COMMON KNOWLEDGE AS A COINDUCTIVE MODALITY

VENANZIO CAPRETTA

Institute for Computing and Information Science (ICIS), Radboud University Nijmegen
e-mail address: venanzio@cs.ru.nl

Abstract. I prove in Coq Aumann's Theorem: In perfect information games, common knowledge of rationality implies backward induction equilibrium. The notion of common knowledge is formalized, using a coinductive definition, as a modality containing an infinite amount of information.

1. The Green-Eyed Amazons

In the land of the Amazons a strange taboo was enforced.[1] None of them was allowed to know the colour of her own eyes. Everybody could see the colour of the eyes of the others, but there were no mirrors or other reflective surfaces on which to check one's own appearance. It was also strictly forbidden to ask somebody else to reveal the colour of one's eyes.

The ironic fact is that every one of them had green eyes. Each of them could see that all her companions were green-eyed, but she was in doubt about herself.

Another peculiarity of the Amazons was that they were all proficient logicians: They could make the most complex reasoning without error.

One day the goddess Artemis came to the village and announced: 'I decided to give a golden bow to every one of you who has green eyes. To apply for the prize you will need to paint your front door green before midnight of any day. Then, the following morning, I will come to check whether your eyes are indeed green. If they are, you will receive the price. But be warned, there will be no tolerance for cheaters: If your eyes turn out to be of a different colour, I will strike you down with one of my arrows. I am quite sure that I didn't come here in vain, because I can see that at least one of you has green eyes.'

My question to you is: Can you predict how the Amazons behaved. Did any of them paint her door green, and when did they do it? The solution is revealed at the end of the article.

[1]This is a variant of the cheating wives [5] (or husbands [6]) puzzle, also known as the muddy children puzzle [2].

REFLECTIONS ON TYPE THEORY, λ-CALCULUS, AND THE MIND
Essays Dedicated to Henk Barendregt on the Occasion of his 60th Birthday

2. Introduction

Epistemic logic is the discipline that studies reasoning about knowledge. One of its goals is to identify correct derivation rules when we deal with formulas of the kind: *'X knows that ...'*. There are several good introduction to the subject, for example [4] and [7].

Robert Aumann applied epistemic logic to game theory [1] to model the knowledge that a player may have of other players' strategies. We assume that a group of agents is playing a perfect information game. This is a game in which the present position is completely known to all players. Chess is a perfect information game while poker isn't, because each player's cards are unknown to the others.

Each agent is rational, which means that she plays in order to maximize her payoff: she never knowingly chooses an action that would give her a worse outcome. The rationality hypothesis is not enough to determine the strategies of the agents, since, without knowing anything about the other agents actions, an agent cannot determine an optimal behavior. However, if every player, beside being rational, knows of the other players rationality, and if this knowledge is common, then Aumann's Theorem tells us that the chosen strategies satisfy a particular form of equilibrium called *Backward Induction*.

My goal is to formalize this result in Coq [8, 3]. There was a previous attempt at the formalization by Vestergaard, Lescanne, and Ono [9]. Rather than using Coq's native logic, they adopted a proof-theoretic method: they used Coq as a metalogic and formalized their own formal system to capture epistemic logic as applied to game theory. The question then arises of whether their object logic is adequate and models appropriately the desired notions. The fact that their result is much stronger that Aumann's original statement leads to the suspicion that they assumed axioms that are too strong. In fact, they come to the conclusion that only rationality, without the need of common knowledge of it, is sufficient to determine Backward Induction. This is patently false in the intended interpretation.

To avoid such problems I decided to formalize directly all the notions from game theory and epistemic logic in Coq. This leads to a Coq development more complex than the one in [9], but with the added certainty that one has to trust only Coq logic and not an ad hoc formal system. The Coq file is available at: `http://www.cs.ru.nl/~venanzio/aumann.v`.

3. Epistemic Logic and Common Knowledge

We follow Aumann in the definition of Common Knowledge. We assume that the possible situations of the world on which the actions of the players may depend is modeled by a set State of states. A state $w \in \mathsf{State}$ completely determines every aspect of the world. Let us take as example the Amazon story given at the beginning. Suppose there are just three Amazons, called Hippolyta, Antiope, and Melanippe. A state w specifies the eye colour of each of them, but also whether they paint their door and on what day.

w_1: Hippolyta has green eyes and paints her door on the second day;
Antiope has blue eyes and doesn't paint her door;
Melanippe has brown eyes and paints her door on the first day;
...

w_2: Hippolyta has brown eyes and paints her door on the third day;
Antiope has green eyes and paints her door on the second day;
Melanippe has green eyes and paints her door on the first day;
...

A state will possibly contain an infinite amount of information: all that is needed to determine completely the world. In practice we restrict its content to what is relevant to the problem at hand.

A statement, for example *'There are two green-eyed Amazons'*, is seen as a property that states may or may not satisfy, that is, a predicate on states. We call such a predicate an *event*: $\mathsf{Event} := \mathsf{State} \to \mathsf{Prop}$. If E_1 is the event stating that there are two green-eyed Amazons, then $E_1\, w_1$ is false and $E_1\, w_2$ is true. If E_2 is the statement *'Melanippe paints her door on the first day'*, then both $E_2\, w_1$ and $E_2\, w_2$ are true.

We introduce some notation for the logic of events in this possible worlds model. (Note that this is quite different from Kripke models, where states represent stages of knowledge). Connectives for events are defined as follows:

$$\begin{aligned} E_1 \sqsubset E_2 &:= \lambda w.(E_1\, w) \to (E_2\, w) \\ E_1 \sqcap E_2 &:= \lambda w.(E_1\, w) \wedge (E_2\, w) \\ E_1 \sqcup E_2 &:= \lambda w.(E_1\, w) \vee (E_2\, w) \\ \neg\, E &:= \lambda w. \sim (E\, w). \end{aligned}$$

Besides, we also need a notation for logical implication and equivalence:

$$\begin{aligned} E_1 \subset E_2 &:= \forall w.(E_1\, w) \to (E_2\, w) \\ E_1 \equiv E_2 &:= (E_1 \subset E_2) \wedge (E_2 \subset E_1). \end{aligned}$$

Using the necessity modality (which we don't really need in this work) we can define logical implication as necessary implication:

$$\begin{aligned} \Box\, E &:= \forall w.E\, w \\ (E_1 \subset E_2) &\Leftrightarrow \Box(E_1 \sqsubset E_2). \end{aligned}$$

We model knowledge in the following way. In a certain state w, a player i will know something about the world. This means that i will know some of the events that are true in w and will not know others. Knowledge, like any other aspect of the world, is determined by the state. So there may be states that differ only for the knowledge that an agent has. Refining the examples above, we could have:

w_{11}: Hippolyta has green eyes and paints her door on the second day;
Antiope has blue eyes and doesn't paint her door;
Melanippe has brown eyes and paints her door on the first day;
Hippolyta knows that Melanippe has brown eyes;
Melanippe knows that Antiope doesn't paint her door;
...

w_{12}: Hippolyta has green eyes and paints her door on the second day;
Antiope has blue eyes and doesn't paint her door;
Melanippe has brown eyes and paints her door on the first day;
Antiope knows that Melanippe paints her door on the first day;
Melanippe knows that Antiope doesn't have green eyes;
...

w_{21}: Hippolyta has brown eyes and paints her door on the third day;
Antiope has green eyes and paints her door on the second day;
Melanippe has green eyes and paints her door on the first day;
Hippolyta knows that there are two green-eyed Amazons;
Antiope knows that Melanippe paints her door on the first day;
...

w_{22}: Hippolyta has brown eyes and paints her door on the third day;
Antiope has green eyes and paints her door on the second day;
Melanippe has green eyes and paints her door on the first day;
Hippolyta knows that she doesn't paint her door in the first two days;
Melanippe knows that her own eyes are green;
...

Our notion of knowledge is strong in two ways. Knowledge is always true: If, in a state w, an agent knows an event E, then E must be true in w. (If this doesn't happen we talk about *belief* rather than knowledge.) We also assume that every agent is aware of her own knowledge: If, in state w, an agent i knows E, then i also knows that i knows i. For example, in state w_{11}, Hippolyta knows that she knows that Melanippe has brown eyes.

For every agent i and state w, we have therefore a predicate on events: For every event E, $\mathsf{Knows}\, i\, w\, E$ is true if in state w, i knows E. However, we don't take the predicate Knows as primitive in our formalization but, following Aumann, adopt a more extensional viewpoint. Agents are assumed to be proficient logicians: If an event E_1 logically implies an event E_2, and i knows E_1, then i also knows E_2. This means that there is no need to explicitly state that i knows E_2. Indeed, it is sufficient to state that i knows a single event: The intersection of all the events that i knows. The rest of i's knowledge will follow by implication. We call this single event $(\mathcal{K}_i \bullet w) : \mathsf{Event}$. Remembering that $\mathsf{Event} = \mathsf{State} \to \mathsf{Prop}$, we have that $\mathcal{K} : \mathsf{Player} \to \mathsf{State} \to \mathsf{State} \to \mathsf{Prop}$. So $\mathcal{K}_i$ is a relation on states, for every agent i.

We use $\mathcal{K}$ as the primitive notion to model knowledge. Here is another way to grasp it intuitively. Suppose the world is in state w. The agent i does not know exactly in what state the world is. She will know some facts but not others. This means that there are other states, let v be one, that i considers possible: All the facts that i knows about the actual state w are also true in v; in other words, i may well think that the world is in state v instead of w. This is the situation that we describe by $\mathcal{K}_i\, v\, w$. For example, in state w_{12}, all that Antiope knows is that Melanippe paints her door on the first day. Since this event occurs also in state w_{21}, Antiope cannot tell whether the world is in state w_{12} or w_{21}: Therefore $\mathcal{K}_{\text{Antiope}}\, w_{12}\, w_{21}$ is true.

The fact that knowledge is always true amounts to $\mathcal{K}_i$ being reflexive: i's knowledge must be compatible with the actual state. We stated that an agent is always aware of her own knowledge. In other words, i cannot think that she may know more or less that what she actually knows. Therefore, if $\mathcal{K}_i\, v\, w$ holds, the knowledge that i has in v is equal to the knowledge that she has in w: $\mathcal{K}_i\, v = \mathcal{K}_i\, w$. From this it is immediate to prove that $\mathcal{K}_i$ is symmetric and transitive.

In conclusion, a complete description of the knowledge of all players is given by a *knowledge system*: $\mathcal{K} : \mathsf{Player} \to \mathsf{Equivalence\ State}$, mapping every player to an equivalence relation on states.

As we said, $\lambda v.\mathcal{K}_i\, v\, w$ (written as $\mathcal{K}_i \bullet w$ above) is the least event that i knows in state w; it implies all other events known to i. We can therefore define:

$$\begin{aligned}&\mathsf{Knows} : \mathsf{Player} \to \mathsf{Event} \to \mathsf{Event}\\ &:= \lambda i E w.\forall v : \mathsf{State}, \mathcal{K}_i\, v\, w \to E\, v.\end{aligned}$$

If E is an event, then the event *'Player i knows E'*, formally $\mathsf{Knows}\, i\, E$, consists of the union of the equivalence classes of $\mathcal{K}_i$ that are contained in E. In fact, in a state w, we can

say that i knows E if E is true and the information available to i determines E; that is, all states in which i thinks the world may be in, satisfy E.

We define an operator K on events such that $\mathsf{K}\,E$ states that all players know the event E: $\mathsf{K} := \lambda E w.\forall i, \mathsf{Knows}\, i\, E\, w$.

Finally, common knowledge of an event E states that everybody knows E, everybody knows that everybody knows E, everybody knows that everybody knows that everybody knows E, and so on. This event can be modeled by a coinductive definition:

$$\begin{aligned}&\mathsf{CoInductive}\ \mathsf{CK}\ : \mathsf{Event} \to \mathsf{Event}\ :=\\ &\quad \mathsf{CKintro} : \forall (E : \mathsf{Event})(w : \mathsf{State}), \mathsf{K}\, E\, w \to \mathsf{CK}(\mathsf{K}\, E)\, w \to \mathsf{CK}\, E\, w.\end{aligned}$$

A number of standard results in epistemic logic can be proved. The proofs of some of them are not standard, because we use the specific rules for coinductive definitions in type theory. What follows is a list of results (corresponding to Lemmas 4–9 of Aumann [1]), with example proofs for a couple of them.

Theorem 3.1. (1) $\forall i E, \mathsf{Knows}\, i\, E \subset E$;
(2) $\forall E i, \mathsf{CK}\, E \equiv \mathsf{Knows}\, i\, (\mathsf{CK}\, E)$;
(3) $\forall E F, (E \subset F) \to \forall i, \mathsf{Knows}\, i\, E \subset \mathsf{Knows}\, i\, F$;
(4) $\forall E F i, (\mathsf{Knows}\, i\, E) \sqcap (\mathsf{Knows}\, i\, F) \equiv \mathsf{Knows}\, i\, (E \sqcap F)$;
(5) $\forall E, \mathsf{CK}\, E \subset E$;
(6) $\forall E i, \neg\, \mathsf{Knows}\, i\, E \equiv \mathsf{Knows}\, i\, (\neg\, \mathsf{Knows}\, i\, E)$.

Proof. We illustrate our proof techniques by showing explicitly the proofs of the left-to-right direction of points 2 and 6. The first states that *if E is common knowledge, then everybody knows that E is common knowledge.* The second states that *if i doesn't know E, then he knows that he doesn't know it* (the Socratic principle).

Proof of the left-to-right direction of 2. We proceed by coinduction: We unfold the definitions of $\subset$ and Knows and assume the statement as coinductive hypothesis:

$$H : \forall E i w, \mathsf{CK}\, E\, w \to \forall v, \mathcal{K}_i\, v\, w \to \mathsf{CK}\, E\, v.$$

We remember that, since CK is a coinductive predicate, we can prove the statement using hypothesis H, as long as the occurrences of H are all guarded by the constructor $\mathsf{CKintro}$. So, assume that E, i, w are given such that $\mathsf{CK}\, E\, w$ holds, and that v is such that $\mathcal{K}_i\, v\, w$ holds; then we must prove $\mathsf{CK}\, E\, v$. Since $\mathsf{CK}\, E\, w$, by definition of CK we must have that $\mathsf{CK}\, (\mathsf{K}\, E)\, w$; again by definition of CK we must also have that $\mathsf{K}\, (\mathsf{K}\, E)\, w$. Now, to prove $\mathsf{CK}\, E\, v$ we use $\mathsf{CKintro}$, therefore we must prove $\mathsf{K}\, E\, v$ and $\mathsf{CK}\, (\mathsf{K}\, E)\, v$. $\mathsf{K}\, E\, v$ follows from $\mathsf{K}\, (\mathsf{K}\, E)\, w$ applied to i and $\mathcal{K}_i\, v\, w$. $\mathsf{CK}\, (\mathsf{K}\, E)\, v$ follows from the coinductive hypothesis H applied to the event $\mathsf{K}\, E$ and from $\mathsf{CK}\, (\mathsf{K}\, E)\, w$. Notice that H was used to prove a subgoal generated by the application of $\mathsf{CKintro}$, therefore the guardedness condition is satisfied.

Proof of the left-to-right direction of 6. Unfolding definitions, we need to prove:

$$\forall E i w, \sim(\mathsf{Knows}\, i\, E\, w) \to \forall v, \mathcal{K}_i\, v\, w \to \sim(\mathsf{Knows}\, i\, E\, v).$$

Assume that E, i, w are such that $\mathsf{Knows}\, i\, E\, w$ doesn't hold and that v is such that $\mathcal{K}_i\, v\, w$ holds. Suppose, towards a contradiction, that $\mathsf{Knows}\, i\, E\, v$ holds. We will prove that also $\mathsf{Knows}\, i\, E\, w$ must hold, contradicting the hypothesis. By definition of Knows, we have that:

$$\mathsf{Knows}\, i\, E\, w \Leftrightarrow \forall u, \mathcal{K}_i\, u\, w \to E\, u.$$

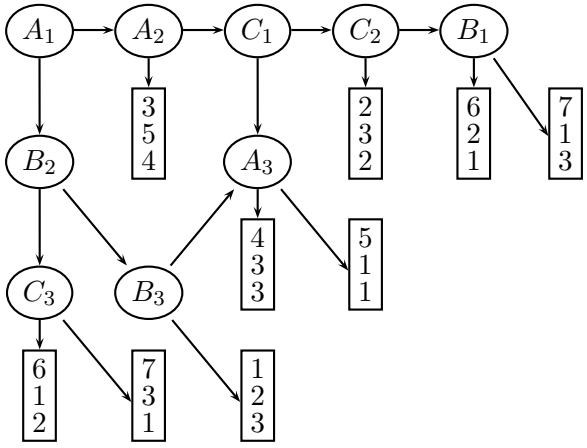
So assume that u is such that $\mathcal{K}_i\, u\, w$ holds. By symmetry and transitivity of $\mathcal{K}_i$, we have that also $\mathcal{K}_i\, u\, v$ must hold; therefore $E\, u$ is true by the assumption $\mathsf{Knows}\, i\, E\, v$. We conclude that $\mathsf{Knows}\, i\, E\, w$ holds, leading to a contradiction, as desired. □

4. Games and Backward Induction

We will consider perfect information (PI) games: there is a set of players taking turns in the play; in each possible position it is determined which player has the initiative and what are the possible moves. Every play of the game is finite, that is, it will end after a finite number of moves. Players have complete knowledge of the position of the game. Therefore we exclude games like poker, in which some information is known to one player but not to the others (that player's cards). Every player has preferences with respect to the final positions. Often, these preferences are given by numerical payoffs.

In the literature, a game is represented by a well-founded finitely branching tree in which the root represents the initial position, nodes represent later positions, and leaves represent the final positions of the game. Nodes are labelled with the player who has the initiative at the corresponding position. Leaves are labelled with payoffs.

I choose a different formalization, that I deem more flexible. A game will be represented by the following elements: a set of positions; an assignment of a player to every position (the one on turn); a set of possible outcomes; for every position, two possible moves to another position or an outcome. We restrict to binary moves, instead of having a finite number of possible moves for every position. This is not a real restriction, since a choice of n moves can be simulated by $n-1$ successive nodes with the same player on turn. Here is an example of a game in this form:



There are three players: A, B, and C. In the position on the left top corner (A_1), player A has two possible moves: the first one leads to the position B_2 in which B is on turn; the second one to the position A_2 in which A is again on turn and has again two possible moves: the first one determines an immediate end of the game with 3 points awarded to A, 5 to B, and 4 to C; the second leads to the position C_1 in which C is on turn. This is equivalent to a position in which A has three possible moves: the first to position B_2, the second to the end position with outcome $(3, 5, 4)$, and the third to position with C_1.

In Coq voice, all this becomes (Pos is the disjoint union of positions and outcomes):

```
Position : Set
Outcome : Set
turn : Position → Player
move : Position → Pos × Pos
```

```
Inductive Pos : Set :=
  ino : Outcome → Pos
  inp : Position → Pos
```

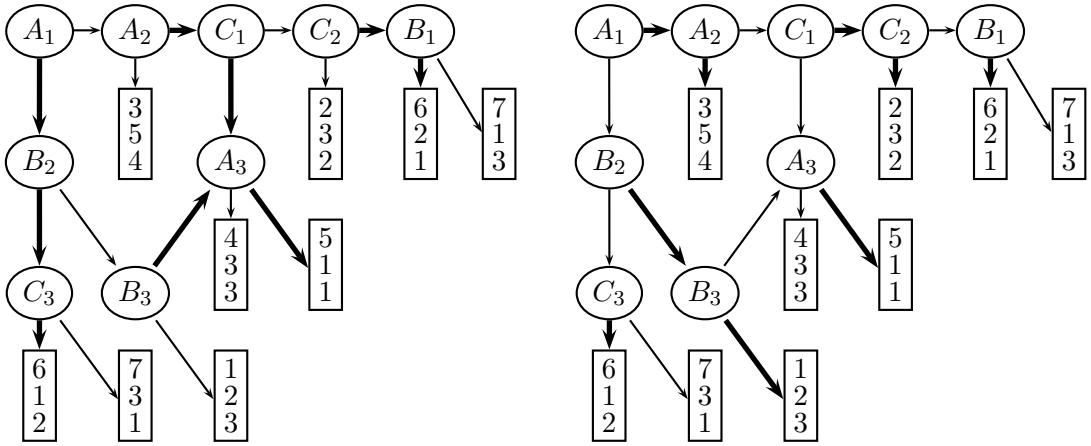
Each player expresses his preferences by a total decidable order on the outcomes:

```
Preference : Player → Outcome → Outcome → Prop
preference_dec : ∀ i p1 p2, {Preference i p1 p2} + {Preference i p2 p1}.
```

Finiteness of the game is expressed by stating that move is wellfounded, that is, every play of the game will certainly terminate after a finite number of moves. I omit the exact definition here (see the Coq file). This condition will allow us to make definitions and write proofs about the game by induction, according to the following principle:

```
game_recursion :  ∀(P : Pos → Type),
                  (∀o : Outcome, P (ino p)) →
                  (∀p : Position, P (first (move p)) → P (second (move p)) → P (inp p)) →
                  ∀𝔭 : Pos, P 𝔭
```

A *strategy* for a player i is a function by which i chooses what move to play in every possible position. A *strategy profile* is a choice of a strategy by every player, that is, a function that determines the move at each position. Here are two possible strategy profiles for the game above. The chosen moves are indicated by a thick arrow.



Given a strategy profile, we can simulate the game starting at any position and determine what the outcome will be. For example, the strategy profile on the left results in the outcome $(6, 1, 2)$ for the position A_1: just follow the thick lines and you obtain the play $A_1, B_2, C_3, (6, 1, 2)$. Formally, defining Action := go_left|go_right, we have:

```
Str_profile : Set := Position → Action
Strategy (i : Player) : Set := ∀p : Position, turn p = i → Action

profile_outcome : Str_profile → Pos → Outcome
profile_outcome π
= game_recursion(λ𝔭.Outcome)(λo.o)(λp o1 o2.Match (π p) with go_left  ↦ o1
                                                              go_right ↦ o2).
```

From now on, I will write functions obtained by recursion using recursive equations, rather that on explicit application of game_recursion:

```
profile_outcome : Str_profile → Pos → Outcome
profile_outcome π (ino o) = o
profile_outcome π (inp p) = Match (π p) with go_left  ↦ profile_outcome (first (move p))
                                             go_right ↦ profile_outcome (second (move p)).
```

Notation: Given a strategy profile π and a strategy σ for player i, we denote by $\pi[i/\sigma]$ the profile in which i adopts strategy σ and every other player adopts the same strategy as in π.

The question now arises: what is the optimal strategy for each player? There is no absolute answer to this question, since the best strategy for a player depends on the strategies adopted by the other players. However, there is a notion of optimal strategy profile, in which each player adopts the locally best move at each position, assuming recursively that the other players adopt the same strategy on the children of the present node. The profile satisfying this condition is called a *Backward Induction Equilibrium* (BIE). We define it recursively in the following way:

$$
\begin{aligned}
&\mathsf{ind_{load}} : \mathsf{Pos} \to \mathsf{Action} \times \mathsf{Outcome}\\
&\mathsf{ind_{load}}\,(\mathsf{ino}\,o) = \langle \mathsf{go\_left}, o\rangle\\
&\mathsf{ind_{load}}\,(\mathsf{inp}\,p) = \mathsf{if}\ (\mathsf{Preference}\,(\mathsf{turn}\,p)\,o_1\,o_2)\ \mathsf{then}\ \langle \mathsf{go\_left}, o_1\rangle\ \mathsf{else}\ \langle \mathsf{go\_right}, o_2\rangle\\
&\qquad\qquad \text{where}\ \langle a_1, o_1\rangle = \mathsf{ind_{load}}\,(\mathsf{first}\,(\mathsf{move}\,p))\\
&\qquad\qquad\qquad\quad \langle a_2, o_2\rangle = \mathsf{ind_{load}}\,(\mathsf{second}\,(\mathsf{move}\,p))\\
&\mathsf{ind} : \mathsf{Str\_profile}\\
&\mathsf{ind}\,p = \mathsf{first}\,(\mathsf{ind_{load}}\,(\mathsf{inp}\,p)).
\end{aligned}
$$

This strategy is optimal for every player, but only if all other players also adopt it. In our example game, the BIE strategy profile is the one above on the right. For short we will call the BIE for a specific player the *inductive strategy.* In general we do not assume that the players know each other strategies, therefore it is not immediately clear that they would choose the backward induction equilibrium. To see what strategy a rational player would actually choose we need some assumptions on what the knowledge of each player is. Aumann's achievement consists in finding an epistemic characterization of BIE based on the notion of common knowledge of rationality.

5. Rationality

Let us combine the development of epistemic logic and of game theoretic notions from the previous two sections. We assume that, in every possible state of the world, each player fixes a strategy by which to play the game:

$$
\begin{aligned}
&\mathsf{str\_profile} : \mathsf{State} \to \mathsf{Str\_profile}\\
&\mathsf{strategy} : \mathsf{State} \to \forall i, \mathsf{Strategy}\,i\\
&:= \lambda w i.(\mathsf{str\_profile}\,w)_i
\end{aligned}
$$

where the subscript i denotes that we are selecting the strategy of player i from the profile strategy $(\mathsf{str\_profile}\,w)$.

For a position p, $\mathsf{Ind}\,p$ denotes the event stating that the inductive strategy is adopted at position p; for $\mathfrak{p} : \mathsf{Pos}$, $\mathsf{CK\_Ind_{pos}}\,\mathfrak{p}$ is the event stating that it is common knowledge that the inductive strategy is adopted at $\mathfrak{p}$ and at every position reachable from it:

$$
\begin{aligned}
&\mathsf{Ind} : \mathsf{Position} \to \mathsf{Event}\\
&\mathsf{Ind}\,p = (\lambda w.\mathsf{str\_profile}\,w\,p = \mathsf{ind}\,p)
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{CK\_Ind_{pos}} : \mathsf{Pos} \to \mathsf{Event}\\
&\mathsf{CK\_Ind_{pos}}\,(\mathsf{ino}\,o) = \top\\
&\mathsf{CK\_Ind_{pos}}\,(\mathsf{inp}\,p) = \mathsf{CK}\,(\mathsf{Ind}\,p) \sqcap (\mathsf{CK\_Ind_{pos}}\,(\mathsf{first}\,(\mathsf{move}\,p)) \sqcap (\mathsf{CK\_Ind_{pos}}\,(\mathsf{second}\,(\mathsf{move}\,p)).
\end{aligned}
$$

We denote simply by BIE the statement that the inductive strategy is adopted at every position:

$$\mathsf{BIE} : \mathsf{Event}$$
$$\mathsf{BIE} = (\lambda w. \forall p, \mathsf{str\_profile}\, w\, p = \mathsf{ind}\, p) = (\lambda w. \forall p, \mathsf{Ind}\, p).$$

Being rational means choosing the best move; but the judgement of what move is best depends on what kind of knowledge the player has about the other players' strategies. In each state w, a player i knows that the strategy profile is one of those associated with any state v such that $(\mathcal{K}\, i\, v\, w)$ holds. The strategy adopted by i at a certain position p is called rational if no other strategy would give a better outcome for sure. This means that no other strategy would have given a better outcome in every state v such that $(\mathcal{K}\, i\, v\, w)$ holds. Formally,

$$\mathsf{better\_strategy} : \forall p, \mathsf{Strategy}\,(\mathsf{turn}\, p) \to \mathsf{Event}$$
$$\mathsf{better\_strategy}\, p\, \sigma\, w = \mathsf{Preference}\,(\mathsf{turn}\, p)\ \ \mathsf{profile\_outcome}\,(\mathsf{Str\_profile}\, w)[i/\sigma]\,(\mathsf{inp}\, p)\ \ \mathsf{profile\_outcome}\,(\mathsf{Str\_profile}\, w)\,(\mathsf{inp}\, p)$$

$$\mathsf{Rational} : \mathsf{Position} \to \mathsf{Event}$$
$$\mathsf{Rational}\, p\, w = \forall \sigma : \mathsf{Strategy}\,(\mathsf{turn}\, p), (\neg\,(\mathsf{Knows}\,(\mathsf{turn}\, p)\,(\mathsf{better\_strategy}\, p\, \sigma)))\, w$$

Intuitively the event $\mathsf{Rational}\, p$ says: *In position p, the player on turn doesn't know whether there is a better strategy than the one he actually adopts.* The rationality event is the statement that players always act rationally:

$$\mathsf{Rationality} : \mathsf{Event}$$
$$\mathsf{Rationality}\, w = \forall p, \mathsf{Rational}\, p\, w.$$

We are now in the position to state and prove Aumann's theorem. Informally it states: *If there is common knowledge of rationality, then every player will adopt the inductive strategy.* The statement is proved by induction on the positions, using a loaded induction predicate that states, at each position $\mathfrak{p}$: *It is common knowledge that the inductive strategy will be adopted at $\mathfrak{p}$ and at every position reachable from $\mathfrak{p}$*, that is $\mathsf{CK\_Ind}_{\mathsf{pos}}\, \mathfrak{p}$. We give an outline of the proof; see [1] for detailed informal proof (slightly different from ours), and the Coq file for the formal one.

Lemma 5.1.

$$\forall w, \mathsf{CK}\, \mathsf{Rationality}\, w \to \forall \mathfrak{p}, \mathsf{CK\_Ind}_{\mathsf{pos}}\, \mathfrak{p}\, w.$$

Proof. Fix the state w and assume that $\mathsf{CK}\, \mathsf{Rationality}\, w$ holds. We prove the statement $\forall \mathfrak{p}, \mathsf{CK\_Ind}_{\mathsf{pos}}\, \mathfrak{p}\, w$ by induction on $\mathfrak{p}$.

If $\mathfrak{p} = \mathsf{ino}\, o$ then the statement $\mathsf{CK\_Ind}_{\mathsf{pos}}\,(\mathsf{ino}\, o)\, w$ is trivially true by definition of $\mathsf{CK\_Ind}_{\mathsf{pos}}$.

If $\mathfrak{p} = \mathsf{inp}\, p$ then we have, by induction hypothesis, that

$$\mathsf{CK\_Ind}_{\mathsf{pos}}\,(\mathsf{first}\,(\mathsf{move}\, p))\, w$$
$$\mathsf{CK\_Ind}_{\mathsf{pos}}\,(\mathsf{second}\,(\mathsf{move}\, p))\, w.$$

That is: It is common knowledge that the inductive strategy will be adopted by every player in the prosecution of the game both if $\mathsf{turn}\, p$ plays to the left or to the right. In particular $\mathsf{turn}\, p$ knows that the inductive strategy will be adopted after his move, therefore he can deduce that his payoff if he moves left or right will be, respectively,

$$o_1 = \mathsf{profile\_outcome}\,(\mathsf{ind}\,(\mathsf{first}\,(\mathsf{move}\, p))) = \mathsf{first}\,(\mathsf{ind}_{\mathsf{load}}\,(\mathsf{first}\,(\mathsf{move}\, p))) \text{ and}$$
$$o_2 = \mathsf{profile\_outcome}\,(\mathsf{ind}\,(\mathsf{second}\,(\mathsf{move}\, p))) = \mathsf{first}\,(\mathsf{ind}_{\mathsf{load}}\,(\mathsf{second}\,(\mathsf{move}\, p))).$$

If $\mathsf{turn}\, p$ is rational, by definition of rationality, he will chose the move that gives him the outcome that he prefers. In other words, we will choose the inductive strategy at position p. We have therefore proved that $\mathsf{Ind}\, p$ holds.

Notice now that Statement 3 of Theorem 3.1 extends to common knowledge, that is, if an event E implies another event F, then $\mathsf{CK}\, E$ implies $\mathsf{CK}\, F$. We just proved that $\mathsf{Rationality}$ implies $\mathsf{Ind}\, p$, therefore $\mathsf{CK\ Rationality}$ implies $\mathsf{CK}\, (\mathsf{Ind}\, p)$. Since $\mathsf{CK\ Rationality}\, w$ holds by hypothesis, we conclude that $\mathsf{CK}\, (\mathsf{Ind}\, p)\, w$ is true.

The other two conjuncts in the statement $\mathsf{CK\_Ind_{pos}}\, (\mathsf{inp}\, p)$ are just the two inductive hypotheses, so we conclude that the statement is true. □

Theorem 5.2 (Aumann)**.**

$$\mathsf{CK\ Rationality} \sqsubset \mathsf{BIE}.$$

Proof. Assume $\mathsf{CK\ Rationality}$

By the previous lemma $\mathsf{CK\_Ind_{pos}}\, (\mathsf{inp}\, p)$ holds for every position p. But $\mathsf{CK\_Ind_{pos}}\, (\mathsf{inp}\, p)$ trivially implies $\mathsf{Ind}\, p$. Therefore $\mathsf{Ind}\, p$ holds for every p, that is, BIE is true. □

6. Back to the Amazons

Here is what happened in the village of the Amazons. Let us call n the number of the Amazons. For the first $n-1$ days after the goddess' announcement nothing happened: Nobody painted her door green. But on the nth day all the Amazons painted their doors and they all got their reward.

That is because for the first $n-1$ days nobody could be sure of the colour of her own eyes, but on the nth day they could all deduce logically that they had green eyes.

They reasoned by induction on the number m of green-eyed Amazons. They know, because the goddess told them, that m is at least 1.

If $m = 1$, then there is only one green-eyed Amazon. She sees that the others have eyes of different colour, so she deduces that she must have green eyes. Therefore, she paints her door on the first night.

If $m = 2$, there are two green-eyed Amazons. Each of them makes this reasoning: 'I can see only one Amazon with green eyes. If she were the only one, then, by the case $m = 1$, she would paint her door on the first night.' So they both wait for one day. Since neither paints their door, they can deduce that $m > 1$. But they can see that there is no other green-eyed Amazon, therefore they know that they have green eyes themselves. Thus they paint their door on the second night.

Now the induction step. Suppose that there are $m + 1$ green-eyed Amazon. Each of them sees m Amazons with green eyes. Each of them thinks: 'If my eyes are not green, then there are exactly m Amazons with green eyes'. By induction hypothesis, she knows that they would all paint their door on the mth night. So she waits for m days to see what happens. Everybody else does the same. So on the mth night no door is painted. Then each concludes that the assumption that her own eyes were not green must be wrong, and she paints her door on the $(m + 1)$st night.

References

[1] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[2] Jon Barwise. Scenes and other situations. *The Journal of Philosophy*, 78(7):369–397, July 1981.

[3] Yves Bertot and Pierre Castéran. *Interactive Theorem Proving and Program Development. Coq'Art: The Calculus of Inductive Constructions*. Springer, 2004.

[4] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. The MIT Press, 1995.

[5] George Gamow and Marvin Stern. *Puzzle Math*. Viking Press, New York, 1958.

[6] Martin Gardner. *Puzzles from Other Worlds*. Vintage, 1984.

[7] Vincent Hendricks and John Symons. Epistemic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2006. `http://plato.stanford.edu/archives/spr2006/entries/logic-epistemic`.

[8] The Coq Development Team. LogiCal Project. *The Coq Proof Assistant. Reference Manual. Version 8*. INRIA, 2004. `http://pauillac.inria.fr/coq/coq-eng.html`.

[9] René Vestergaard, Pierre Lescanne, and Hiroakira Ono. The inductive and modal proof theory of Aumann's theorem on rationality. technical report: JAIST/IS-RR-2006-009, `http://www.jaist.ac.jp/~vester/Writings/index.html#formal-GT`, 2006.