Computational Logic Seminar, GC CUNY

To believe, or not to believe: that is the question.

Sergei Artemov

New York City, August 31, 2010

To believe, or not to believe

Two rational players, Ann and Bob, with common belief of rationality playing a game of perfect information. The course of the game depends on how Bob would react to being surprised by Ann's hypothetical irrational move. There are various options:

1. Bob revises his belief about Ann's rationality for the remainder of the game.

2. Bob maintains his belief in Ann's rationality for the remainder of the game.

Stalnaker describes what happens when (1) is allowed.
We claim that (2) also makes perfect sense and study this case.

Does one mistake disqualify?

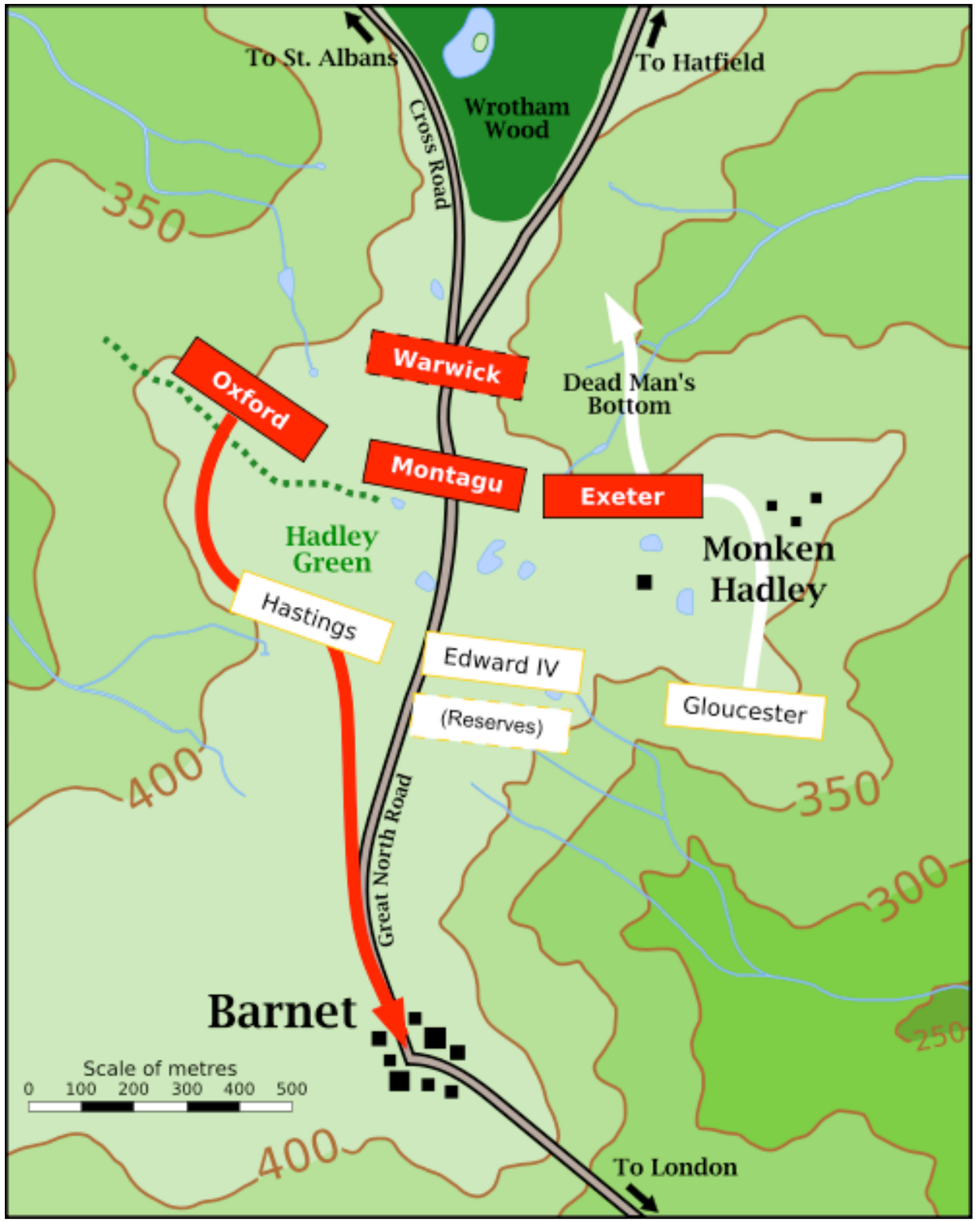
Not really:

1. Random error
2. Friendly fire
3. Learning mistakes
4. Age-related mistakes
5. Communication errors
6. Implementation errors
7. etc.,

are all examples of non-disqualifying mistakes.

Losing trust can be lethal

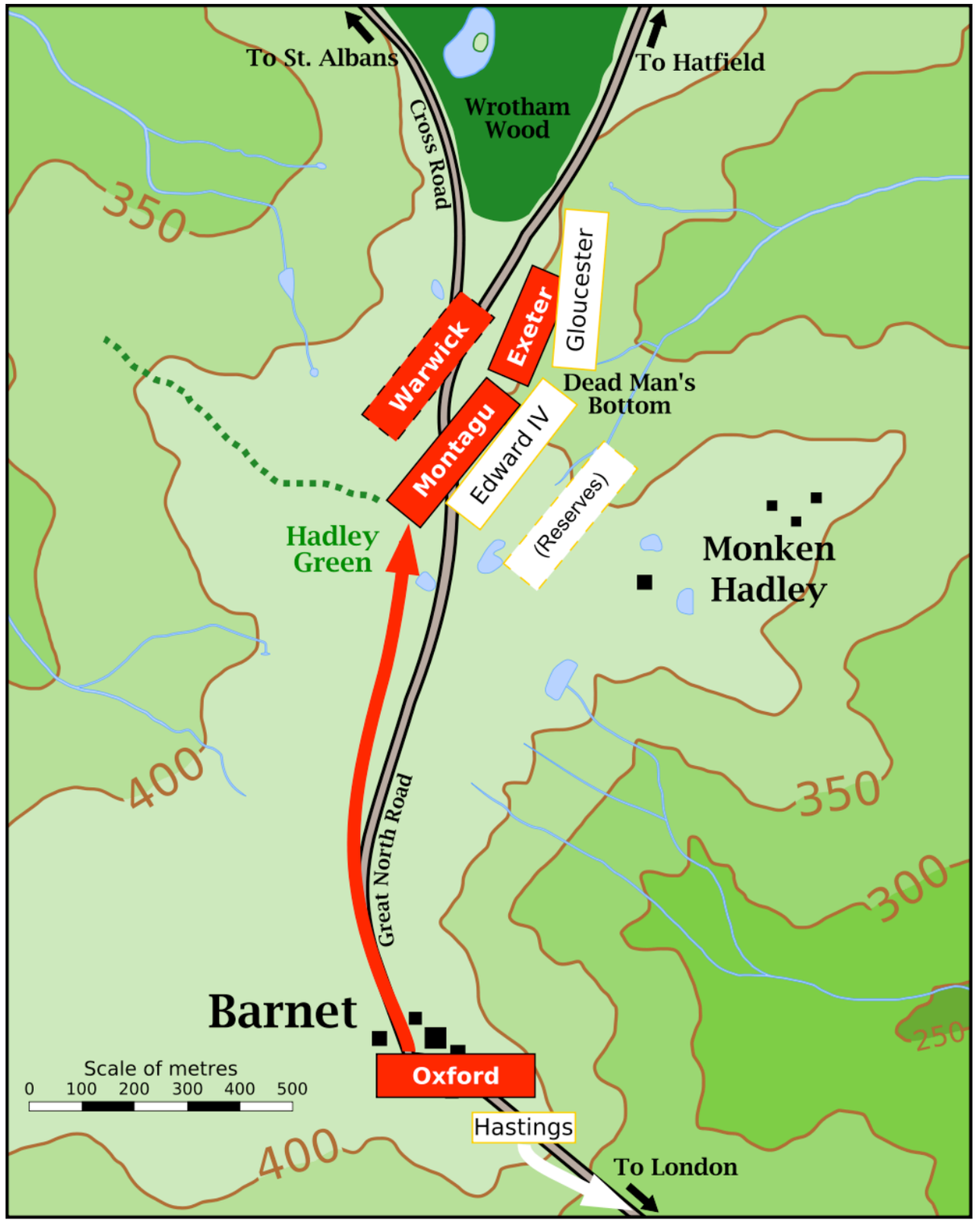
In **the Battle of Barnet**, April 14, 1471, Edward IV led the House of York in a fight against the House of Lancaster, which backed Henry VI for the throne. The Battle ended in Edward's victory and became a decisive turn of fortune in **the Wars of the Roses**.



Oxford quickly overwhelmed Hastings and then retraced his steps through the fog back to the fight. His group arrived, unexpectedly, at Montagu's rear. Obscured by fog, Montagu's men assumed their allies were Edward's reserves and unleashed a volley of arrows. Oxford and his men immediately cried treachery, struck back and began withdrawing from the battle.

Losing trust can be lethal

In **the Battle of Barnet**, April 14, 1471, Edward IV led the House of York in a fight against the House of Lancaster, which backed Henry VI for the throne. The Battle ended in Edward's victory and became a decisive turn of fortune in **the Wars of the Roses**.



Oxford quickly overwhelmed Hastings and then retraced his steps through the fog back to the fight. His group arrived, unexpectedly, at Montagu's rear. Obscured by fog, Montagu's men assumed their allies were Edward's reserves and unleashed a volley of arrows. Oxford and his men immediately cried treachery, struck back and began withdrawing from the battle.

The Hunt for Red October

Officer: *Combat tactics, Mr. Ryan. By turning into the torpedo, the captain closed the distance before it could arm itself.*

Jack Ryan (Alec Baldwin): *So that's it?*

Capitain Ramius (Sean Connery): *Not quite. Right now, Captain Tupolev is removing the safety features on all his weapons.* ***He won't make the same mistake twice.***

Experience is worth it

Russian proverb:
За одного битого двух небитых дают.

Literal:
A beaten one is worth two unbeaten ones.

"...his father saw him and was filled with compassion for him; he ran to his son, threw his arms around him and kissed him."

– Luke 15:17-20, NIV

Rembrandt, Return of the Prodigal Son, 1662,
(Hermitage Museum, St Petersburg)

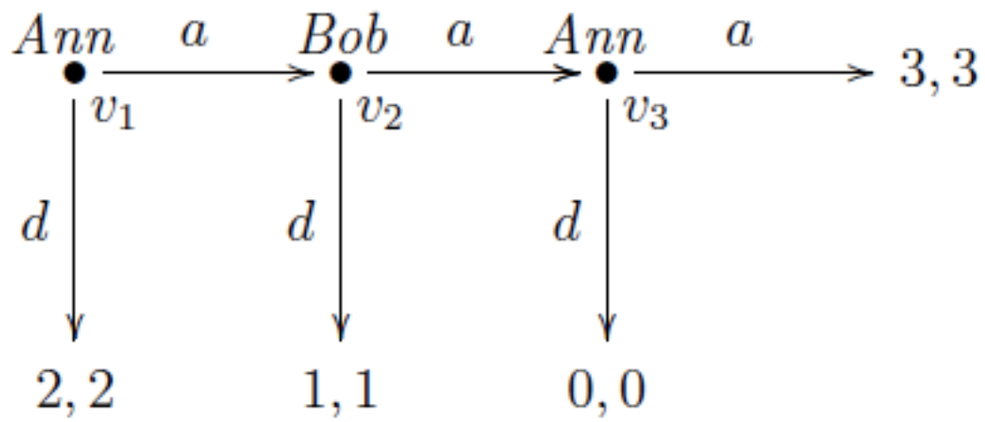
Does one mistake disqualify?

Perhaps, a complete disqualification after a single mistake is the exception rather than the rule.

We need an analysis of games that models a certain degree of error-tolerance.

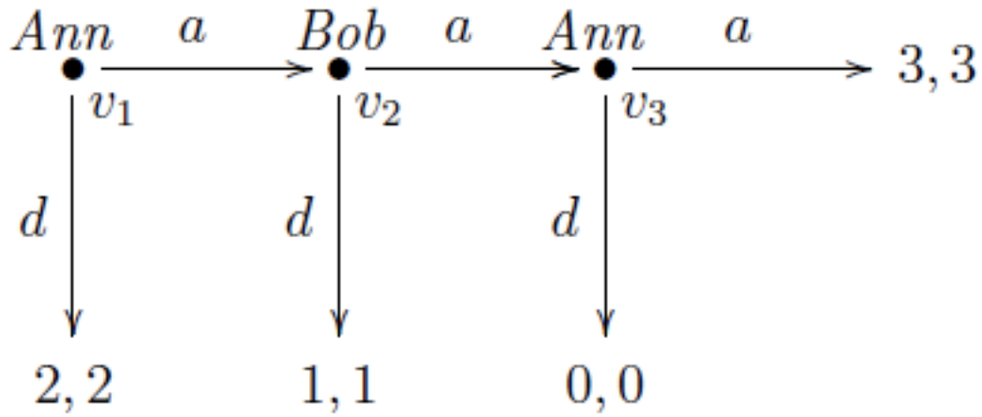
Belief revision in games

Stalnaker's approach to games of perfect information (PI games) introduces belief revision into players' reasoning. The paradigmatic example is given by this simple common interest game. In Aumann's setting, given common knowledge of rationality, players play the



backward induction solution, i.e., *across* at all three nodes. Stalnaker's approach, with 'the same' epistemic assumptions, claims that the solution (dda), i.e., Ann plays *down* at v_1, Bob plays *down* at v_2, and Ann plays *across* at v_3, if commonly known is also commonly known to be rational.

Stalnaker reasoning



We assume that (*dda*) is commonly known and check that players are rational at each node.

- Ann is rational at v_3 by the game tree.
- Bob is rational at v_2 since if Ann were to play across (an obviously irrational move by Ann given her knowledge that Bob is playing *down*), then Bob revises his initial belief of Ann's rationality and no longer assumes that Ann will play *across* at v_3. Under these circumstances, playing *down* at v_2 is not irrational for Bob.
- Ann is rational at v_1 since she knows that Bob is playing *down*.

Stalnaker reasoning

In this proof, the heart of the matter is how Bob would react to being surprised by Ann's irrational move *across* at v_1. Stalnaker describes what happens when Bob revises his beliefs about Ann's rationality for the remainder of the game, which makes good sense. This case was cast in a formal logical framework by Halpern in 2001.

What is good about Stalnaker's approach?

It made belief revision an issue in Game Theory.

Stalnaker reasoning: reservations

1. ***An artificial example.*** The aforementioned Ann-Bob game is a PI game, with the unmotivated epistemic constraint that *dda* is commonly known.

2. ***A made-up juxtaposition with Aumann's Theorem on Rationality.*** Epistemology usually attributes to knowledge a certain indefeasibility (infallibility, reliability, truth-tracking, necessity, etc.). What is known, is true in a robust way and is not subject to revision. Aumann's assumption *CKR*

*common **knowledge** of players' rationality,*

does not suggest the possibility of revising the rationality condition.
The Stalnaker setup is a fit for a different well-known assumption *RCBR*

*players' rationality and common **belief** of player' rationality*.

3. ***Does not accommodate other revision policies***, e.g., robust belief of rationality, error-tolerance, virtue of experience, etc.

Extensive Games

Let us recap basic terminology ([2, 6]). An extensive game consists of the following components.

1. A finite set $N = \{1, 2, \ldots, n\}$ of players.

2. A finite rooted tree H. Each node has a unique path from the root called the history of this node. The leaves of the game tree are called terminal nodes, or outcomes. The set of all terminal nodes is called Z.

3. A player function P that assigns a player (who makes a move) to each nonterminal node.

4. For each player i, a payoff function u_i defined on Z.

The root node is the starting point of the game. At any node $v \in (N \setminus Z)$, player $P(v)$ chooses one of the successor nodes (move).

There is no specification of epistemic state of players yet!
This leaves a room for more studies, paradoxes, speculations, etc., and we are entering this room now...

Aumann Models

An **Aumann model** is a tuple $\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$, where Ω is a set of "epistemic states" of the world, $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are knowledge partitions of Ω corresponding to players $1, 2, \ldots, n$, and $\mathbf{s}$ is a mapping from Ω to the set of all strategy profiles: for a state ω,

$$\mathbf{s}(\omega) = (s_1, \ldots, s_n).$$

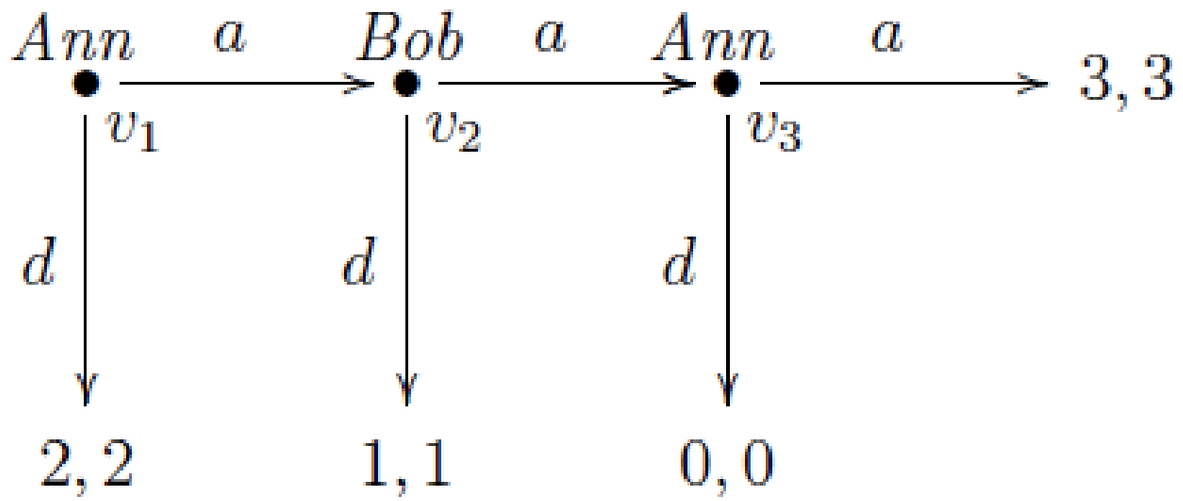
We write $\mathbf{s}_i(\omega)$ for i's component of the strategy profile $\mathbf{s}(\omega)$, i.e., s_i. Also, let (s_{-i}, s^i) be the strategy profile obtained from s by replacing s_i by s^i, $h_i^v(s)$ be i's conditional payoff if strategy profile s is followed starting at v, and $\mathcal{K}_i(\omega)$ be the cell in $\mathcal{K}_i$ that includes ω. We assume that players know their strategies ("measurability" property), that means that if $\omega' \in \mathcal{K}_i(\omega)$, then $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$; that is, i uses the same strategy at all the states in a cell of $\mathcal{K}_i$.

Given AM, each epistemic logical formula *F* built from "Ann knows," "Bob knows," and specific move propositions "player *i* chooses move *j* at node *v*," receives a definitive truth value in any given state ω

$$\omega \vDash F \quad \text{or} \quad \omega \nvDash F .$$

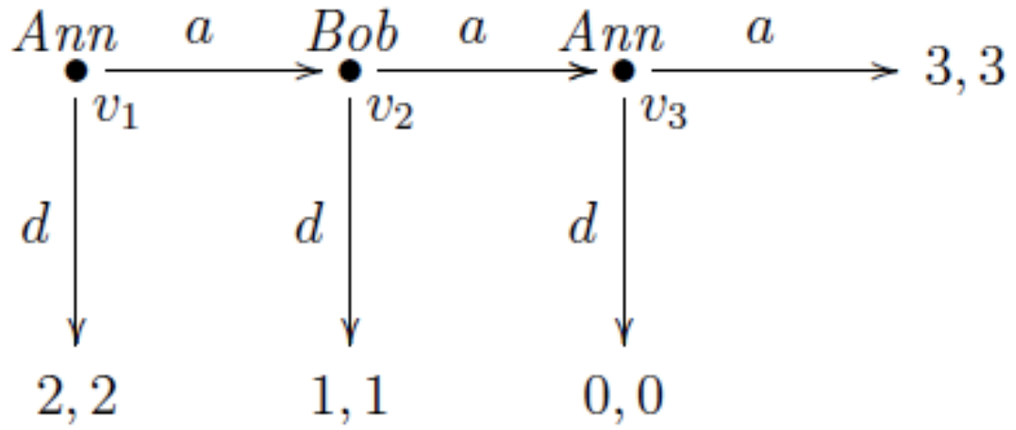
Aumann Models: Example

Model AM_1:



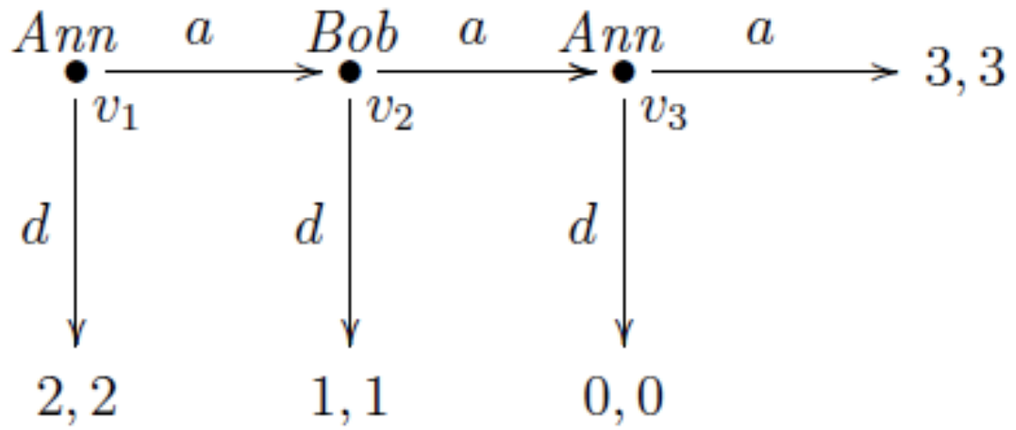
- s^1 is the strategy profile (dda), i.e., Ann plays *down* at v_1, Bob plays *down* at v_2, and Ann plays *across* at v_3;
- s^2 is the strategy profile (ada);
- s^3 is the strategy profile (add);
- s^4 is the strategy profile (aaa);
- s^5 is the strategy profile (aad).
- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$; $\mathbf{s}(\omega_j) = s^j$ for j =1–5

Aumann Models: Example



- s^1 is the strategy profile (dda), i.e., Ann plays *down* at v_1, Bob plays *down* at v_2, and Ann plays *across* at v_3;
- s^2 is the strategy profile (ada);
- s^3 is the strategy profile (add);
- s^4 is the strategy profile (aaa);
- s^5 is the strategy profile (aad).
- $\mathcal{K}_{Ann} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;
- $\mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;

Aumann Models: Example



- s^1 is the strategy profile (dda), i.e., Ann plays *down* at v_1, Bob plays *down* at v_2, and Ann plays *across* at v_3;
- s^2 is the strategy profile (ada);
- s^3 is the strategy profile (add);
- s^4 is the strategy profile (aaa);
- s^5 is the strategy profile (aad).
- $\mathcal{K}_{Ann} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;
- $\mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;

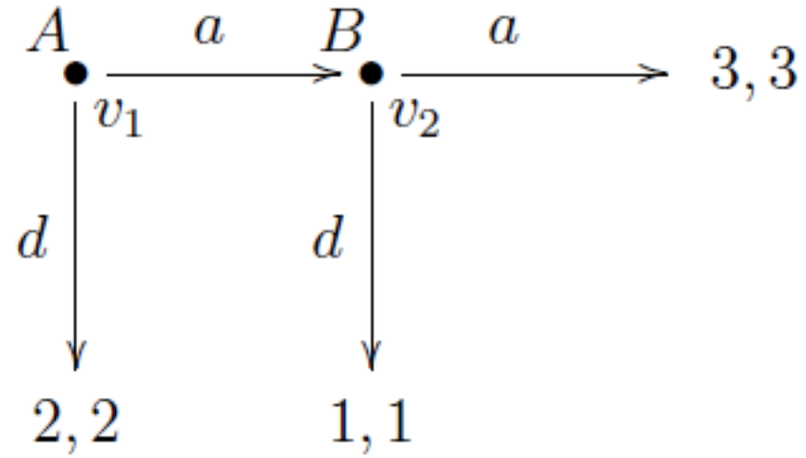
In state ω_1, it is common knowledge that Ann plays *down* at v_1, *across* at v_3, and that Bob plays *down* at v_2.

In state ω_2, Ann knows that Bob plays *down*, Bob knows that Ann plays *down* at v_1 but considers either move by Ann at v_3 possible.

Aumann Models: Pros and Cons

Aumann models are the game-theoretical equivalent of canonical models = collections of maximal S5-consistent sets. Like canonical models, Aumann models are capable of representing any epistemic condition concerning the moves propositions.

An Aumann model does not specify the game.
Consider model ***AM₂*** with
$\mathbf{s}(\omega_1) = dd$, $\mathbf{s}(\omega_2) = da$, $\mathbf{s}(\omega_3) = aa$,
$\mathcal{K}_{Ann} = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$,
$\mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$.



The question of whether Ann knows that Bob will play *across* depends on the state: YES in ω_3, NO in ω_1, ω_2. On the other hand, in a real game with real players, this question should have a definitive answer.

An Aumann model + a state overspecifies the game: it describes not only players' knowledge, but also the players' choices which are not necessarily determined by the game rules.

Aumann Models: Pros and Cons

Aumann models allow for problematic games. Consider AM_3 which is defined as AM_1 but with $\mathcal{K}_{Ann} = \mathcal{K}_{Bob} = \{\{\omega_1\}\}$. This game may be regarded as the result of a public announcement that *dda* is played.

This "game" looks problematic since each player's strategy is common knowledge. There is no "game" here.

Perhaps it is worth studying 'regular games' in which epistemic conditions are limited to player's rationality rather than specific moves.

Aumann's account of rationality

Informal Account = Rationality as Reputation

Rationality of a player means that he is a habitual pay-off maximizer: that no matter where he finds himself – at which vertex – he will not knowingly continue with a strategy that yields him less than he could have gotten with a different strategy.

This is a remarkably epistemic approach, "knowingly" being a key word. This account is rather about "rationality as reputation" though further formalizations deviate considerably from this spirit.

Definitions of Aumann rationality

An informal definition first:

Definition 1 *Player i is rational at vertex v if there is no strategy that i could have used that he knows would net him a conditional higher payoff than the strategy he actually uses.*

A completely formal definition:

Definition 2 *Player i is rational at vertex v in state ω if, for all strategies $s^i \neq \mathbf{s}_i(\omega)$,*

$$h_i^v(\mathbf{s}(\omega')) \geq h_i^v(\mathbf{s}_{-i}(\omega'), s^i)$$

for some $\omega' \in \mathcal{K}_i(\omega)$. Player i is rational in state ω if i is rational at any node in ω.

Definitions of Aumann rationality

Definition 2 *Player i is rational at vertex v in state ω if, for all strategies $s^i \neq \mathbf{s}_i(\omega)$,*

$$h_i^v(\mathbf{s}(\omega')) \geq h_i^v(\mathbf{s}_{-i}(\omega'), s^i)$$

for some $\omega' \in \mathcal{K}_i(\omega)$. Player i is rational in state ω if i is rational at any node in ω.

Rather, it defines '**irrationality**.' Player *i* is *irrational* (at a node *v* in state ω) if *i* can do strictly better by using some other strategy against all the strategy profiles of the other players that he considers possible at ω.

There is a ***hidden independency assumption***: for any strategy s^i and any possible state ω', strategy profile $(\mathbf{s}_{-i}(\omega'), s^i)$ is deemed possible by *i* at *v*.

Definitions of Aumann rationality

Note that these characterizations of rationality (Informal Account, Definition 1, and Definition 2) lead to a different analysis.

Informal Account = Rationality as Reputation is close to the condition "holds in each state" typical in public announcements, the universal modality, the McCarthy "any fool knows" modality, justified common knowledge, etc., which is quite different from statewise rationality, even with the common knowledge assumptions.

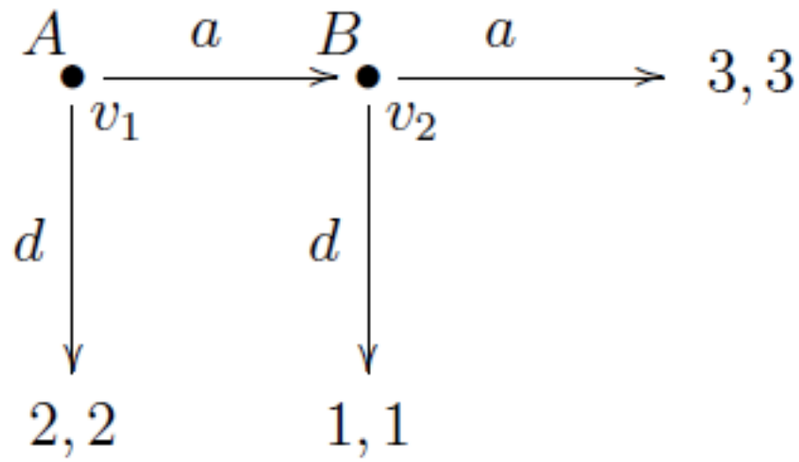
Definition 1 is statewise, hence different from Rationality as Reputation.

Definition 2 is also statewise, but works only under special epistemic assumptions, e.g., the Independence Condition (see the next two slides), and fails in some other natural PI games.

Aumann Models: Pros and Cons

Aumann models do not represent some reasonable PI games.

Game 2 has this game tree and the following commonly known epistemic conditions:
1. if Ann plays *across*, then Bob plays *across*;
2. if Ann plays *down*, then Bob plays *down*.



This extensive game does not seem to be fairly represented by any Aumann model. Indeed, there are two possible profiles: *dd*, *aa*, (states ω_1 and ω_3).

$$\mathcal{K}_{Ann} = \{\{\omega_1\}, \{\omega_3\}\}.$$

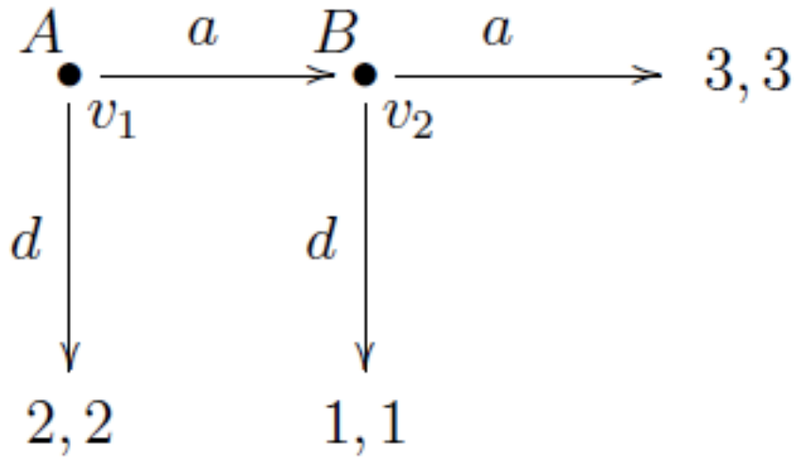
Let us try to answer the question of whether it is rational for Ann to play *down*. The intuitive answer, as well as the answer suggested by Definition 1 is NO. Indeed, Ann knows that playing *across* will net her 3, *down* only 2.

Definition 2 applied to $\mathcal{K}_{Ann}$, however, gives a different account: ***Ann is rational in both states***. In ω_1, Ann's choice is *down*. The alternative choice is *across*, which could bring Ann only 1 since Bob chooses *down*.

Aumann Models: Pros and Cons

Aumann models do not represent some reasonable games.

Game 2 has this game tree and the following commonly known epistemic conditions:
1. if Ann plays *across*, then Bob plays *across*;
2. if Ann plays *down*, then Bob plays *down*.



An Aummann model proponent could say that Bob's strategy is defined ***under the assumption that Bob's node is reached***. Since, by (1), Bob plays *across* if v_2 is reached, Ann's condition (2) becomes impossible: even if Ann plays *down*, she knows that Bob's choice at v_2 is *across*. The resulting Aumann model for Game 2 could then be ***AM₃***:

$$\mathcal{K}_{Ann} = \{\{\omega_2\}, \{\omega_3\}\}; \quad \mathcal{K}_{Bob} = \{\{\omega_2, \omega_3\}\}$$

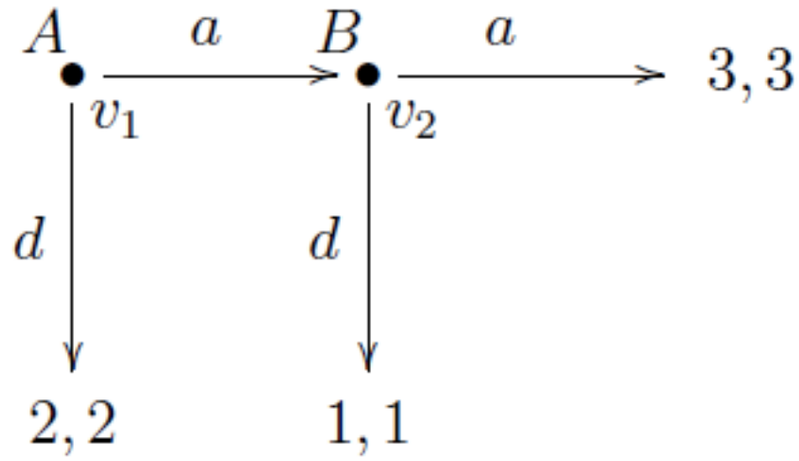
which does not reflect Game 2, although having the same rational solution *aa* as Game 2.

In Aumann models, a player chooses his/her strategy at the beginning of the game and hence this choice does not depend on actual moves by other players (Independence Condition). This is a limitation of the model.

Aumann Models: Pros and Cons

Aumann models do not fairly represent some reasonable games.

Game 2 has this game tree and the following commonly known epistemic conditions:
1. if Ann plays *across*, then Bob plays *across*;
2. if Ann plays *down*, then Bob plays *down*.



It might appear that the Independence Condition does not influence the game analysis: if a vertex is not reached, then choices at said vertex do not alter the game path/outcome. This argument does not work if the rationality analysis is involved. Definition 2 considers arbitrary combinations of any i-th player's strategy s^i with any other players' strategies $\mathbf{s}_{-i}(\omega')$ deemed possible by i and this requires the Independence Condition.

For example, in Game 2, the Independence Condition does not hold and Definition 2 applied to a simplistic Aumann model for Game 2 leads to a counter-intuitive conclusion that playing *down* is rational for Ann.

Definition 1 is applied even when the Independence Condition does not hold.

Belief revision models

Extended models formalize Stalnaker's representation of counterfactuals via the selection function "the closest world where a given vertex is reached." In a formal setting, the extended model is a tuple

$$\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$$

where $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$ is an Aumann model and a selection function f maps pairs of states and vertices to states. The intended reading of $f(\omega, v) = \omega'$ is

ω' is the closest state to ω in which vertex v is reached.

It is assumed that f satisfies the following conditions:

F1. Vertex v is reached in $f(\omega, v)$.

F2. If v is reached in ω, then $f(\omega, v) = \omega$.

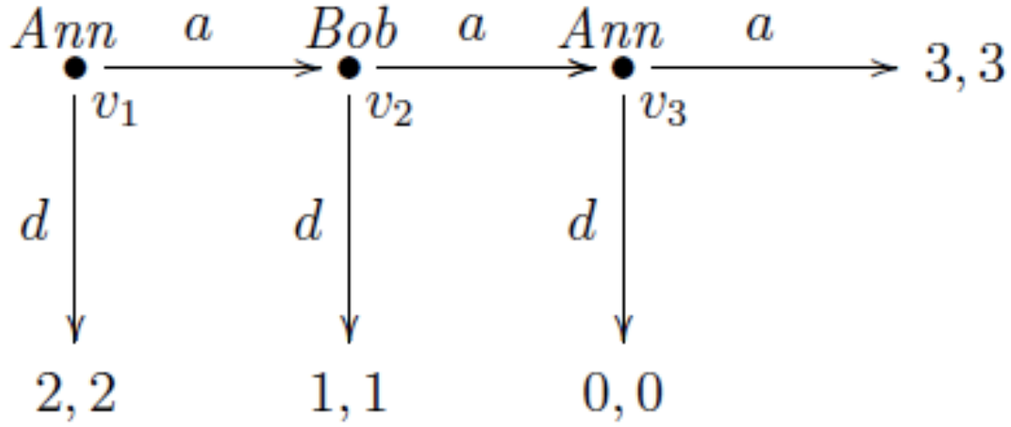
F3. $\mathbf{s}(f(\omega, v))$ and $\mathbf{s}(\omega)$ agree on the subtree of the game tree at and below v.

Belief revision models: example

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,

$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$

$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



There exists a unique selection function here:

$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

Definition 3 (Halpern) *Player i is* **Stalnaker-rational** *in state ω at vertex v if i is rational at vertex v in $f(\omega, v)$. Player i is Stalnaker-rational in state ω if i is Stalnaker-rational at any of its vertices in ω.*

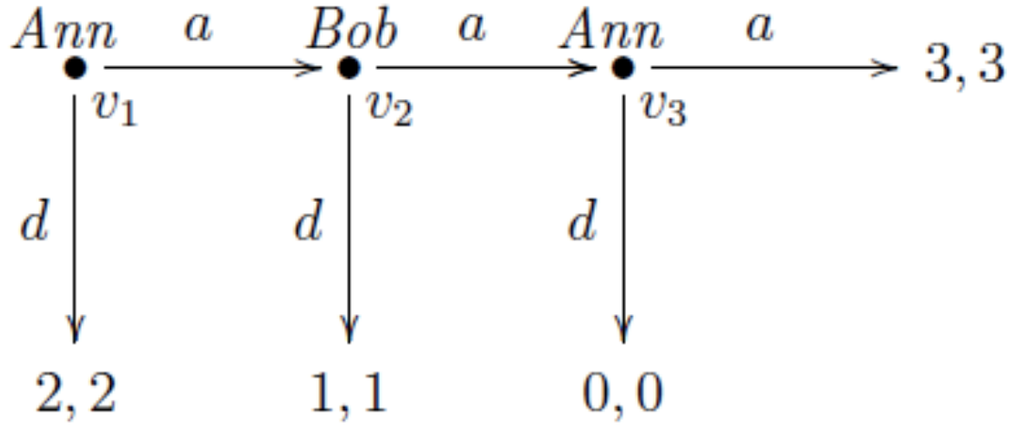
Stalnaker rationality spills over epistemic reachability - state $f(\omega,v)$ can be unreachable from ω - which is an indication that reachability-based common knowledge may be insufficient here.

Belief revision models: example

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,

$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$

$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



There exists here a unique selection function:

$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

Definition 3 (Halpern) *Player i is* **Stalnaker-rational** *in state ω at vertex v if i is rational at vertex v in $f(\omega, v)$. Player i is Stalnaker-rational in state ω if i is Stalnaker-rational at any of its vertices in ω.*

Bob is Stalnaker-rational at v_2 in state ω_1 if Bob is (Aumann-)rational at $f(\omega_1,v_2) = \omega_2$ which is not reachable from ω_1.

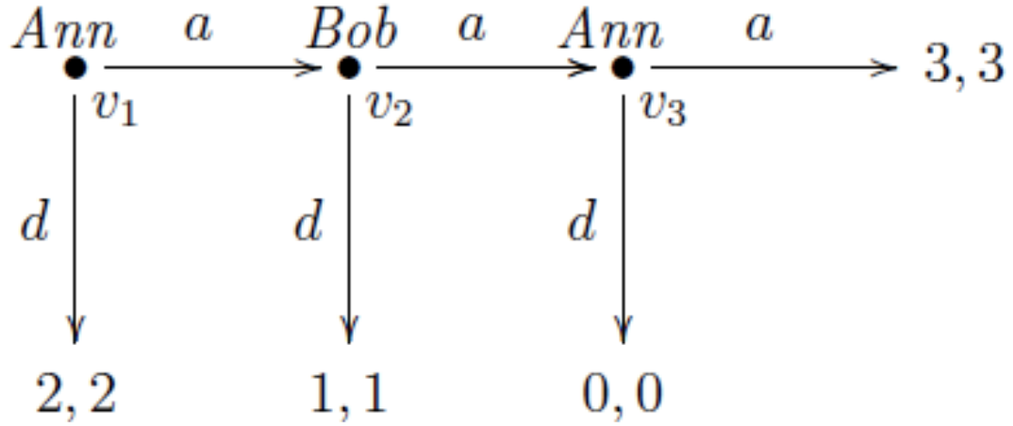
Stalnaker rationality spills over common knowledge as reachability.

Stalnaker reasoning, formally

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,

$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$

$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$,
and $f(\omega, v) = \omega$ in all other situations.

Stalnaker rationality is common knowledge in ω_1. (3)

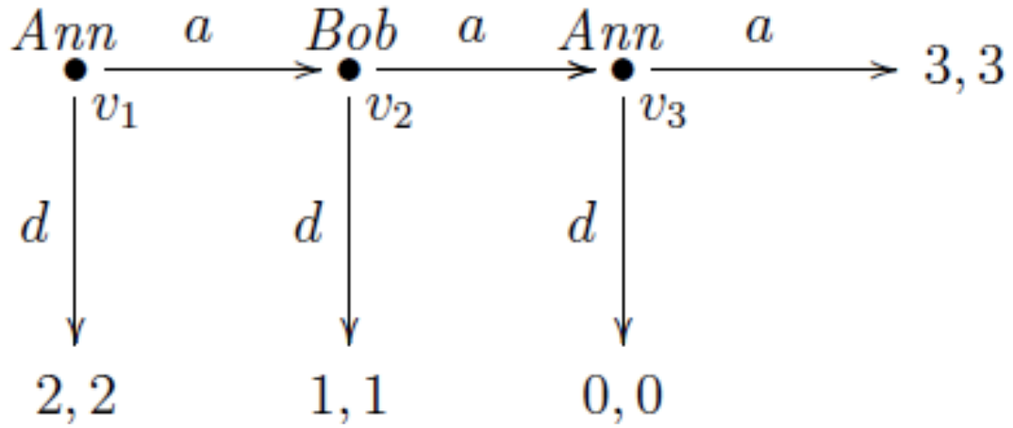
Indeed, since $\mathcal{K}_{Ann}(\omega_1) = \mathcal{K}_{Bob}(\omega_1) = \{\omega_1\}$, everything that is true in ω_1 is common knowledge in ω_1. Stalnaker rationality of both players holds in ω_1, in particular, Bob is Stalnaker-rational in ω_1 at v_2. Selection function f reduces this question to the claim that Bob is (Aumann-) rational in ω_2 at vertex v_2 which is established by direct application of Definition 2.

Stalnaker reasoning, formally

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,

$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$

$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$

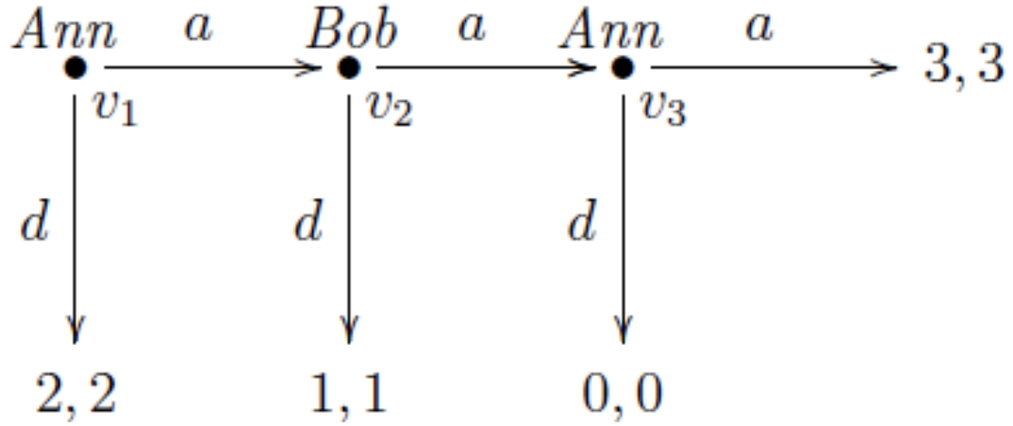


$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

The problem is that in state ω_2 at vertex v_2 **Bob cannot know that Ann is Stalnaker-rational**. Indeed, Ann is not Stalnaker-rational in ω_3 (since $f(\omega_3, v_3) = \omega_5$ and Ann in not rational in ω_5 at v_3), and $\omega_3 \in \mathcal{K}_{Bob}(\omega_2)$. Speaking informally, following selection function $f(\omega_1, v_2) = \omega_2$, Bob in ω_1 revises his belief that Ann plays *down* at v_1 and considers the case ω_2 in which Ann plays *across* at v_1. Accidentally, Bob also forfeits his knowledge of Ann's rationality at v_3, thus treating this knowledge as mere belief.

Stalnaker reasoning, formally

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,
$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$
$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$,
and $f(\omega, v) = \omega$ in all other situations.

The formal result: "***common knowledge of Stalnaker rationality does not yield backward induction"*** is correct. However, its interpretation as "***common knowledge of rationality does not yield backward induction"*** is not entirely convincing. Common knowledge of Stalnaker rationality holds at the beginning of the game, but is forfeited after the first move, i.e., behaves as **belief** rather than **knowledge**. Informally, the Stalnaker example is a fit for

'***rationality and common belief of rationality'***

rather than

'***common knowledge of rationality***.'

Common knowledge is too weak

The initial assumption of common knowledge as reachability is too weak in the belief revision models. The selection function that determines the way rationality is calculated does not respect reachability and hence this 'common knowledge' can simply disappear in the process of the game.

What grounds could one find for deriving the backward induction solution if its principal source, common knowledge of rationality at every induction step, is no longer valid?

The Stalnaker theorem states the expected: NONE, and provides an example.

We now consider a belief revision model in which common knowledge of rationality for the remainder of the game is maintained throughout the game. For this we will need a stronger notion of common knowledge.

The Initial Format I

We introduce a notion of *robust knowledge of Stalnaker rationality* in which Stalnaker rationality holds in all relevant situations. This notion captures the essence of belief revision under which knowledge of rationality at a vertex is maintained whenever possible. This can also be considered as a case study which sketches a general framework for different sorts of rationality: (Aumann) rationality is required for some sets X of situations, i.e., pairs *(state,vertex)*, and the choice of X is used to specify the corresponding notion of rationality. In particular,

1. knowledge of rationality in state ω: $X = \{(\omega, v) \mid v \textit{ is a vertex}\}$;
2. common knowledge of rationality in state ω:
   $X = \{(\omega', v) \mid \omega' \textit{ is reachable from } \omega, v \textit{ is a vertex}\}$;
3. Stalnaker rationality in state ω: $X = \{(f(\omega, v), v) \mid v \textit{ is a vertex}\}$.

The Initial Format II

Given an extended model $\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$, a **situation** is a pair (ω, v) consisting of a state ω and vertex v of the game tree. We define a notion of **relevant situation** which reflects our goal to maintain common knowledge of Stalnaker rationality for the remainder of the game at any depth of the belief revision process. The set of situations relevant in (ω, v) is closed under belief revision, epistemic reachability, and advancing to a later moment in the game. **Robust knowledge of Stalnaker rationality** (Definition 4) is defined then via rationality in any relevant situation.

Relevant Situations

A situation (ω', v') is *relevant in* (ω, v), if there is a finite sequence of situations with $m \geq 1$

$$(\omega, v) = (\omega_0, v_0),\ (\omega_1, v_1),\ (\omega_2, v_2),\ \ldots, (\omega_m, v_m) = (\omega', v')$$

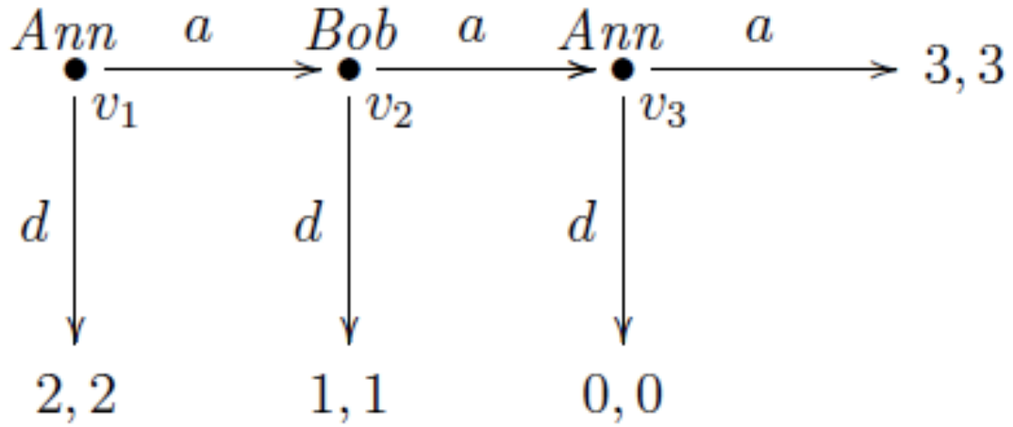
such that for each $k = 0, \ldots, m-1$,

1. $v_k \preceq v_{k+1}$, i.e., v_{k+1} is a future node with respect to v_k;
2. $\omega_{k+1} = f(\widetilde{\omega_k}, v_{k+1})$ for some $\widetilde{\omega_k}$ reachable from ω_k.

It is easy to see that to get from (ω_k, v_k) to (ω_{k+1}, v_{k+1}), one has to pick a state $\widetilde{\omega_k}$ reachable from ω_k (e.g., $\widetilde{\omega_k} = \omega_k$) and a future vertex v_{k+1} (e.g., $v_{k+1} = v_k$), and advance to the revised state $f(\widetilde{\omega_k}, v_k)$. Iteration of this procedure generates all relevant situations.

Relevant Situations: example

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,
$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$
$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

Example 1 In model $\mathcal{A}$, the set U of situations relevant in (ω_1, v_3) is $U = \{(\omega_4, v_3)\}$. The set V of situations relevant in (ω_1, v_2) is $V = U \cup \{(\omega_2, v_2), (\omega_3, v_2), (\omega_5, v_3)\}$. The set W of situations relevant in (ω_1, v_1) is $W = V \cup \{(\omega_1, v_1)\}$. Intuitively, Stalnaker rationality in state ω_1 is determined by (Aumann) rationality in five situations from W.

Robust Knowledge of Rationality

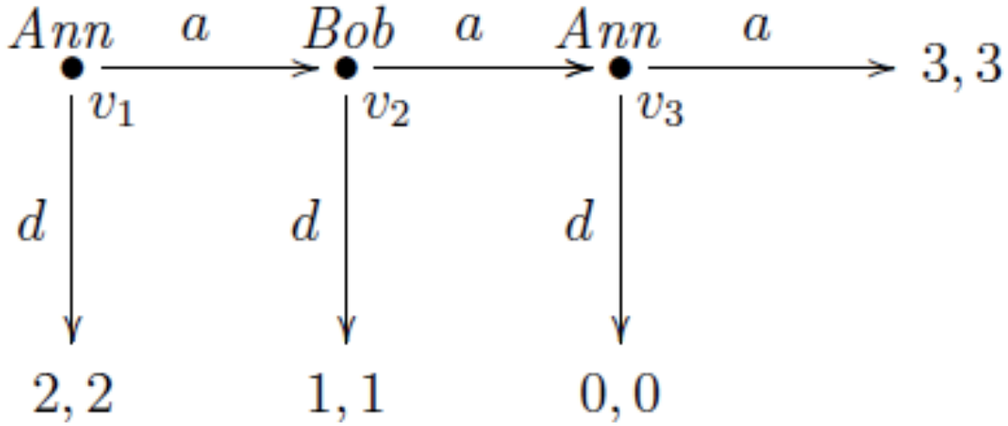
Definition 4 *Robust knowledge of Stalnaker rationality in state ω at vertex v means that in any situation (ω', v') relevant in (ω, v), player $P(v')$ is rational in ω' at v'. Robust knowledge of Stalnaker rationality in state ω means robust knowledge of Stalnaker rationality in state ω at v for each vertex v.*

This definition justifies the notion of a 'relevant situation': robust knowledge of Stalnaker rationality guarantees common knowledge of 'Stalnaker rationality is maintained for the remainder of the game' in any relevant situation.

Robust Knowledge of Rationality

Definition 4 *Robust knowledge of Stalnaker rationality in state ω at vertex v means that in any situation (ω', v') relevant in (ω, v), player $P(v')$ is rational in ω' at v'. Robust knowledge of Stalnaker rationality in state ω means robust knowledge of Stalnaker rationality in state ω at v for each vertex v.*

$\mathbf{s}(\omega_1) = dda$, $\mathbf{s}(\omega_2) = ada$, $\mathbf{s}(\omega_3) = add$,
$\mathbf{s}(\omega_4) = aaa$, $\mathbf{s}(\omega_5) = aad$,
$K_{Ann} = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_5\}\}$
$K_{Bob} = \{\{\omega_1\},\{\omega_2,\omega_3\},\{\omega_4\},\{\omega_5\}\}$



$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

Example 2 In model $\mathcal{A}$, Stalnaker rationality is common knowledge in ω_1. However, robust knowledge of Stalnaker rationality does not hold in (ω_1, v_1). Indeed, situation (ω_5, v_3) is relevant in (ω_1, v_1), but Ann is not rational in ω_5 at v_3.

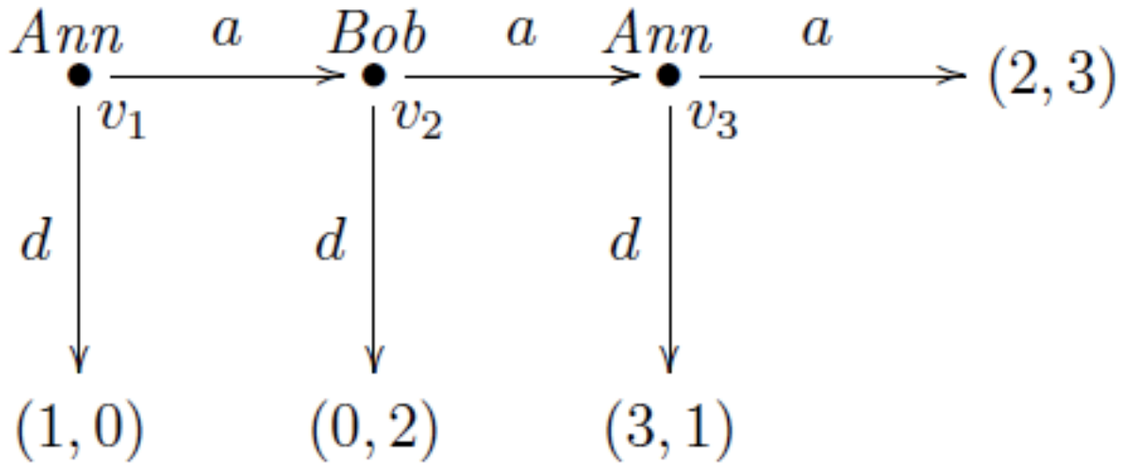
Robust Knowledge and Centipede

$\Omega = \{\omega_1, \omega_2, \omega_3\}$;

$\mathcal{K}_{Ann} = \mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$;

$\mathbf{s}(\omega_1) = (ddd)$, $\mathbf{s}(\omega_2) = (add)$, $\mathbf{s}(\omega_3) = (aad)$;

$f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = f(\omega_2, v_3) = \omega_3$.



Stalnaker rationality does not hold in some situations, e.g., (ω_2, v_1) and (ω_3, v_2). However, such 'bad' situations are irrelevant in 'real' state ω_1 and robust Stalnaker rationality holds in ω_1. Indeed, relevant situations in ω_1 for all possible v's are

$$\{(\omega_1, v_1),\ (\omega_2, v_2),\ (\omega_3, v_3)\},$$

and the corresponding players are rational in all of them.

In Centipede, backward induction survives belief revision given robust knowledge of rationality in the initial state.

Robust Knowledge and BI

Theorem 1 *In extended models over generic game trees, robust knowledge of Stalnaker rationality yields backward induction.*

Proof. Let

$$\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$$

be an extended model such that robust knowledge of Stalnaker rationality holds in state ω of $\mathcal{M}$. This yields that a corresponding player is rational in ω' at v' for each situation (ω', v') relevant in (ω, v_0) where v_0 is the root vertex. We claim that for every relevant situation (ω', v'), restriction of profile $\mathbf{s}(\omega')$ on the subtree Γ below v' coincides with *BI*. Theorem 1 follows from this claim since (ω, v_0) is relevant in itself and the subtree Γ below v_0 is the entire game tree.

Robust Knowledge and BI

To prove the claim, assume the opposite, i.e., that $\mathbf{s}(\omega') \neq BI$ on the subtree Γ below v' for some relevant situation (ω', v'). Let (ω', v') be such a situation with the lowest non-terminal vertex v'. Also, let i be the player making a choice at v'.

Note that $\mathbf{s}(\omega')$ coincides with BI at any vertex v'' strictly below v'. Indeed, situation $(f(\omega', v''), v'')$ for any vertex v'' strictly below v' is relevant, by the definition. By choice of (ω', v'), $\mathbf{s}(f(\omega', v''))$ coincides with BI on v''. By condition F3 on the selection function, $\mathbf{s}(f(\omega', v''))$ agrees with $\mathbf{s}(\omega')$ on v'', hence $\mathbf{s}(\omega')$ coincides with BI on v''.

Then i is not Aumann-rational at v' in ω'. Indeed, the backward induction at v' chooses the best move for i given BI-moves at all other nodes of the subtree Γ below v'. Since the choice of $\mathbf{s}(\omega')$ at v' is different from those of BI and the game tree is generic, it can only be strictly worse. By Definition 1, i is not rational in ω' at v'. $\square$

Discussion

Extended models treat knowledge as defeasible: players revise not only their beliefs in other players' moves but also their 'knowledge' of rationality for the remainder of the game. However, in epistemology, 'knowledge' is usually understood as non-defeasible, and not subject to revision. In this respect, the Stalnaker example reflects the assumption 'rationality and common belief of rationality' rather than 'common knowledge of rationality.'

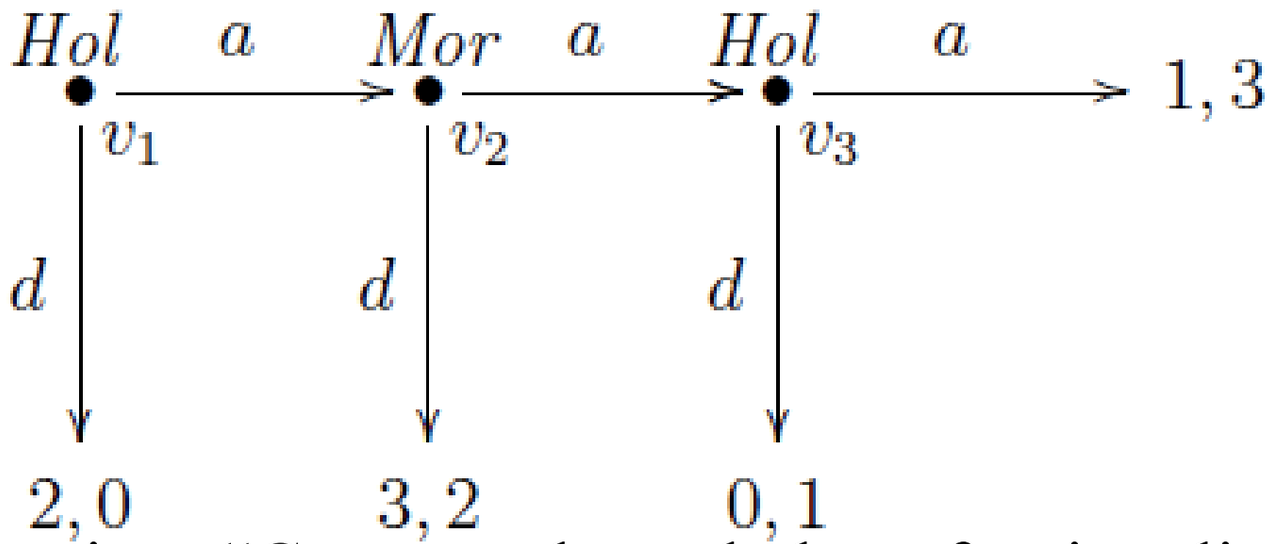
The notion of robust knowledge of Stalnaker rationality reflects the idea of common knowledge of Stalnaker rationality for the remainder of the game at any depth of the belief revision process; it necessarily goes beyond reachability-based common knowledge.

Discussion

For games with a 'small' number of irrational moves, robust knowledge of Stalnaker rationality can be justified by the strong *a priori* rationality reputation of players, their history of rational behavior, etc. An isolated irrational move can be viewed as a technical error. However, trust in rationality fades with each irrational move and given a 'large' number of such moves, robust knowledge of Stalnaker rationality becomes unfeasible. More realistic models of robust rationality should include a bound on the number of errors (e.g., one) allowed for each player.

Vicious Circle of Stalnaker Revisions

Holmes vs Moriarty Game.



Assumption: "Common knowledge of rationality."

Robust players: Backward Induction solution *daa.*

Stalnaker players: ***No commonly known solution.***

Case 1 - *d* is rational. Then Moriarty at v_2 forfeits his belief in Holmes's rationality and plays *down* (by Harsanyi's maximin principle of rationality) which makes Holmes's choice at v_1 not rational.

Case 2 - *a* is rational. Then Moriarty maintains his belief in Holmes's rationality and plays *across,* which renders Holmes's choice at v_1 not rational.

By the spirit, robust rationality models appears more appropriate here.