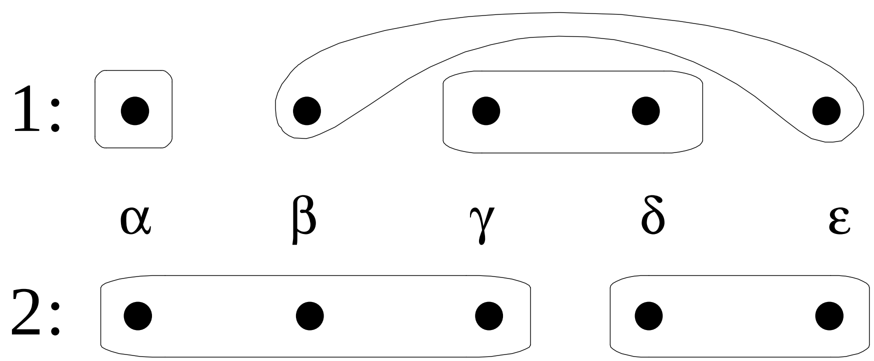


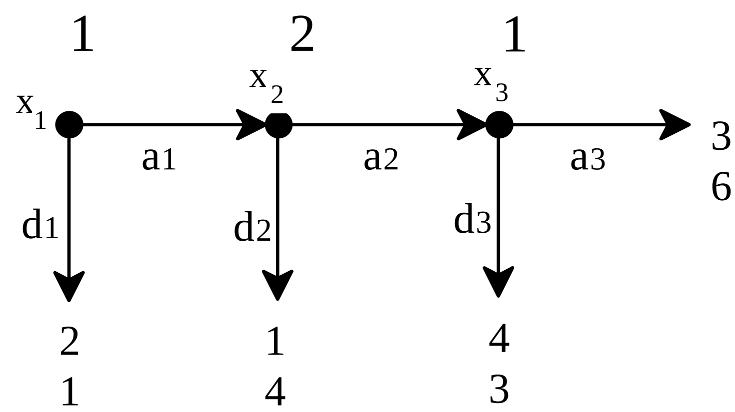


We add two more requirements:

(3) for every play there is at least one state where that play is realized

(4) if, at a state, node *x* of player *i* is reached and he takes action *a* there, then he knows that if *x* is reached he takes action *a*: $\|a\| \subseteq K_i\left(\|x\| \to \|a\|\right)$





a				

$$\|x_2\| = \{\alpha, \beta, \varepsilon\}, \ \|a_2\| = \{a, \beta\}$$

$$\|x_2\| \to \|a_2\| = \neg\|x_2\| \cup \|a_2\| = \{\alpha, \beta, \gamma, \delta\}$$

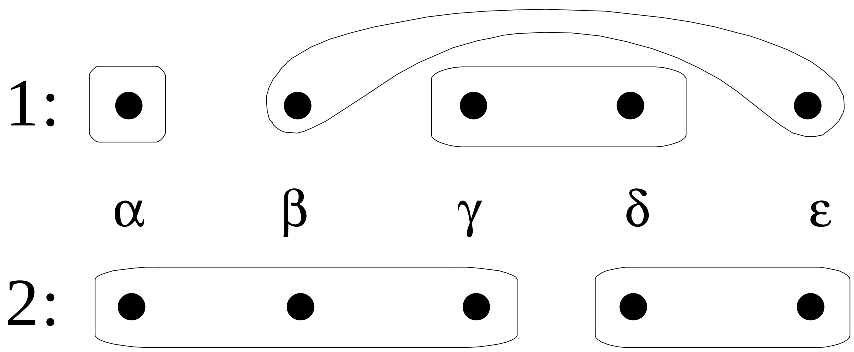
$$K_2\left(\|x_2\| \to \|a_2\|\right) = \{\alpha, \beta, \gamma\}$$

EXTRACTING STRATEGIES FROM A MODEL

Given a model we can extract a strategy profile at every state as follows.

If s_i is a strategy of player i and x_i is a decision node of player i, denote by $s_i(x_i)$ the choice prescribed by s_i at x_i .

Define $\sigma_i(\omega)$ as follows: $\sigma_i(\omega)(x_i) = c_i$ if and only if $\omega \in \|x_i\| \hookrightarrow \|c_i\|$



***a*₁*a*₂*d*₃**	***a*₁*a*₂*a*₃**	***d*₁**	***d*₁**	***a*₁*d*₂**
α	β	γ	δ	ε
β	γ	β	ε	δ
γ	α	α	γ	γ
δ	ε	δ	α	α
ε	δ	ε	β	β

$\sigma_1(\alpha) = a_1 d_3$, $\sigma_1(\beta) = a_1 a_3$

$\sigma_1(\gamma) = d_1 a_3$ (for node x_3 we use state β)

$\sigma_1(\delta) = d_1 d_3$ (for node x_3 we use state α)

$\sigma_1(\varepsilon) = a_1 d_3$ (for node x_3 we use state α)

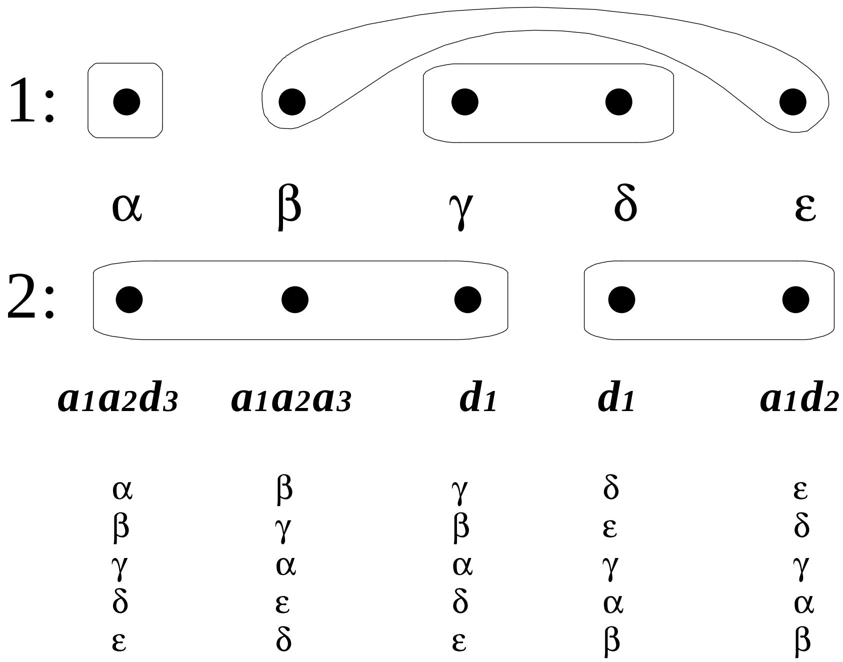
$\sigma_2(\alpha) = a_2$, $\sigma_2(\beta) = a_2$

$\sigma_2(\gamma) = a_2$ (for node x_2 we use state β)

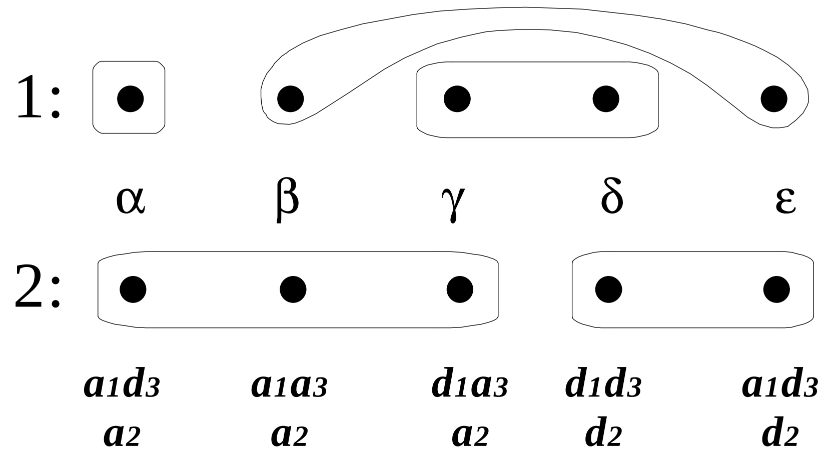
$\sigma_2(\delta) = d_2$ (for node x_2 we use state ε)

$\sigma_2(\varepsilon) = d_2$

From



We get



In this model it is not true that players know their own strategies. E.g. player 1 at state γ

In order for a counterfactual model to give rise to a standard model based on strategies, we need to impose a further condition:

$$(5)\quad \left(\|x_i\| \hookrightarrow \|c_i\|\right) \to K_i\left(\|x_i\| \hookrightarrow \|c_i\|\right)$$

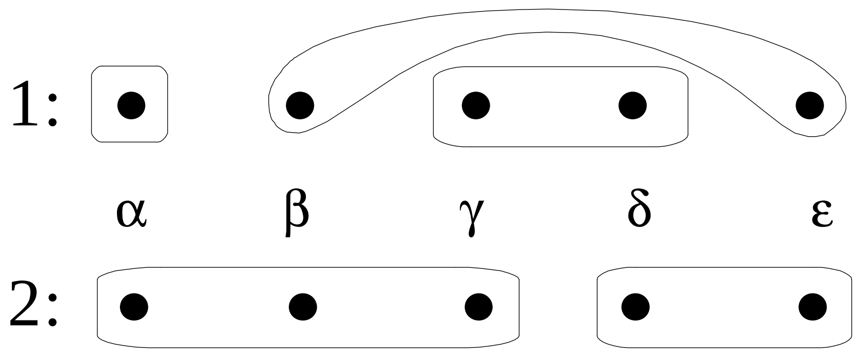
RE-DEFINING RATIONALITY AT REACHED NODES

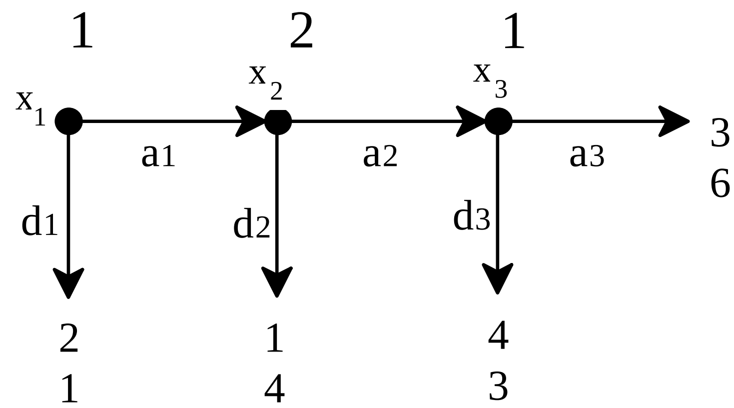
Let x_i be a decision node of player i and c_i and c_i' be two choices of player i at x_i.

If m is a number, let $\|\pi_{\mathrm{i}} = m\|$ be the event that player i's payoff is m.

If k and ℓ are numbers, let $\|k > \ell\| = \Omega$ if $k > \ell$ and $\|k > \ell\| = \varnothing$ otherwise.

$$\|c_i\| \cap \|\pi_i = k\| \cap K_i\left(\|x_i\| \to \left(\| c_i' \| \hookrightarrow \|\pi_i = \ell\|\right)\right) \cap \|\ell > k\| \subseteq \neg \boldsymbol{R}_{\boldsymbol{i}}^{\boldsymbol{RN}}$$

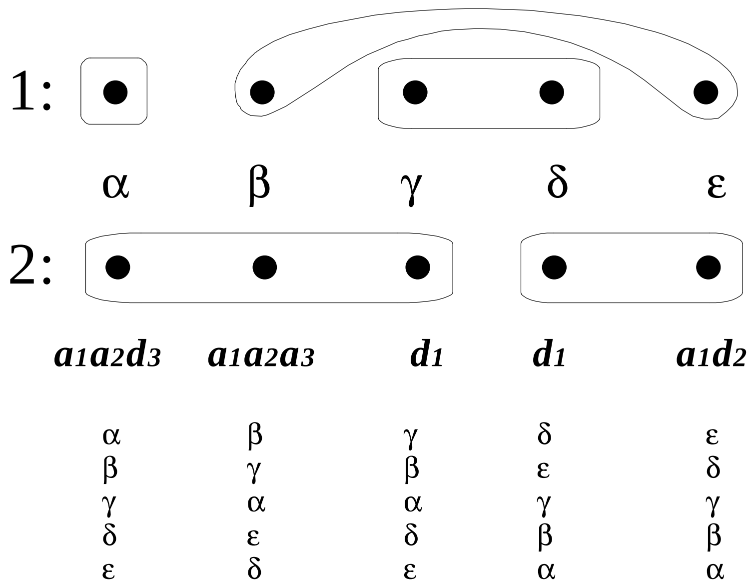


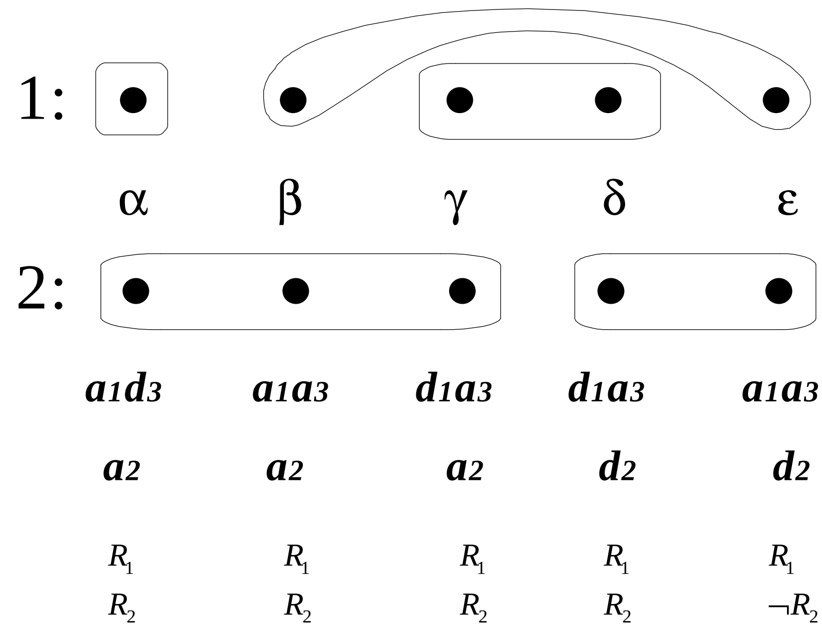


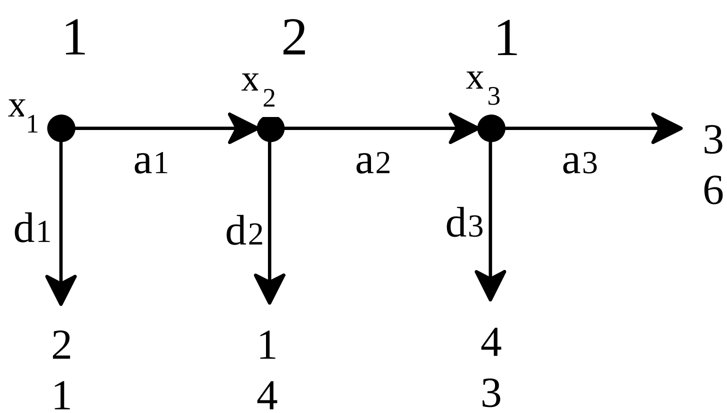
$a_1a_2d_3$	***$a_1a_2a_3$***	***d_1***	***d_1***	***a_1d_2***
α	β	γ	δ	ε
β	γ	β	ε	δ
γ	α	α	γ	γ
δ	ε	δ	β	β
ε	δ	ε	α	α

$\pi_1 = 4$	$\pi_1 = 3$	$\pi_1 = 2$	$\pi_1 = 2$	$\pi_1 = 1$
$d_1 \hookrightarrow \pi_1 = 2$	$d_1 \hookrightarrow \pi_1 = 2$	$a_1 \hookrightarrow \pi_1 = 3$	$a_1 \hookrightarrow \pi_1 = 1$	$d_1 \hookrightarrow \pi_1 = 2$
$\pi_2 = 3$	$\pi_2 = 6$	$\pi_2 = 1$	$\pi_2 = 1$	$\pi_2 = 4$
$d_2 \hookrightarrow \pi_2 = 4$	$d_2 \hookrightarrow \pi_2 = 4$	no choices by 2	no choices by 2	$a_2 \hookrightarrow \pi_2 = 6$
R_1	R_1	R_1	R_1	R_1
R_2	R_2	R_2	R_2	$\neg R_2$
		use β	use ε	use β

Thus no common knowledge of rationality at any state.







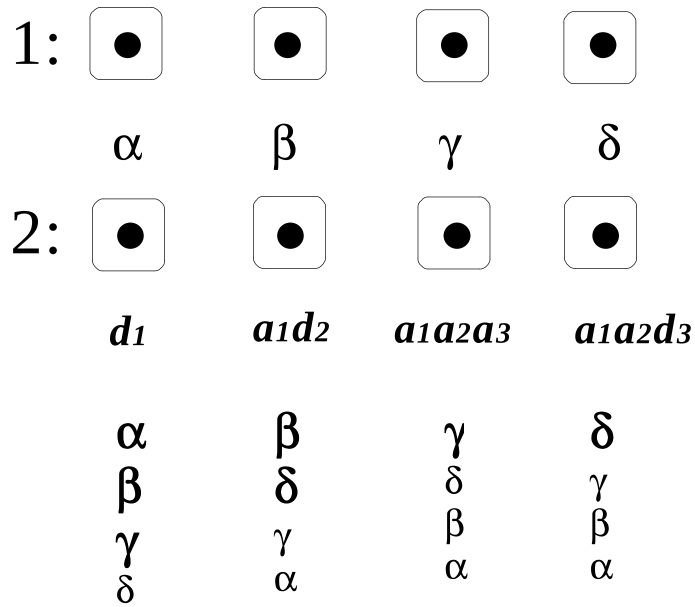
The corresponding strategy-based model

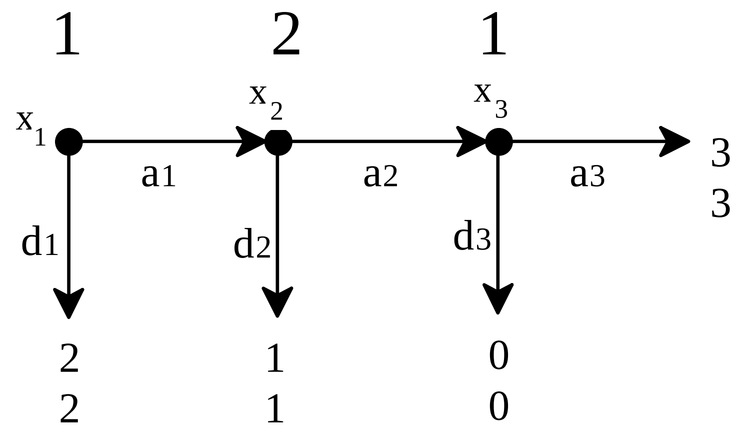
Redefining substantive rationality (Stalnaker's notion)

$$\boldsymbol{R_i^{SR}} = \bigcap_{x_i \in X_i} \left(\| x_i \| \hookrightarrow \boldsymbol{R_i^{RN}} \right)$$

rationality at all nodes: reached and un-reached

Does common knowledge of substantial rationality so defined imply the backward-induction play?





$$\boldsymbol{R_1^{RN}} = \{\alpha, \gamma\}$$

$$\boldsymbol{R_2^{RN}} = \{\alpha, \beta, \gamma\}$$

At state α there is common knowledge of substantive rationality. The following is true at α:

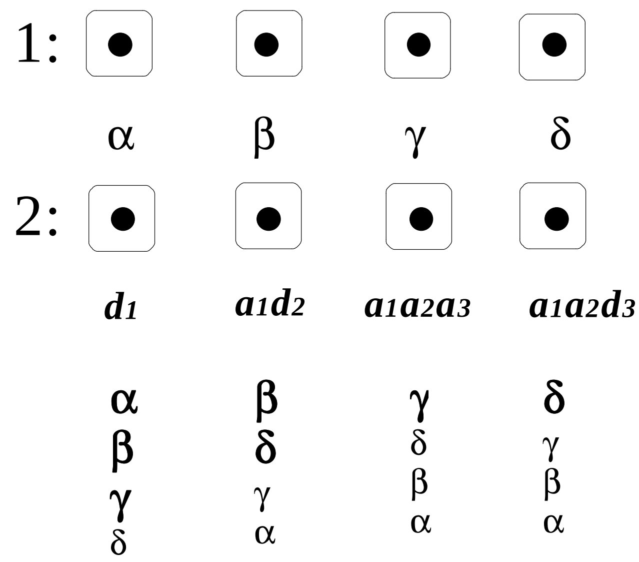
(1) 1 is materially rational at x_1 : 1 knows that if he played a_1 then 2 would play d_2. **[state β]**

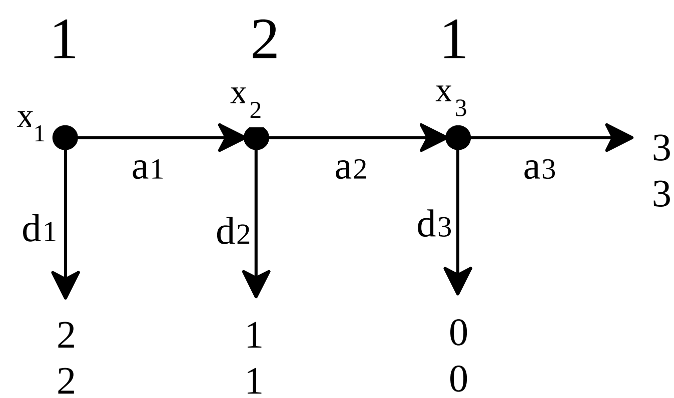
(2) 2 is materially rational (does not do anything) but also substantively rational: if x_2 were reached **[state β]** then player 2 would be materially rational (she would play d_2 knowing that if she played a_2 then 1 would play d_3) **[state δ]**.

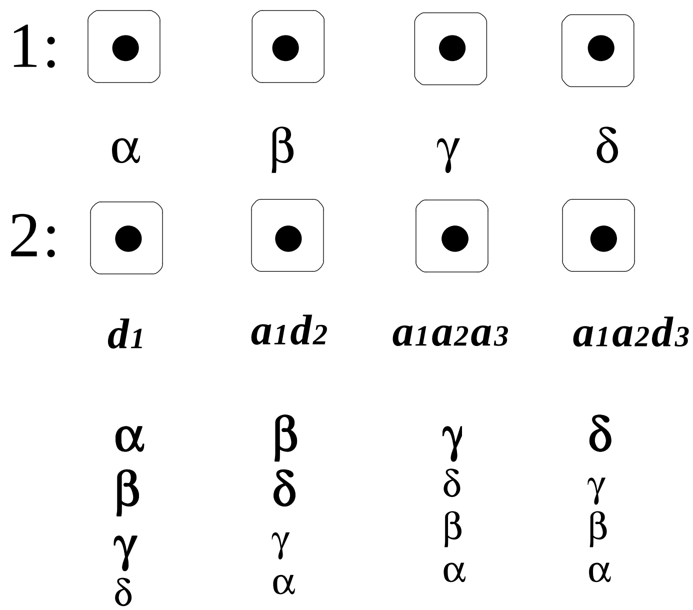
(3) 1 is substantively rational at x_3 : if x_3 were reached he would play a_3 **[state γ]**.

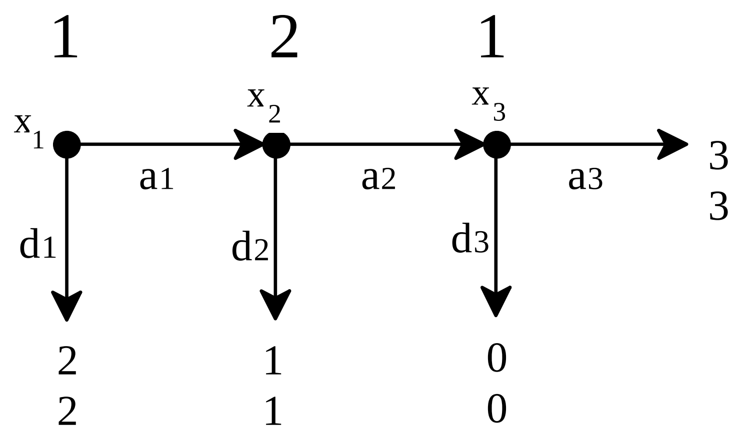
Stalnaker (1998 p. 48)

Player 2 has the following initial belief: player 1 would choose a_3 on her second move *if* she had a second move. This is a causal 'if' – an 'if' used to express 2's opinion about 1's *disposition to act* in a situation that they both know will not arise. Player 2 knows that since player 1 is rational, if she somehow found herself at her second node, she would choose a_3 . But to ask what player 2 would believe about player 1 *if* he learned that he was wrong about 1's first choice is to ask a completely different question – this 'if' is epistemic; it concerns player 2's belief revision policies, and not player 1's disposition to be rational. No assumption about player 1's substantive rationality, or about player 2's knowledge of her substantive rationality, can imply that player 2 should be disposed to maintain his belief that she will act rationally on her second move even were he to learn that she acted irrationally on her first.

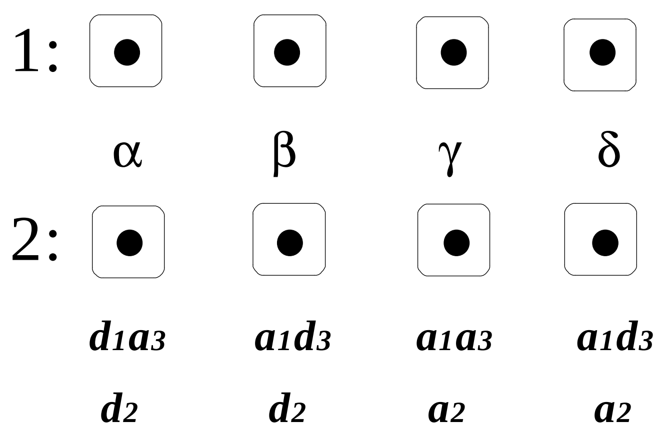




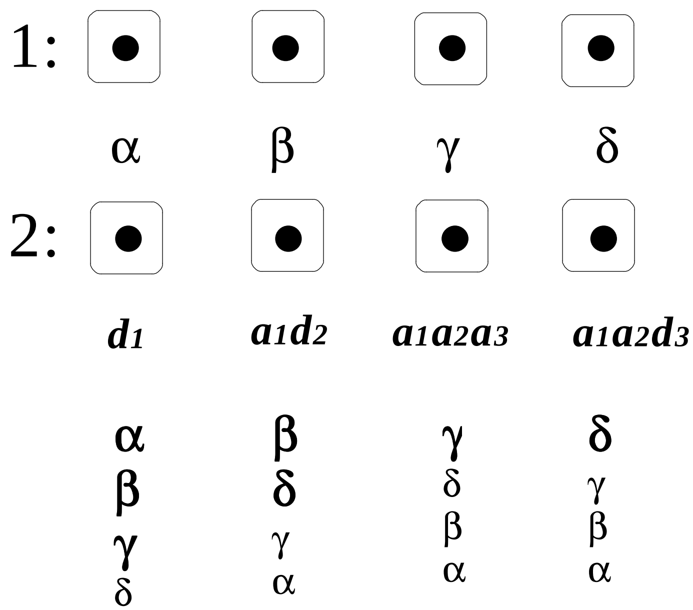


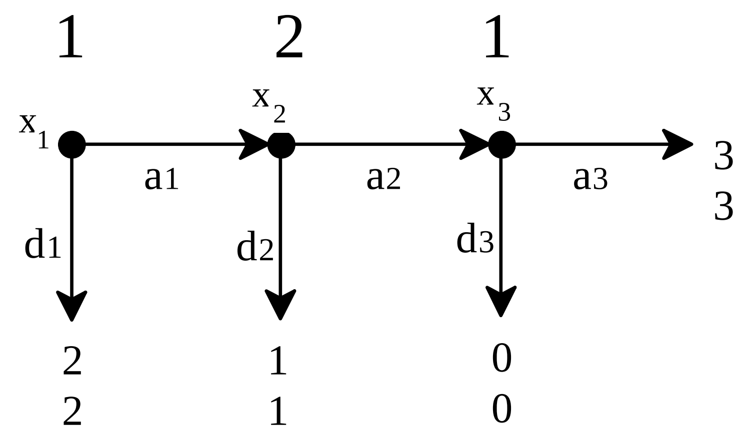


The corresponding strategy-based model is:



According to Aumann, player 2 is not substantively rational at α: player 2 is planning to play d_2 knowing that player 1 would play a_3.





$$\alpha \in K_2\left(\|x_3\| \hookrightarrow a_3\right) \text{ and also } \alpha \in \|x_2\| \hookrightarrow K_2\left(\|x_3\| \hookrightarrow d_3\right)$$

Thus what player 2 believes about player 1's behavior in the hypothetical world where node x_3 is reached changes going from node x_1 (where the game ends without node x_2 being reached) to the hypothetical world where x_2 is reached. *If one imposes the constraint that such changes cannot happen, then common knowledge of substantive rationality implies the backward-induction play.*

ADDITIONAL REFERENCES

Aumann, R., Backward induction and common knowledge of rationality, *Games and Economic Behavior*, 1995, 8: 6-19.

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

Samet, D., Hypothetical knowledge and games with imperfect information, *Games and Economic Behavior*, 1996, 17: 230-251.

Stalnaker, R., Belief revision in games: forward and backward induction, *Mathematical Social Sciences*, 1998, 36: 31–56