Ten Puzzles and Paradoxes about Knowledge and Belief

ESSLLI 2013, Düsseldorf

Wes Holliday **Eric Pacuit**

August 14, 2013

Ten Puzzles and Paradoxes

1. Surprise Exam
2. The Knower
3. Logical Omniscience/Knowledge Closure
4. Lottery Paradox & Preface Paradox
5. Margin of Error Paradox
6. Fitch's Paradox
7. Aumann's Agreeing to Disagree Theorem
8. Brandenburger-Keisler Paradox
9. Absent-Minded Driver
10. Backward Induction
11. A puzzle about the sure-thing principle
12. Modeling awareness

Puzzles about *interactive* knowledge and beliefs

$K_i E$: "i knows that E"

$K_i K_j E$: "i knows that j knows that E"

Alternative history...

J. Harsanyi. *Games with incomplete information played by "Bayesian" players I-III. Management Science Theory* **14***: 159-182, 1967-68.*

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

Alternative history...

J. Harsanyi. *Games with incomplete information played by "Bayesian" players I-III. Management Science Theory* **14***: 159-182, 1967-68.*

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

R. Aumann. *Interactive Epistemology I & II*. International Journal of Game Theory (1999).

P. Battigalli and G. Bonanno. *Recent results on belief, knowledge and the epistemic foundations of game theory*. Research in Economics (1999).

R. Myerson. *Harsanyi's Games with Incomplete Information*. Special 50th anniversary issue of *Management Science*, 2004.

Harsanyi Type Space

John C. Harsanyi, nobel prize winner in economics, developed a theory of games with **incomplete information**.

J. Harsanyi. *Games with incomplete information played by "Bayesian" players I-III. Management Science Theory* **14**: *159-182, 1967-68.*

Harsanyi Type Space

John C. Harsanyi, nobel prize winner in economics, developed a theory of games with **incomplete information**.

1. incomplete information: uncertainty about the *structure* of the game (outcomes, payoffs, strategy space)
2. imperfect information: uncertainty *within the game* about the previous moves of the players

J. Harsanyi. *Games with incomplete information played by "Bayesian" players I-III. Management Science Theory* **14***: 159-182, 1967-68.*

Harsanyi's Problem

A natural question following any game-theoretic analysis is

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*?

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*? How do we completely specify such a model?

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*? How do we completely specify such a model?

1. Suppose there is a parameter that some player i does not know

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*? How do we completely specify such a model?

1. Suppose there is a parameter that some player i does not know
2. i's uncertainty about the parameter must be included in the model (first-order beliefs)

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*? How do we completely specify such a model?

1. Suppose there is a parameter that some player i does not know
2. i's uncertainty about the parameter must be included in the model (first-order beliefs)
3. this is a new parameter that the other players may not know, so we must specify the players beliefs about this parameter (second-order beliefs)

Harsanyi's Problem

A natural question following any game-theoretic analysis is *how would the players react if some parameters of the model are not known to the players*? How do we completely specify such a model?

1. Suppose there is a parameter that some player i does not know
2. i's uncertainty about the parameter must be included in the model (first-order beliefs)
3. this is a new parameter that the other players may not know, so we must specify the players beliefs about this parameter (second-order beliefs)
4. but this is a new parameter, and so on....

Harsanyi's Problem

A (game-theoretic) **type** of a player summarizes everything the player knows privately at the beginning of the game which could affect his beliefs about payoffs in the game and about all other players' types.

(Harsanyi argued that all uncertainty in a game can be equivalently modeled as uncertainty about payoff functions.)

Information in games situations

- imperfect information about the play of the game
- incomplete information about the structure of the game

Information in games situations

- imperfect information about the play of the game
- incomplete information about the structure of the game
- strategic information (what will the other players do?)
- higher-order information (what are the other players thinking?)

Epistemic Game Theory

Formally, a game is described by its strategy sets and payoff functions.

Epistemic Game Theory

Formally, a game is described by its strategy sets and payoff functions. But in real life, may other parameters are relevant; there is a lot more going on. Situations that substantively are vastly different may nevertheless correspond to precisely the same strategic game.

Epistemic Game Theory

Formally, a game is described by its strategy sets and payoff functions. But in real life, may other parameters are relevant; there is a lot more going on. Situations that substantively are vastly different may nevertheless correspond to precisely the same strategic game.... The difference lies in the attitudes of the players, in their expectations about each other, in custom, and in history, though the rules of the game do not distinguish between the two situations. (pg. 72)

R. Aumann and J. H. Dreze. *Rational Expectations in Games*. American Economic Review 98 (2008), pp. 72-86.

The Epistemic Program in Game Theory

"...the analysis constitutes a fleshing-out of the textbook interpretation of equilibrium as 'rationality plus correct beliefs.'...this suggests that equilibrium behavior cannot arise out of strategic reasoning alone. "

E. Dekel and M. Siniscalchi. *Epistemic Game Theory*. manuscript, 2013.

A. Brandenburger. *The Power of Paradox*. International Journal of Game Theory, 35, pgs. 465 - 492, 2007.

EP and O. Roy. *Epistemic Game Theory*. Stanford Encyclopedia of Philosophy, forthcoming, 2013.

Doesn't such talk of what Ann believes Bob believes about her, and so on, suggest that some kind of self-reference arises in games, similar to the well-known examples of self-reference in mathematical logic.

A. Brandenburger and H. J. Keisler. *An Impossibility Theorem on Beliefs in Games*. *Studia Logica* (2006).

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false?

$*$ A **strongest belief** is a belief that implies all other beliefs.

A. Brandenburger and H. J. Keisler. *An Impossibility Theorem on Beliefs in Games*. *Studia Logica* (2006).

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose Yes.

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose Yes.

Then, Ann believes that it's not the case that Ann believes that Bob's strongest belief is false.

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose Yes.

Then, Ann believes that it's not the case that Ann believes that Bob's strongest belief is false.

So, it's not the case that Ann believes that Bob's strongest belief is false. $(B\neg B\varphi \rightarrow \neg B\varphi)$

A Paradox

Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose Yes.

Then, Ann believes that it's not the case that Ann believes that Bob's strongest belief is false.

So, it's not the case that Ann believes that Bob's strongest belief is false. ($B\neg B\varphi \to \neg B\varphi$)

So, the answer is no.

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose No.

A Paradox

Ann believes that Bob's strongest belief is
that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose No.

Then, it's not the case that Ann believes it's not the case that Ann believes that Bob's strongest belief is false.

A Paradox

Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose No.

Then, it's not the case that Ann believes it's not the case that Ann believes that Bob's strongest belief is false.

So, Ann believes that Bob's strongest belief is false.
$(\neg B \neg B\varphi \to B\varphi)$

A Paradox

Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.

Does Ann believe that Bob's strongest belief is false? Suppose No.

Then, it's not the case that Ann believes it's not the case that Ann believes that Bob's strongest belief is false.

So, Ann believes that Bob's strongest belief is false.
$(\neg B \neg B\varphi \rightarrow B\varphi)$

So, the answer must be yes.

- strongest belief

- strongest belief
- weakest belief

- strongest belief
- weakest belief
- craziest belief

- strongest belief
- weakest belief
- craziest belief
- all of Bob's belief

Is there a space of all possible interactive beliefs of a game?

Is there a space of all possible interactive beliefs of a game?

Two questions

Is there a space of all possible interactive beliefs of a game?

Two questions

- What exactly does "all possible" mean?

Is there a space of all possible interactive beliefs of a game?

Two questions

- What exactly does "all possible" mean?
  (Complete, Canonical, Universal)

Is there a space of all possible interactive beliefs of a game?

Two questions

- What exactly does "all possible" mean?
(Complete, Canonical, Universal)
- Who cares?

Who Cares?

A. Brandenburger and E. Dekel. *Hierarchies of Beliefs and Common Knowledge*. Journal of Economic Theory (1993).

A. Heifetz and D. Samet. *Knoweldge Spaces with Arbitrarily High Rank*. Games and Economic Behavior (1998).

L. Moss and I. Viglizzo. *Harsanyi type spaces and final coalgebras constructed from satisfied theories*. EN in Theoretical Computer Science (2004).

A. Friendenberg. *When do type structures contain all hierarchies of beliefs?*. working paper (2007).

Who cares?

We think of a particular incomplete structure as giving the "context" in which the game is played.

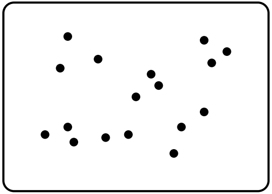
Who cares?

We think of a particular incomplete structure as giving the "context" in which the game is played. In line with Savage's Small-Worlds idea in decision theory [...], who the players are in the given game can be seen as a shorthand for their experiences before the game.

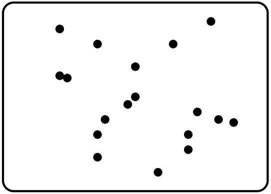
Who cares?

> *We think of a particular incomplete structure as giving the "context" in which the game is played. In line with Savage's Small-Worlds idea in decision theory [...], who the players are in the given game can be seen as a shorthand for their experiences before the game. The players' possible characteristics — including their possible types — then reflect the prior history or context. (Seen in this light, complete structures represent a special "context-free" case, in which there has been no narrowing down of types.) (pg. 319)*

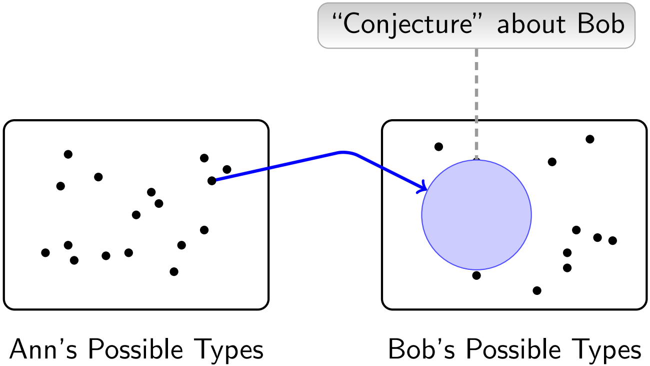
A. Brandenburger, A. Friedenberg, H. J. Keisler. *Admissibility in Games*. Econometrica (2008).

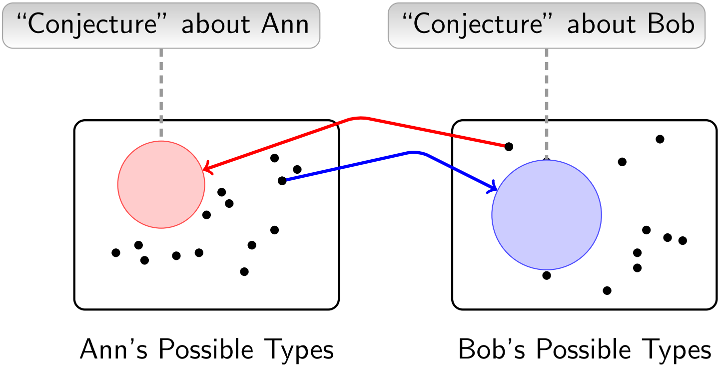


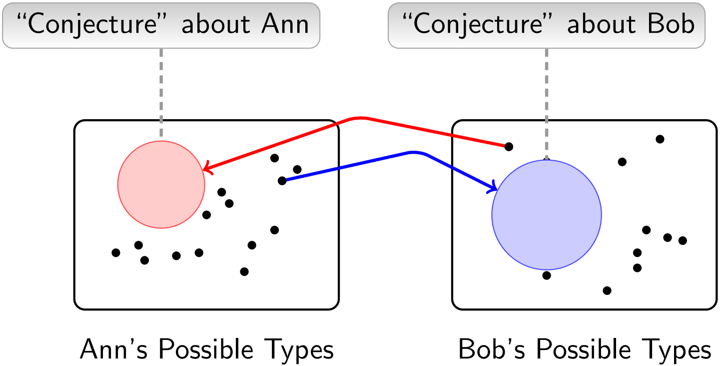
Ann's Possible Types



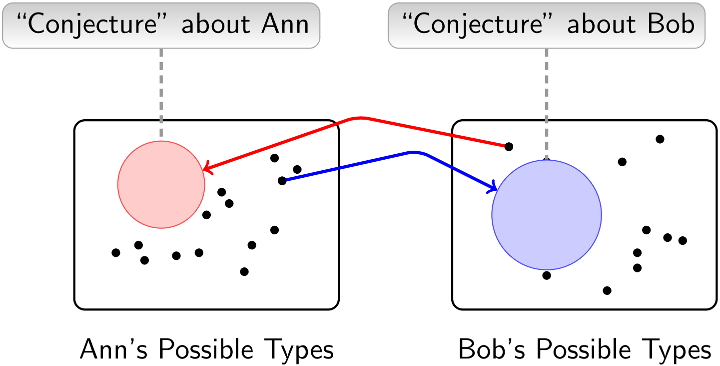
Bob's Possible Types


"Conjecture" about Bob
Ann's Possible Types
Bob's Possible Types


"Conjecture" about Ann
"Conjecture" about Bob
Ann's Possible Types
Bob's Possible Types



Is there a space where every *possible* conjecture is considered by *some* type?



Is there a space where every *possible* conjecture is considered by *some* type? It depends...

S. Abramsky and J. Zvesper. *From Lawvere to Brandenburger-Keisler: interactive forms of diagonalization and self-reference*. Proceedings of LOFT 2010.

EP. *Understanding the Brandenburger Keisler Pardox*. Studia Logica (2007).

Impossibility Results

Language: the (formal) language used by the players to formulate conjectures about their opponents.

Impossibility Results

Language: the (formal) language used by the players to formulate conjectures about their opponents.

Completeness: A model is **complete for a language** if every (consistent) statement in a player's language about an opponent is *considered* by some type.

Qualitative Type Spaces: $\langle T_a, T_b, \lambda_a, \lambda_b \rangle$

$\lambda_a : T_a \to \wp(T_b)$
$\lambda_b : T_b \to \wp(T_a)$

Qualitative Type Spaces: $\langle T_a, T_b, \lambda_a, \lambda_b \rangle$

$\lambda_a : T_a \rightarrow \wp(T_b)$
$\lambda_b : T_b \rightarrow \wp(T_a)$

x **believes** a set $Y \subseteq T_b$ if $\lambda_a(x) \subseteq Y$

x **assumes** a set $Y \subseteq T_b$ if $\lambda_a(x) = Y$

Impossibility Results

Impossibility 1 There is no complete interactive belief structure for the *powerset language*.

Proof. Cantor: there is no onto map from X to the nonempty subsets of X.

Impossibility Results

Impossibility 1 There is no complete interactive belief structure for the *powerset language*.

Proof. Cantor: there is no onto map from X to the nonempty subsets of X.

Impossibility 2 (Brandenburger and Keisler) There is no complete interactive belief structure for *first-order logic*.

Suppose that $\mathcal{C}_A \subseteq \wp(T_A)$ is a set of *conjectures* about Ann and $\mathcal{C}_B \subseteq \wp(T_B)$ a set of conjectures about Bob states.

Assume For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: "in state x_0, Ann has consistent beliefs"
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: "in state x_0, Ann believes that Bob's strongest belief is that X"

Lemma. Under the above assumption, for each $X \in \mathcal{C}_A$ there is an x_0 such that

$$x_0 \in X \text{ iff there is a } y \in T_B \text{ such that } y \in \lambda_A(x_0) \text{ and } x_0 \in \lambda_B(y)$$

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$. By 2., we have $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in X = \lambda_B(y_0)$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$. By 2., we have $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in X = \lambda_B(y_0)$.

Suppose that there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y_0)$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$. By 2., we have $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in X = \lambda_B(y_0)$.

Suppose that there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y_0)$. By 2., $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$.

Claim. $x_0 \in X$ iff $\exists y \in T_B$, $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Assumption: For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$

Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.

Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$. By 2., we have $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in X = \lambda_B(y_0)$.

Suppose that there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y_0)$. By 2., $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in \lambda_B(y_0) = X$.

Consider a first-order language $\mathcal{L}$ containing binary relational symbols $R_A(x, y)$ and $R_B(x, y)$ defining λ_A and λ_B, respectively.

Consider a first-order language $\mathcal{L}$ containing binary relational symbols $R_A(x, y)$ and $R_B(x, y)$ defining λ_A and λ_B, respectively.

$\mathcal{L}$ is interpreted over qualitative type structures where the interpretation of R_A is $\{(t, s) \mid t \in T_A, s \in T_B, \text{ and } s \in \lambda_A(t)\}$.

Consider a first-order language $\mathcal{L}$ containing binary relational symbols $R_A(x, y)$ and $R_B(x, y)$ defining λ_A and λ_B, respectively.

$\mathcal{L}$ is interpreted over qualitative type structures where the interpretation of R_A is $\{(t, s) \mid t \in T_A, s \in T_B, \text{ and } s \in \lambda_A(t)\}$.

Consider the formula φ in $\mathcal{L}$:

$$\varphi(x) := \exists y(R_A(x, y) \wedge R_B(y, x))$$

Consider a first-order language $\mathcal{L}$ containing binary relational symbols $R_A(x, y)$ and $R_B(x, y)$ defining λ_A and λ_B, respectively.

$\mathcal{L}$ is interpreted over qualitative type structures where the interpretation of R_A is $\{(t, s) \mid t \in T_A, s \in T_B, \text{ and } s \in \lambda_A(t)\}$.

Consider the formula φ in $\mathcal{L}$:

$$\varphi(x) := \exists y(R_A(x, y) \wedge R_B(y, x))$$

$\neg\varphi(x) := \forall y(R_A(x, y) \to \neg R_B(y, x))$: "Ann believes that Bob's strongest belief is *false*."

Proof of the Theorem

Suppose that $X \in \mathcal{C}_A$ is defined by the formula
$\neg\varphi(x) := \neg\exists y(R_A(x, y) \wedge R_B(y, x))$.

Proof of the Theorem

Suppose that $X \in \mathcal{C}_A$ is defined by the formula
$\neg\varphi(x) := \neg\exists y(R_A(x,y) \wedge R_B(y,x))$.

There is an $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: Ann's beliefs at x_0 are consistent.
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: At x_0, Ann believes that Bob's strongest belief is that $X = \{x \mid \neg\varphi(x)\}$ (i.e., Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.)

Proof of the Theorem

Suppose that $X \in \mathcal{C}_A$ is defined by the formula $\neg\varphi(x) := \neg\exists y(R_A(x,y) \wedge R_B(y,x))$.

There is an $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: Ann's beliefs at x_0 are consistent.
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: At x_0, Ann believes that Bob's strongest belief is that $X = \{x \mid \neg\varphi(x)\}$ (i.e., Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.)

$\neg\varphi(x_0)$ is true iff (def. of X) $x_0 \in X$

Proof of the Theorem

Suppose that $X \in \mathcal{C}_A$ is defined by the formula $\neg\varphi(x) := \neg\exists y(R_A(x,y) \wedge R_B(y,x))$.

There is an $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: Ann's beliefs at x_0 are consistent.
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: At x_0, Ann believes that Bob's strongest belief is that $X = \{x \mid \neg\varphi(x)\}$ (i.e., Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.)

$\neg\varphi(x_0)$ is true	iff (def. of X)	$x_0 \in X$
	iff (Lemma)	there is a $y \in T_B$ with $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

Proof of the Theorem

Suppose that $X \in \mathcal{C}_A$ is defined by the formula $\neg\varphi(x) := \neg\exists y(R_A(x,y) \wedge R_B(y,x))$.

There is an $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: Ann's beliefs at x_0 are consistent.
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: At x_0, Ann believes that Bob's strongest belief is that $X = \{x \mid \neg\varphi(x)\}$ (i.e., Ann believes that Bob's strongest belief is that Ann believes that Bob's strongest belief is false.)

$\neg\varphi(x_0)$ is true	iff (def. of X)	$x_0 \in X$
	iff (Lemma)	there is a $y \in T_B$ with $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$
	iff (def. of $\varphi(x)$)	$\varphi(x_0)$ is true.

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

"A group of agents cannot agree to disagree"

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

"A group of agents cannot agree to disagree"

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

"A group of agents cannot agree to disagree"

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

"*Common Knowledge*" is informally described as what any fool would know, given a certain situation: It encompasses what is relevant, agreed upon, established by precedent, assumed, being attended to, salient, or in the conversational record.

"*Common Knowledge*" is informally described as what any fool would know, given a certain situation: It encompasses what is relevant, agreed upon, established by precedent, assumed, being attended to, salient, or in the conversational record.

It is not Common Knowledge who "defined" Common Knowledge!

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study*. 1969.

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study*. 1969.

Fixed-point definition: $\gamma :=$ *i* and *j* know that (φ and γ)

G. Harman. *Review of* Linguistic Behavior. Language (1977).

J. Barwise. *Three views of Common Knowledge*. TARK (1987).

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study*. 1969.

Fixed-point definition: $\gamma :=$ i and j know that (φ and γ)

G. Harman. *Review of* Linguistic Behavior. Language (1977).

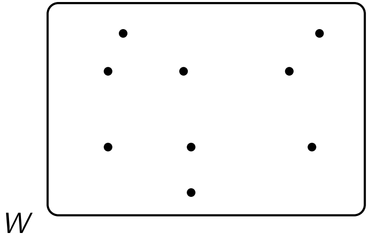
J. Barwise. *Three views of Common Knowledge*. TARK (1987).

Shared situation: There is a *shared situation* s such that (1) s entails φ, (2) s entails everyone knows φ, plus other conditions

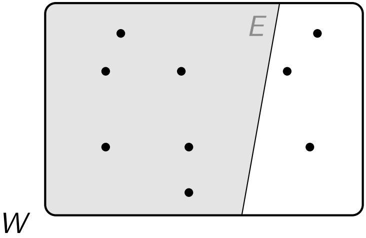
H. Clark and C. Marshall. *Definite Reference and Mutual Knowledge*. 1981.

M. Gilbert. *On Social Facts*. Princeton University Press (1989).

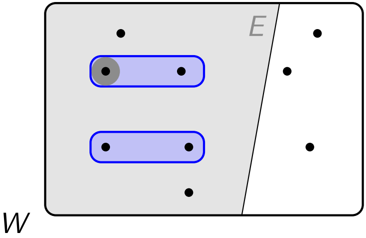
P. Vanderschraaf and G. Sillari. *"Common Knowledge", The Stanford Encyclopedia of Philosophy (2009)*.
http://plato.stanford.edu/entries/common-knowledge/.



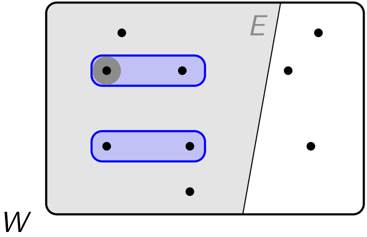
W is a set of **states** or **worlds**.



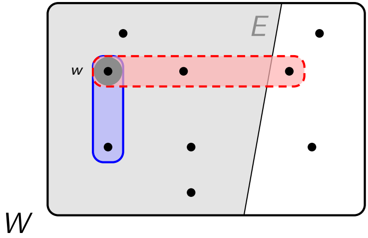
An **event/proposition** is any (definable) subset $E \subseteq W$



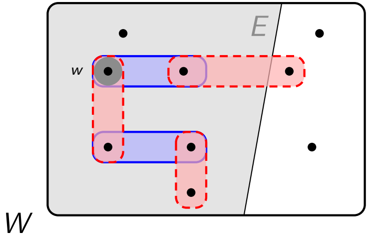
The agents receive signals in each state. States are considered equivalent for the agent if they receive the same signal in both states.



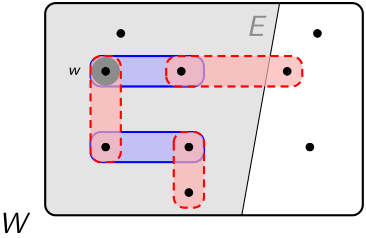
Knowledge Function: $K_i : \wp(W) \to \wp(W)$ where $K_i(E) = \{w \mid R_i(w) \subseteq E\}$



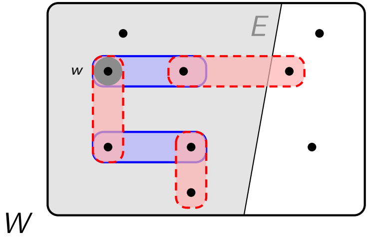
$w \in K_A(E)$ and $w \notin K_B(E)$



The model also describes the agents' **higher-order knowledge/beliefs**

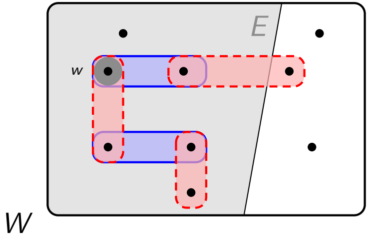


Everyone Knows: $K(E) = \bigcap_{i \in \mathcal{A}} K_i(E)$, $K^0(E) = E$, $K^m(E) = K(K^{m-1}(E))$

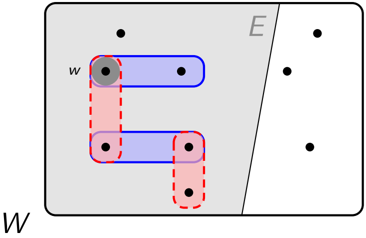


Common Knowledge: $C : \wp(W) \to \wp(W)$ with

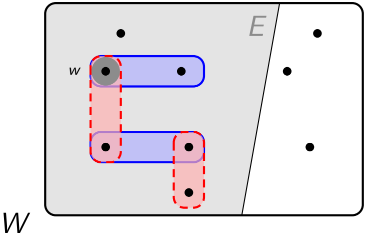
$$C(E) = \bigcap_{m \geq 0} K^m(E)$$



$w \in K(E)$ $\qquad$ $w \notin C(E)$



$w \in C(E)$



Fact. $w \in C(E)$ if every finite path starting at w ends in a state in E

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n + 1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

Do the agents know there numbers are less than 1000?

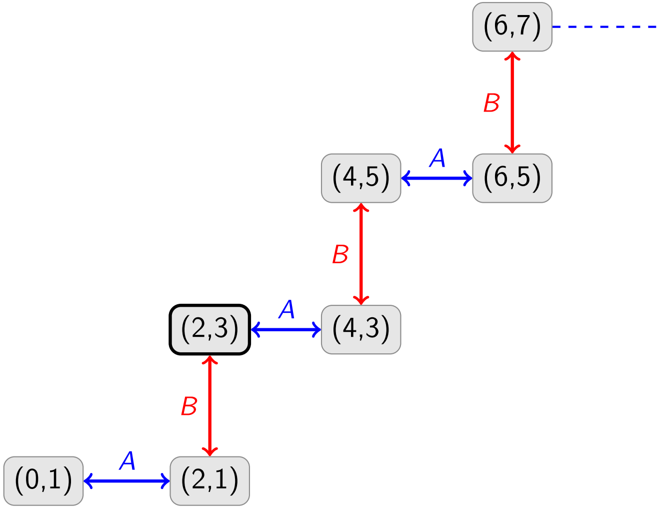
An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

Do the agents know there numbers are less than 1000?

Is it common knowledge that their numbers are less than 1000?


(6,7)
B
(4,5)
A
(6,5)
B
(2,3)
A
(4,3)
B
(0,1)
A
(2,1)

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

S. Morris. *The common prior assumption in economic theory*. Economics and Philosophy, 11, pgs. 227 - 254, 1995.

Generalized Aumann's Theorem

Qualitative versions: *like-minded individuals cannot agree to make different decisions*.

M. Bacharach. *Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge*. Journal of Economic Theory (1985).

J.A.K. Cave. *Learning to Agree*. Economic Letters (1983).

D. Samet. *Agreeing to disagree: The non-probabilistic case*. Games and Economic Behavior, Vol. 69, 2010, 169-174.

The Framework

Knowledge Structure: $\langle W, \{\Pi_i\}_{i \in \mathcal{A}} \rangle$ where each Π_i is a partition on W ($\Pi_i(w)$ is the cell in Π_i containing w).

Decision Function: Let D be a nonempty set of **decisions**. A decision function for $i \in \mathcal{A}$ is a function $\mathbf{d}_i : W \to D$. A vector $\mathbf{d} = (d_1, \ldots, d_n)$ is a decision function profile. Let $[\mathbf{d}_i = d] = \{w \mid \mathbf{d}_i(w) = d\}$.

The Framework

Knowledge Structure: $\langle W, \{\Pi_i\}_{i \in \mathcal{A}} \rangle$ where each Π_i is a partition on W ($\Pi_i(w)$ is the cell in Π_i containing w).

Decision Function: Let D be a nonempty set of **decisions**. A decision function for $i \in \mathcal{A}$ is a function $\mathbf{d}_i : W \to D$. A vector $\mathbf{d} = (d_1, \ldots, d_n)$ is a decision function profile. Let $[\mathbf{d}_i = d] = \{w \mid \mathbf{d}_i(w) = d\}$.

(A1) Each agent knows her own decision:

$$[\mathbf{d}_i = d] \subseteq K_i([\mathbf{d}_i = d])$$

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

$w \in [j \succeq i]$ then j knows at w every event that i knows there.

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

$w \in [j \succeq i]$ then j knows at w every event that i knows there.

$[j \sim i] = [j \succeq i] \cap [i \succeq j]$

The Sure-Thing Principle

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

The Sure-Thing Principle

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

The Sure-Thing Principle

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

The Sure-Thing Principle

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

Interpersonal Sure-Thing Principle (ISTP)

For any pair of agents i and j and decision d,

$$K_i([j \succeq i] \cap [\mathbf{d}_j = d]) \subseteq [\mathbf{d}_i = d]$$

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions. Yet, being the students of the same academy, the method by which they arrive at their conclusions is the same.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions. Yet, being the students of the same academy, the method by which they arrive at their conclusions is the same. Suppose now that detective Bob, a father of four who returns home every day at five oclock, collects all the information about the case at hand together with detective Alice.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is. Suppose that he also knows what Alices decision is.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is. Suppose that he also knows what Alices decision is. Since Alice uses the same investigation method as Bob, he knows that had he been in possession of the more extensive knowledge that Alice has collected, he would have made the same decision as she did. Thus, this is indeed his decision.

Implications of ISTP

Proposition. If the decision function profile $\mathbf{d}$ satisfies ISTP, then

$$[i \sim j] \subseteq \bigcup_{d \in D} ([\mathbf{d}_i = d] \cap [\mathbf{d}_j = d])$$

ISTP Expandability

Agent i is an **epistemic dummy** if it is always the case that all the agents are at least as knowledgeable as i. That is, for each agent j,

$$[j \succeq i] = W$$

A decision function profile $\mathbf{d}$ on $\langle W, \Pi_1, \dots, \Pi_n \rangle$ is **ISTP expandable** if for any expanded structure $\langle W, \Pi_1, \dots, \Pi_{n+1} \rangle$ where $n+1$ is an epistemic dummy, there exists a decision function $\mathbf{d}_{n+1}$ such that $(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{n+1})$ satisfies ISTP.

ISTP Expandability: Illustration

Suppose that after making their decisions, Alice and Bob are told that another detective, one E.P. Dummy, who graduated the very same police academy, had also been assigned to investigate the same case.

ISTP Expandability: Illustration

Suppose that after making their decisions, Alice and Bob are told that another detective, one E.P. Dummy, who graduated the very same police academy, had also been assigned to investigate the same case. In principle, they would need to review their decisions in light of the third detectives knowledge: knowing what they know about the third detective, his usual sources of information, for example, may impinge upon their decision.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy. The news that he had been assigned to the same case is completely irrelevant to the conclusions that Alice and Bob have reached. Obviously, based on the information he gets from the media, Dummy also makes a decision.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy. The news that he had been assigned to the same case is completely irrelevant to the conclusions that Alice and Bob have reached. Obviously, based on the information he gets from the media, Dummy also makes a decision. We may assume that the decisions made by the three detectives satisfy the ISTP, for exactly the same reason we assumed it for the two detectives decisions

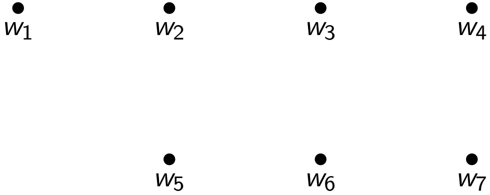
Generalized Agreement Theorem

If $\mathbf{d}$ is an ISTP expandable decision function profile on a partition structure $\langle W, \Pi_1, \ldots, \Pi_n \rangle$, then for any decisions $d_1, \ldots, d_n$ which are not identical, $C(\bigcap_i [\mathbf{d}_i = d_i]) = \emptyset$.

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

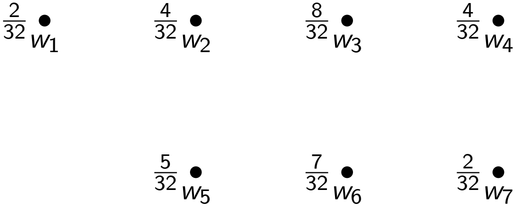
Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

2 Scientists Perform an Experiment



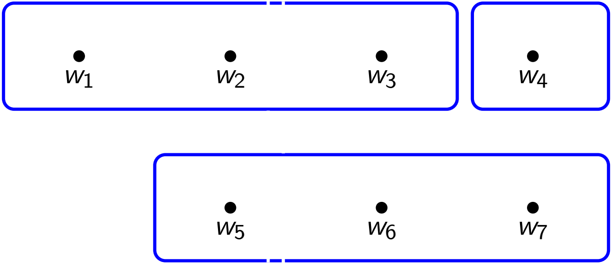
They agree the true state is one of seven different states.

2 Scientists Perform an Experiment



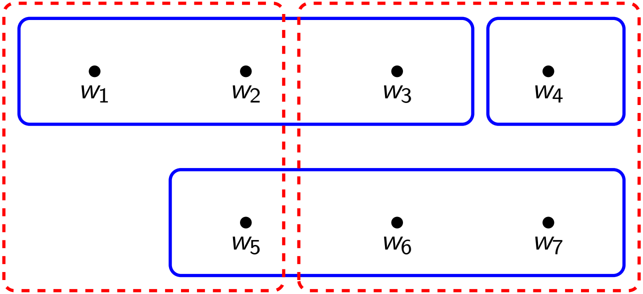
They agree on a common prior.

2 Scientists Perform an Experiment



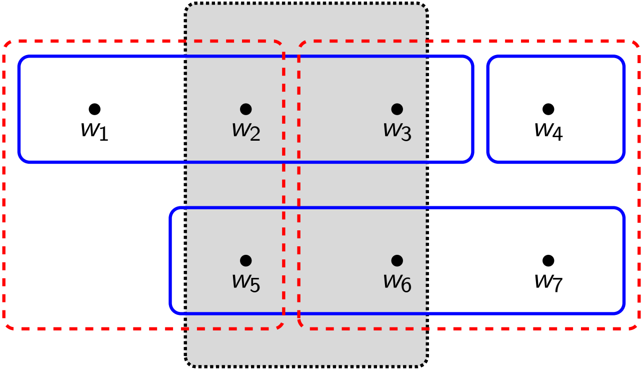
They agree that Experiment 1 would produce the blue partition.

2 Scientists Perform an Experiment



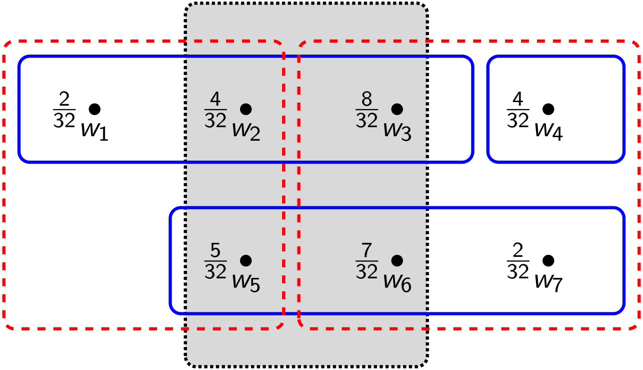
They agree that Experiment 1 would produce the blue partition and Experiment 2 the red partition.

2 Scientists Perform an Experiment



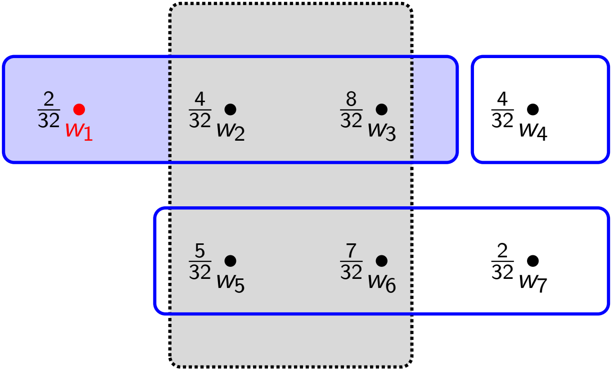
They are interested in the truth of $E = \{w_2, w_3, w_5, w_6\}$.

2 Scientists Perform an Experiment



So, they agree that $P(E) = \frac{24}{32}$.

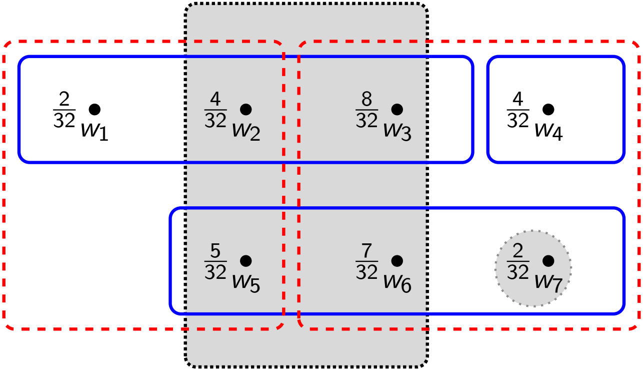
2 Scientists Perform an Experiment



Also, that *if the true state is* w_1, then Experiment 1 will yield

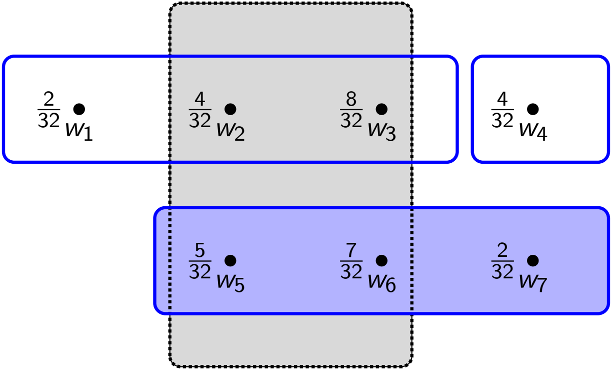
$$P(E|I) = \frac{P(E \cap I)}{P(I)} = \frac{12}{14}$$

2 Scientists Perform an Experiment



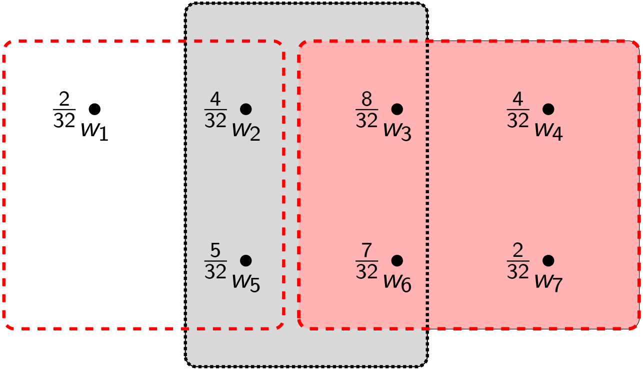
Suppose the true state is w_7 and the agents preform the experiments.

2 Scientists Perform an Experiment



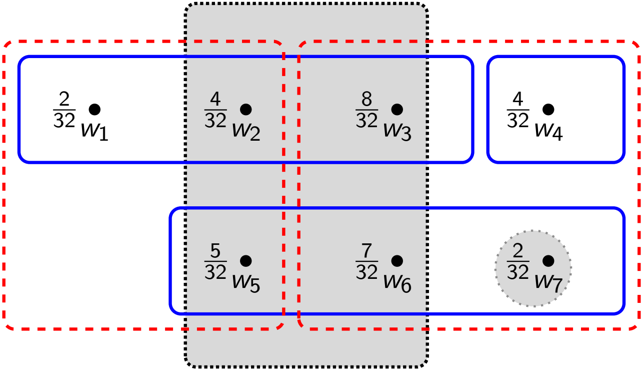
Suppose the true state is w_7, then $Pr_1(E) = \frac{12}{14}$

2 Scientists Perform an Experiment



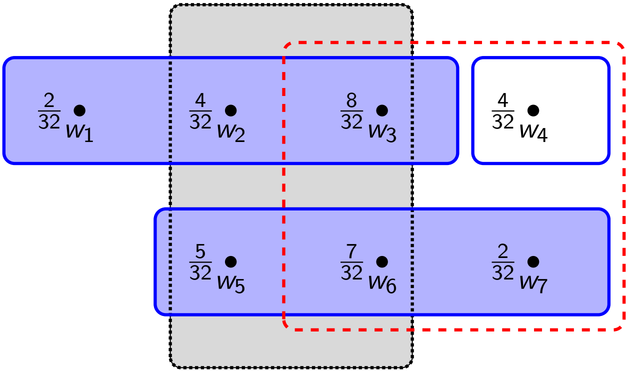
Then $Pr_1(E) = \frac{12}{14}$ and $Pr_2(E) = \frac{15}{21}$

2 Scientists Perform an Experiment



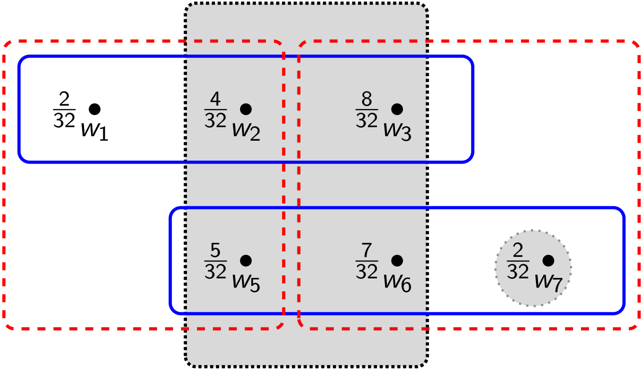
Suppose they exchange emails with the new subjective probabilities: $Pr_1(E) = \frac{12}{14}$ and $Pr_2(E) = \frac{15}{21}$

2 Scientists Perform an Experiment



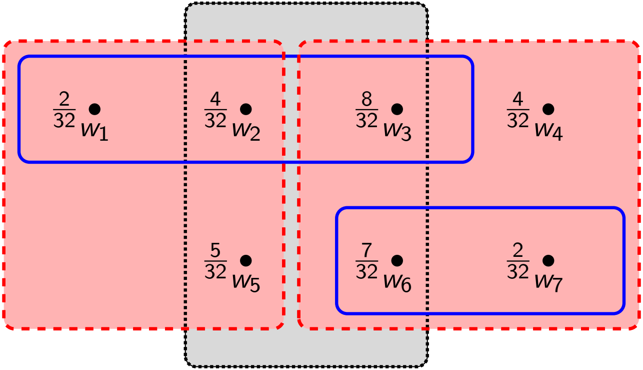
Agent 2 learns that w_4 is **NOT** the true state (same for Agent 1).

2 Scientists Perform an Experiment



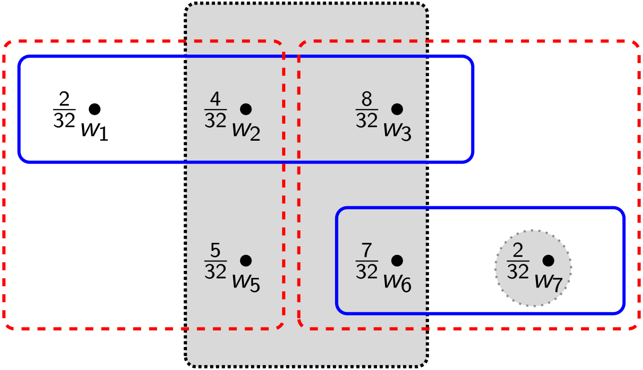
Agent 2 learns that w_4 is **NOT** the true state (same for Agent 1).

2 Scientists Perform an Experiment



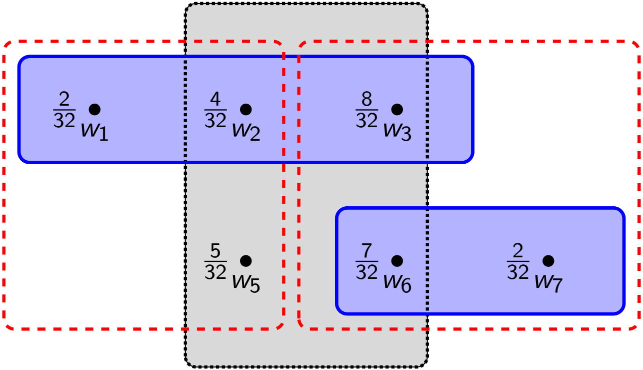
Agent 1 learns that w_5 is **NOT** the true state (same for Agent 1).

2 Scientists Perform an Experiment



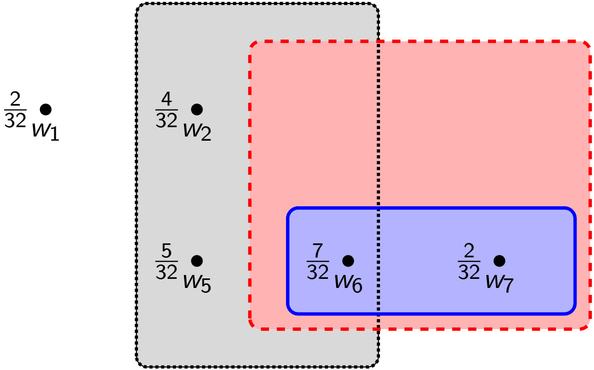
The new probabilities are $Pr_1(E|I') = \frac{7}{9}$ and $Pr_2(E|I') = \frac{15}{17}$

2 Scientists Perform an Experiment



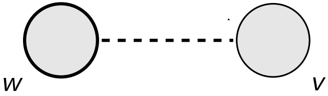
After exchanging this information ($Pr_1(E|I') = \frac{7}{9}$ and $Pr_2(E|I') = \frac{15}{17}$), Agent 2 learns that w_3 is **NOT** the true state.

2 Scientists Perform an Experiment



No more revisions are possible and the agents agree on the posterior probabilities.

Models of Hard and Soft Information

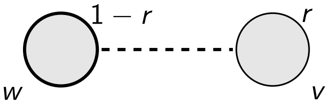


$\mathcal{M} = \langle W, \{\Pi_i\}_{i \in \mathcal{A}} \rangle$

Π_i is agent i's partition with $\Pi_i(w)$ the partition cell containing w.

$K_i(E) = \{w \mid \Pi_i(w) \subseteq E\}$

Models of Hard and Soft Information



$\mathcal{M} = \langle W, \{\Pi_i\}_{i\in\mathcal{A}}, \{p_i\}_{i\in\mathcal{A}}\rangle$
for each i, $p_i : W \to [0,1]$ is a probability measure

$$B^p(E) = \{w \mid p_i(E \mid \Pi_i(w)) = \frac{\pi_i(E \cap \Pi_i(w))}{p_i(\Pi_i(w))} \geq p\}$$

1. $B_i^p(B_i^p(E)) = B_i^p(E)$

2. If $E \subseteq F$ then $B_i^p(E) \subseteq B_i^p(F)$

3. $\pi(E \mid B_i^p(E)) \geq p$

Common p-belief

The typical example of an event that creates common knowledge is a **public announcement**.

Common *p*-belief

The typical example of an event that creates common knowledge is a **public announcement**.

Shouldn't one always allow for some small probability that a participant was absentminded, not listening, sending a text, checking facebook, proving a theorem, asleep, ...

Common p-belief

The typical example of an event that creates common knowledge is a **public announcement**.

Shouldn't one always allow for some small probability that a participant was absentminded, not listening, sending a text, checking facebook, proving a theorem, asleep, ...

"We show that the weaker concept of "common belief" can function successfully as a substitute for common knowledge in the theory of equilibrium of Bayesian games."

D. Monderer and D. Samet. *Approximating Common Knowledge with Common Beliefs*. Games and Economic Behavior (1989).

Common p-belief: definition

$$B_i^p(E) = \{w \mid p(E \mid R_i(w)) \geq p\}$$

Common p-belief: definition

$$B_i^p(E) = \{w \mid p(E \mid R_i(w)) \geq p\}$$

An event E is **evident p-belief** if for each $i \in \mathcal{A}$, $E \subseteq B_i^p(E)$

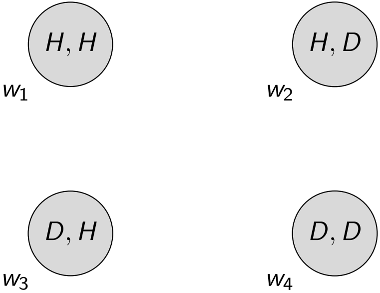
Common p-belief: definition

$$B_i^p(E) = \{w \mid p(E \mid R_i(w)) \geq p\}$$

An event E is **evident p-belief** if for each $i \in \mathcal{A}$, $E \subseteq B_i^p(E)$

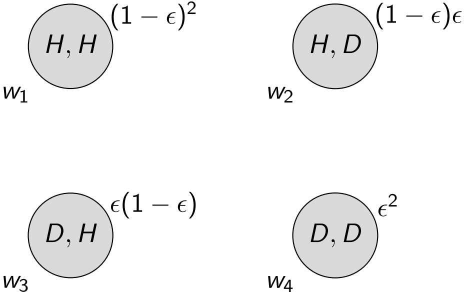
An event F is **common p-belief** at w if there exists and evident p-belief event E such that $w \in E$ and for all $i \in \mathcal{A}$, $E \subseteq B_i^p(F)$

Common p-belief: example



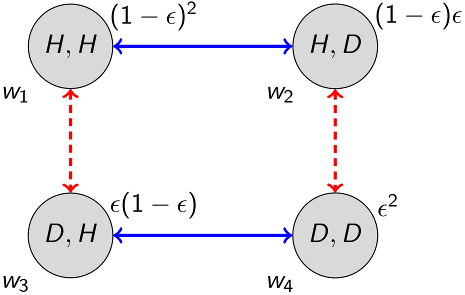
Two agents either hear (H) or don't hear (D) the announcement.

Common p-belief: example



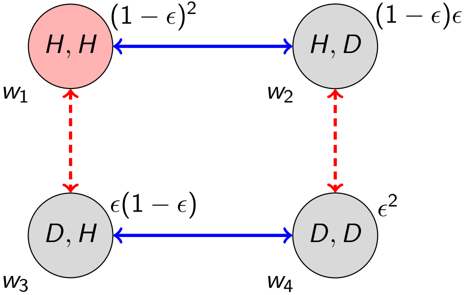
The probability that an agent hears is $1-\epsilon$.

Common p-belief: example



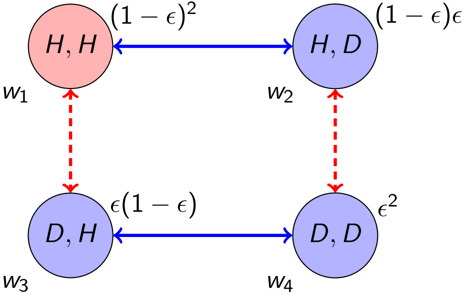
The agents *know* their "type".

Common p-belief: example



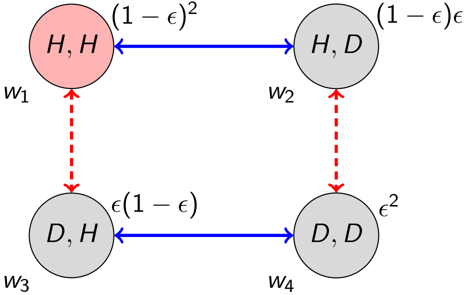
The event "everyone hears" ($E = \{w_1\}$)

Common p-belief: example



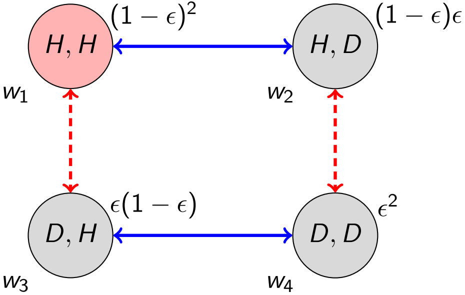
The event "everyone hears" ($E = \{w_1\}$) is **not** common knowledge

Common p-belief: example



The event "everyone hears" ($E = \{w_1\}$) is **not** common knowledge, but it is common $(1-\epsilon)$-belief

Common p-belief: example



The event "everyone hears" ($E = \{w_1\}$) is **not** common knowledge, but it is common $(1-\epsilon)$-belief:
$B_i^{(1-\epsilon)}(E) = \{w \mid p(E \mid \Pi_i(w)) \geq 1-\epsilon\} = \{w_1\} = E$, for $i = 1, 2$

Common p-belief

Theorem. If the posteriors of an event X are common p-belief at some state w, then any two posteriors can differ by at most $2(1 - p)$.

D. Samet and D. Monderer. *Approximating Common Knowledge with Common Beliefs*. Games and Economic Behavior, Vol. 1, No. 2, 1989.