Thought: A Journal of Philosophy (2019)
doi: 10.1002/tht3.435

Two Tales of Epistemic Models

YANG LIU
University of Cambridge

This short paper has two parts. First, we prove a generalisation of Aumann's surprising impossibility result in the context of rational decision making. We then move, in the second part, to discuss the interpretational meaning of some formal setups of epistemic models, and we do so by means of presenting an interesting puzzle in epistemic logic. The aim is to highlight certain problematic aspects of these epistemic systems concerning first/third-person asymmetry which underlies both parts of the story. This asymmetry, we argue, reveals certain limits of what epistemic models can be.

1. Introduction

In a well-known article titled "Agreeing to Disagree," Aumann (1976) proved a surprising impossibility result. The result was given in a Bayesian setting and was accompanied by an extraordinary cover story: If a group of people have common priors, and their posteriors of a given event are also common knowledge, then no matter what respective evidence they each obtains in updating their probabilities on the event it must be that their posteriors are identical. In other words, it is *impossible* for these agents to form different (probabilistic) opinions about any given event.

Aumann's result prompted a long and fruitful series of studies on the epistemic foundations of game theory in the economic literature. To uncover the basic logic behind the so-called agreeing-to-disagree type argument, in Section 2, we provide a generalisation of this impossibility result in the context of rational decision making. Our generalisation carries a similar element of surprise: If a group of (human or machine) decision makers have similar backgrounds, follow the same decision procedure in decision situations, and that, in making any particular decision, their decisions are commonly shared; then, no matter what respective private information they each possesses in making their final decisions, it is *impossible* that these decisions differ.

Dramatic as it may sound, this generalisation, like many agreeing-to-disagree type arguments, relies on the interpretation of the formalism involved. The latter, however, often contains various highly idealised and, in some cases, notoriously ambiguous assumptions. These idealisations in turn may give rise to various mysteries or even inconsolable disputes among experts.

Indeed, along with rapid developments of game theory in the past four decades or so we have seen increasing and persistent disagreements among philosophers and game theorists on foundational issues concerning the interpretational meaning of various basic concepts in game theory. The heated exchange between Kadane and Larkey (1982a,b) and Harsanyi (1982a,b) is one such example of clash of opinions among experts, where both sides provide strong yet opposing views on how basic notions such as Nash equilibrium and common priors should be understood.[1] The broader philosophical aim of this paper is to continue these conversations on foundational issues, and to provide an analysis of the interpretational value of some specific aspects of epistemic models.

To broaden and to put our discussion on a more concrete footing, in Section 3 we highlight an interesting puzzle in epistemic logic. The latter is an area of research in philosophy closely related to epistemic game theory in economics. The puzzle can be dubbed along the following lines: Unless an agent is epistemically ideal (to be made precise below) it is *impossible* for them to know their own epistemic limitations. We argue that this simple puzzle reveals a certain shortcoming of epistemic modelling, that is, within the confines of how notions like knowledge, beliefs, and information are represented in epistemic models, many suboptimal epistemic properties do not seem to be characterisable in these models.

As we will see, there is a common theme that runs through both stories presented here, and it concerns *the first/third-person asymmetry* commonly seen in epistemic models, i.e., the contrast between the perspective of the agents being modelled and that of the theorists who carry out the modelling. This observation is not new, but it is not widely appreciated as it should be. We shall demonstrate this asymmetry in a simple and streamlined manner. Our aim is to stress the significance, as well as the implications, of this asymmetry in a broader context of epistemic and decision/game-theoretic modelings in general, and to highlight a certain limit of what epistemic models can be.

In what follows, we presuppose some knowledge in epistemic logic. But to make the paper more or less self-contained, we will introduce various basic concepts and formal definitions. All trivial proofs are omitted. Readers who are not interested in the formalism presented in Sections 2.1 and 2.2 may skip directly to the 'story time' in Section 2.3 without missing much of the philosophical content of the paper.

2. Impossible to Disagree

2.1. Preliminaries

Let Ω be a (finite) set, referred to as the state space. A Kripke model (of knowledge or beliefs) over Ω is a relational structure distinguished by a binary relation $\rightsquigarrow \subseteq \Omega \times \Omega$, where $\rightsquigarrow$ is often referred to as an *accessibility relation* among possible states. Intuitively,

1. Another notable example can be seen in a standard graduate textbook in game theory by Osborne and Rubinstein (1994), where, from time to time, the authors disagree with one another on various foundational issues and record their respective views in parallel.

$\omega \rightsquigarrow^i \omega'$ says that, from the perspective of player i, ω' is considered doxastically possible in state ω. We also say that state ω' is $\rightsquigarrow^i$*-accessible* from ω. The following is a list of properties of $\rightsquigarrow$ that are commonly adopted in a Kripke frame: for any $\omega, \omega', \omega'' \in \Omega$,

Seriality for each ω there exists an ω' such that $\omega \rightsquigarrow \omega'$.

Reflexivity $\omega \rightsquigarrow \omega$.

Transitivity if $\omega \rightsquigarrow \omega'$and $\omega' \rightsquigarrow \omega''$ then $\omega \rightsquigarrow \omega''$.

Euclid if $\omega \rightsquigarrow \omega'$ and $\omega \rightsquigarrow \omega''$ then $\omega' \rightsquigarrow \omega''$.

As standard definitions in epistemic logic, i is said to be a **S4** agent if $\rightsquigarrow^i$ is reflexive and transitive, a **S5** agent if it is also Euclidean. The latter is taken to be *epistemically ideal* as they satisfies both the positive and the negative introspection conditions.

Information structure. Given $\rightsquigarrow^i$ above, define function $\mathcal{I}^i : \Omega \to 2^\Omega$ as

$$\mathcal{I}^i(\omega) := \Big\{\omega' \in \Omega \mid \omega \rightsquigarrow^i \omega'\Big\}. \tag{2.1}$$

Then $\mathcal{I}^i(\omega)$ is the set of states that are $\rightsquigarrow^i$-accessible from ω, call $\mathcal{I}^i(\omega)$ the *information set* of i in state ω. The interpretation is that $\mathcal{I}^i(\omega)$ contains all the relevant information that is accessible by i at ω. For any *event* (or *proposition*) E (a set of states), denote by $\mathcal{I}^i(E)$ the set of all states that are $\rightsquigarrow^i$-accessible from all states in E, i.e., $\mathcal{I}^i(E) = \bigcup_{\omega \in E} \mathcal{I}^i(\omega)$. Next, define $\mathcal{I}^i = \{\mathcal{I}^i(\omega) \mid \omega \in \Omega\}$ to be the *information structure* of player i. Clearly, $\mathcal{I}^i$ provides a complete description of what information player i has in each state.[2]

Player i is said to be *informed of* (the occurrence of) a certain event E in state ω just in case $\mathcal{I}^i(\omega) \subseteq E$, i.e., if the information i possesses at ω is contained in E. If the underlying model is intended to be a representation of agents' knowledge (beliefs) then $\mathcal{I}^i(\omega) \subseteq E$ is interpreted as saying that i *knows* (*believes in*) E in ω.

Alternatively, one can take the information function $\mathcal{I}^i : \Omega \to 2^\Omega$ of player i as primitive and define i's accessibility relation $\rightsquigarrow^i$ over Ω by

$$\rightsquigarrow^i := \Big\{(\omega, \omega') \in \Omega \times \Omega \;\Big|\; \omega' \in \mathcal{I}^i(\omega)\Big\}. \tag{2.2}$$

Consider the following properties of an information structure $\mathcal{I}$: for any $\omega \in \Omega$,

Viability $\mathcal{I}(\omega) \neq \emptyset$.

Factivity $\omega \in \mathcal{I}(\omega)$.

Inclusion if $\omega' \in \mathcal{I}(\omega)$ then $\mathcal{I}(\omega') \subseteq \mathcal{I}(\omega)$.

2. In order not to overburden the symbolism in use, here we adopt a systematic ambiguity of using $\mathcal{I}^i$ to denote the information structure of player i, $\mathcal{I}^i(\omega)$ to denote the information i possesses at ω, and $\mathcal{I}(E)$ to denote the collective information i possesses at all states in E.

Figure 2.1: Player i's information structure with one blindspot.

Mutuality if $\omega', \omega'' \in \mathcal{I}(\omega)$ then $\omega'' \in \mathcal{I}(\omega')$ and $\omega' \in \mathcal{I}(\omega'')$.

Proposition 2.1. Let $\mathcal{I}$ be an information structure and $\rightsquigarrow$ be the corresponding accessibility relation for which (2.1) and (2.2) are satisfied, then

(1) $\rightsquigarrow$ is serial if and only if $\mathcal{I}$ is viable,

(2) $\rightsquigarrow$ is reflexive if and only if $\mathcal{I}$ is factive,

(3) $\rightsquigarrow$ is transitive if and only if $\mathcal{I}$ is inclusive,

(4) $\rightsquigarrow$ is Euclidean if and only if $\mathcal{I}$ is mutual.

Definition 2.2. An information structure $\mathcal{I}$ is said to be *divisible* if it is (a) viable, (b) inclusive, and (c) mutual; $\mathcal{I}$ is *partitional* if it is divisible and $\mathcal{I}(\Omega) = \Omega$.

It is easy to see that if $\mathcal{I}$ is divisible then, for any $\omega, \omega' \in \Omega$, either $\mathcal{I}(\omega) \cap \mathcal{I}(\omega') = \emptyset$ or $\mathcal{I}(\omega) = \mathcal{I}(\omega')$. Further, if $\mathcal{I}$ is divisible then, by Proposition 2.1, the corresponding accessibility relation $\rightsquigarrow$ forms an equivalence relation over Ω, in which case we have that $\mathcal{I}(\omega) = [\omega]_{\rightsquigarrow}$ for all $\omega \in \Omega$.

Doxastic blindspot. The above construction enables a formal classification of states in Ω. Note that if $\mathcal{I}(\Omega)$ is a proper subset of Ω, i.e., if $\Omega - \mathcal{I}(\Omega) \neq \emptyset$, then it follows that there are states that do not belong to any information set. In other words, given definitions in (2.1) and (2.2), members of $\Omega - \mathcal{I}(\Omega)$ are states that are *not* $\rightsquigarrow^i$-accessible for i from *any* state in Ω. We refer to such states as player i's doxastic blindspots. Formally, for any $\omega \in \Omega$, ω is said to be a *doxastic blindspot* (or *blindspot* for short) of player i if there does not exist any $\nu \in \Omega$ such that $\omega \in \mathcal{I}^i(\nu)$. Denote the set of all blindspots of i by $\mathcal{B}^i(\Omega)$.

Example 2.3. Let $\Omega = \{\omega_1, \omega_2\}$ and $\mathcal{I}^i(\omega_1) = \mathcal{I}^i(\omega_2) = \{\omega_2\}$. Then ω_1 is a doxastic blindspot for i (see Figure 2.1).

A doxastic blindspot ω can also be interpreted as saying that the agent may falsely believe $\mathcal{I}^i(\omega)$ as they do not consider ω as an epistemic possibility. Then, from the definition, it is clear that the concept of blindspots is only intelligible when it is modelled from the third person point of view – a lesson we learn from G. E. Moore. We shall return to this point on *first/third person asymmetry* in the next section. The following are some simple properties of doxastic blindspots.

Proposition 2.4. $\mathcal{I}^i(\Omega) \cap \mathcal{B}^i(\Omega) = \emptyset$ and $\mathcal{I}^i(\Omega) \cup \mathcal{B}^i(\Omega) = \Omega$.

Note that the existence of blindspots differentiates a doxastic information model from an epistemic one. The latter is widely employed in representing knowledge where it is assumed that $\omega \in \mathcal{I}^i(\omega)$ for all $\omega \in \Omega$, that is, it is assumed that it is impossible that the players' information sets exclude "the true state of the world" (and hence there is no blindspot). The assumption is often referred to as the "truth condition" of information sets which is connected to the assumption that knowledge is infallible.

This assumption is by no means uncontroversial, especially when viewed from the agent's first person perspective – it is unclear how blindspots can be eliminated by stipulating that players' information be always truthful. As we shall see, contrary to what many had thought, this infallibility condition is in fact not essential to the agreeing-to-disagree type arguments, to which we now turn.

A generalised sure-thing principle. To arrive at a generalisation of Aumann's impossibility result for rational decision making, we invoke a general decision rule. Here we are guided by the "sure-thing principle" of Savage (1972).

Recall that Savage's sure-thing principle – formulated as the second postulate (P2) in his axiomatic system – is derived from, what he calls, a "loose" version of the sure-thing principle (STP) which says that if a decision maker prefers one act over another assuming either that a certain event obtains or that it does not obtain, then their preference ranking over the two acts should remain unchanged regardless how this given event transpires.[3]

This "loose" version of STP is sometimes referred to as the *dominance principle*. The latter captures an intuitive idea of reasoning by cases: if one option is weakly preferred to others in all situation under which these options are compared then it should be weakly preferred simpliciter. This consideration then gives rise to the following decision rule which generalises the sure-thing principle.

General Sure-Thing Principle (GSTP): If a decision maker makes the same decision conditioning on the information they possesses in all possible decision situations, then they should make the same decision unconditionally.

Formally, let D be a nonempty set with an unspecified domain. We take that an agent's *decision algorithm* – i.e., the manner with which the agent makes decision – is determined by a function f mapping from 2^Ω to D.

Definition 2.5. A decision algorithm $f : 2^\Omega \to D$ is said to be an *informational decision function for player i* if, for any $S \subseteq \Omega$, the following condition holds:

$$f\big(\mathcal{I}^i(\nu)\big) = d \ \text{ for all } \nu \in S \quad \Longrightarrow \quad f\Big(\bigcup_{\nu \in S} \mathcal{I}^i(\nu)\Big) = d. \tag{GSTP}$$

3. For an analysis of the subtle difference between the sure-thing principle in its original form and its formulation as P2 in Savage's axiomatic system see Liu (2017).

Intuitively, $f\big(\mathcal{I}^i(\nu)\big) = d$ is a decision made by player i according to the decision algorithm f based on their information in state ν (i.e., $\mathcal{I}^i(\nu)$), and S is a set of possible decision situations. Then (GSTP) says that if player i makes the same decision d in all possible situations in S (i.e., i makes the same decision d based on information $\mathcal{I}^i(\nu)$ for all $\nu \in S$), then they decide on *d simpliciter*.

Group information. In the interactive situation, let $\rightsquigarrow^N$ be the smallest transitive relation that contains all the $\rightsquigarrow^i$ relations, that is,

$$\rightsquigarrow^N := \mathrm{TC}\Big(\bigcup_{i\in N} \rightsquigarrow^i\Big), \tag{2.3}$$

where 'TC' stands for the transitive closure operator. Then, relation $\rightsquigarrow^N$ represents the maximum reachability relation of all $\rightsquigarrow^i$'s (cf. Proposition 2.6(1) below). Call $\rightsquigarrow^N$ the *group accessibility relation* of N. For any $(\omega, \omega') \in \rightsquigarrow^N$, we also say that ω' is $\rightsquigarrow^N$-*accessible from* ω. From the group accessibility relation $\rightsquigarrow^N$ a corresponding notion of *group information function* $\mathcal{I}^N : \Omega \to 2^\Omega$ can be defined by

$$\mathcal{I}^N(\omega) = \Big\{\omega' \in \Omega \;\Big|\; \omega \rightsquigarrow^N \omega'\Big\}. \tag{2.4}$$

And let $\mathcal{I}^N$ be the *group information structure* such that $\mathcal{I}^N = \{\mathcal{I}^N(\omega) \mid \omega \in \Omega\}$.

Alternatively, one can take players' information structures $\mathcal{I}^1, \ldots, \mathcal{I}^n$ as primitive and define group information structure $\mathcal{I}^N$ as the *meet* of the $\mathcal{I}^i$s, i.e., as it is standardly defined in Boolean algebra,

$$\mathcal{I}^N = \bigwedge_{i\in N} \mathcal{I}^i. \tag{2.5}$$

Then define group accessibility relation $\rightsquigarrow^N$ by

$$\rightsquigarrow^N := \Big\{(\omega, \omega') \in \Omega \times \Omega \;\Big|\; \omega' \in \mathcal{I}^N(\omega)\Big\}. \tag{2.3$'$}$$

Let E be any event, say that E is *common information* among members of group N at ω, if $\mathcal{I}^N(\omega) \subseteq E$. The following is a list of basic properties of the group accessibility relation $\rightsquigarrow^N$ and the group information structure $\mathcal{I}^N$.

Proposition 2.6. Let $\mathcal{I}^i, \rightsquigarrow^N$, and $\mathcal{I}^N$ be defined as above, then we have

(1) For any $(\omega, \omega') \in \rightsquigarrow^N$, there corresponds a sequence $i_1, i_2, \ldots, i_k \in N$ and a sequence of states $\omega_0, \omega_1, \ldots, \omega_k \in \{\nu \mid \omega \rightsquigarrow^N \nu\}$ with $\omega_0 = \omega$ and $\omega_k = \omega'$ such that $\omega_0 \rightsquigarrow^{i_1} \omega_1 \rightsquigarrow^{i_2} \cdots \rightsquigarrow^{i_k} \omega_k$, where $0 \leq k < \infty$.

(2) For any $\omega \in \Omega$ we have $\mathcal{I}^i(\omega) \subseteq \mathcal{I}^N(\omega)$.

(3) For any $\omega \in \Omega$ and for any $i \in N$, we have $\mathcal{I}^i\big(\mathcal{I}^N(\omega)\big) \subseteq \mathcal{I}^N(\omega)$.

(4) Given any $\omega \in \Omega$, if, for any $i, j \in N$, $\mathcal{I}^i$ and $\mathcal{I}^j$ are divisible and $\mathcal{I}^i(\Omega) = \mathcal{I}^j(\Omega)$, then $\mathcal{I}^N(\omega) = \mathcal{I}^i\big(\mathcal{I}^N(\omega)\big)$.

2.2. Agreeing-to-disagree generalised

With all the preparations given above, we are now in the position to prove the following result, which is a generalisation of Aumann's impossibility result in the context of decision making guided by a generalised sure-thing principle (GSTP).

Theorem 2.7. Let $\Omega, N, \mathcal{I}^i$ be defined as above and ω be the actual state of the world. Suppose that, for any $i, j \in N$,

1. $\mathcal{I}^i, \mathcal{I}^j$ are divisible;
2. $\mathcal{B}^i(\Omega) = \mathcal{B}^j(\Omega)$;
3. f is an informational decision function for all players in N; and
4. i's decision d^i is common information shared among members of N at ω.

Then, $d^i = d^j$ for all $i, j \in N$.

Proof. For any $i \in N$, consider the event

$$E^i = \left\{ \nu \in \Omega \mid f\big(\mathcal{I}^i(\nu)\big) = d^i \right\}, \quad d^i \in D \tag{2.6}$$

where E^i is the set of possible states in which f yield d^i given player i's information $\mathcal{I}^i(\nu)$ at ν. It is plain that $\omega \in E^i$ for all $i \in N$. By definition, the assumption that d^i's are common information at ω amounts to

$$\mathcal{I}^N(\omega) \subseteq \bigcap_{i \in N} E^i. \tag{2.7}$$

Then, $\nu \in \mathcal{I}^N(\omega)$ implies $\nu \in E^i$ via (2.7), and hence, by (2.6),

$$f\big(\mathcal{I}^i(\nu)\big) = d^i. \tag{2.8}$$

Note that each $\mathcal{I}^i$ is assumed to be divisible, this implies, by Proposition 2.6(4), that, for each $i \in N$,

$$\mathcal{I}^N(\omega) = \bigcup_{\nu \in \mathcal{I}^N(\omega)} \mathcal{I}^i(\nu). \tag{2.9}$$

Finally, since f is an informational decision function for each i, apply (GSTP) to (2.8) and (2.9), we get

$$f\big(\mathcal{I}^N(\omega)\big) = d^i.$$

Therefore, $d^i = d^j$ for all $i, j \in N$. □

Remark 2.8. If f is a conditional probability of a given event A (i.e., if $f(\cdot) = \Pr(A \mid \cdot)$) and $\mathcal{I}^i(\Omega) = \Omega$ (and hence $\mathcal{B}^i(\Omega) = \emptyset$ by Proposition 2.4), then the impossibility result of Aumann (1976) becomes a special case of this theorem.

2.3. Story time

Now let us turn the premises and the conclusion of Theorem 2.7 into a cover story: For any group of decision makers, *if*

1. the organisation of their information are well structured (i.e., satisfying the divisibility requirement, that is, for any player and any states ω, ν, the respective information $\mathcal{I}(\omega)$ and $\mathcal{I}(\nu)$ on which their decisions are based in these states are either identical or completely disjoint);

2. they have similar biases (i.e., sharing the same blindspots);

3. they follow the same decision procedures in decisions situations (i.e. using the same decision algorithm/function that respects GSTP); and

4. for any given pending decision, if the decisions each member independently made are shared as common information among the entire group.

Then, regardless what individual information they each possessed in making their respective decision, it is *impossible* that they make different decisions in the end!

Presumably, with a touch of imagination, our cover story can be transformed into a telling tale. Take, for instance, the currently popular subject on the long-term impacts of AI technologies. A stirring story can be told about a group of powerful AIs. Imagine that each of these AIs was created by means of certain "copy-and-paste" procedure (conditions 1-3), and they are highly coordinated and constantly communicate with one another about their life choices (condition 4); but quite unexpectedly, so the story goes, these AIs start consistently making the same bad decisions against the humans (a dramatic yet conceivable possibility that is consistent with the theorem proved above), and eventually drive humanity to the edge of extinction.

3. What Epistemic Models Cannot Be

Fictional stories aside, what interests us besides the theorem proved above is how and from what standpoint should this type of results be interpreted. Note that, in the case of our result, we were helpfully aided by the concept of doxastic blindspots (in conditions 2) which, by its very nature, is not something that is accessible by players themselves. This indicates that our result can only be made intelligible from an entirely *detached point of view*. In fact, all of conditions 1-4 can been seen as *meta-theoretic* information gathered and formulated by a theorist, and Theorem 2.7 is entirely proved and interpreted from this theorist's external point of view – indeed, that is precisely what we, qua theorists, just did throughout Section 2.

A natural question can however be raised here: If *we*, as external theorists, can theorise and model the situation, why cannot the players *themselves* do the same modelling? Of course, given the particular theorem proved above, it makes no sense for players to observe and model their own blindspots. But supposedly there are situations where these agents are more epistemically capable, so to speak, and be able to

Figure 3.1: A model of **S4**

model the game-theoretic situation from their own first-person perspectives, in which case the theorist and the agent have an integrated single standpoint.

After all, let's face it, if game theory is to be 'useful' at all, it must be that the players themselves can somehow use the theory to guide their actions in game situations – it would otherwise be disappointing if all the comings and goings of a game go to the onlookers while the players themselves are largely clueless about what is going on with the game.

3.1. A puzzle

Given this challenge from our critic, let us move to empower our agents with more epistemic capabilities: Let us assume that our agents are as epistemically capable as a **S4** agent. The latter was in fact studied by Hintikka in his original work on epistemic logic (Hintikka, 1962), where the agents – being **S4** – are not assumed to have any blindspots. In addition, to match the capacity of external theorists, let us assume that the agents also have access to the meta-theoretical information about how their own epistemic structures are organised!

Ambitious and capable as our agent now is, we however would like to point out a puzzling phenomenon for this way of modelling agents' epistemic corpus. More precisely, we show that within the standard possible world semantics the following cannot hold simultaneously:

i. The agent is aware of their current epistemic situation.

ii. The agent has access to their own epistemic structure.

iii. The agent is an **S4** agent.

To illustrate, let us continue to use $\mathcal{I}(\cdot)$ and $\mathcal{I}$ to denote the information function and information structure defined in Section 2.1. Then, condition (i) means that, for any given state ω, the agent is aware of all the alternative states that are accessible from ω, i.e., $\mathcal{I}(\omega)$ – or, to use our terminology, the agent is informed of $\mathcal{I}(\omega)$ because, trivially, $\mathcal{I}(\omega) \subseteq \mathcal{I}(\omega)$. Conditions (ii) and (iii) are our assumptions.

Now, by truth definition in a Kripke frame, a proposition p is said to be known by the agent at ω, denoted by $\omega \vDash Kp$, if p is true in all the worlds in $\mathcal{I}(\omega)$. Further, as what is characteristic of an **S4** agent, $\mathcal{I}$ is assumed to be viable and inclusive (or equivalently, the underlying accessibility relation $\rightsquigarrow$ is assumed to be reflexive and transitive). In a model of knowledge, this translates into the assumption of the truth axiom ($Kp \to p$) and that of the positive introspection axiom ($Kp \to KKp$).

Consider the simple example illustrated in Figure 3.1, where Ω contains two possible worlds and the agent's alternativeness relation is represented by the directed graph

with $\mathcal{I}(\omega_1) = \{\omega_1, \omega_2\}$ and $\mathcal{I}(\omega_2) = \{\omega_2\}$. Trivially, this is a model of **S4**. Suppose that ω_1 is the real world and that p is true in ω_1 but false in ω_2. Then, by truth definition, p is *not* known by the agent at ω_1.

So far, nothing in what we have described is out of ordinary. However, observe that, at ω_1, the agent may as well have the following (meta)-reasoning:

> "My epistemic structure is $\mathcal{I} = \{\mathcal{I}(\omega_1), \mathcal{I}(\omega_2)\}$, where $\mathcal{I}(\omega_1)$ and $\mathcal{I}(\omega_2)$ are distinctive with $\mathcal{I}(\omega_1) = \{\omega_1, \omega_2\}$ and $\mathcal{I}(\omega_2) = \{\omega_2\}$. Now given that I currently consider both ω_1 and ω_2 as epistemically possible, then the only possible world in which I could be in such a state is world ω_1, therefore ω_1 must be the real world!"

This consideration then leads the agent to exclude ω_2 as an alternative. As a consequence of this realisation, p becomes known by the agent at ω_1 given the truth definition. *But this is very puzzling because the change of the agent's knowledge of p from unknown to known rests on no further evidence but mere reflection on their own epistemic structure!*

The puzzle can also be stated in the following form. Let Ann be an **S4** but not **S5** agent and let p be any proposition, which Ann does not know, but also does not know that she does not know (i.e., $\neg K_a \neg K_a p$). For example, let p be expressed by the sentence, say, "Michael Atiyah claimed to have proved the Riemann conjecture." Suppose that Bill, who knows that Ann is an **S4** agent, asks Ann whether she knows that p. Assume, moreover, that the exchange is subject to the rule that the speaker will not assert any proposition that she does not know to be true. Furthermore, the agents are fully cooperative and will provide the fullest information they have that is relevant to the question. In that case, Ann cannot give any honest yes/no answer to Bill's question. Bill can then deduce from the fact that Ann could not answer 'No' that Ann does not know that p. But if Bill can deduce it, why cannot Ann herself deduce it? In fact, why can't Ann ask this question herself, "Do I know that p?" and then, if she knows that she is an **S4** agent, she can derive, from her inability to answer 'No,' the fact that she does not know. Thus she becomes an **S5** agent. If she knows that she is an **S4** agent, this deduction requires only a minimal meta-reflection on herself. The conclusion seems to be that an **S4**, but not **S5** agent, cannot know that she is an **S4** agent.[4]

To be sure, the crux of the puzzlement lies in assumption (ii), where we made an attempt to empower our agent with the ability of accessing the meta-theoretic information of their own epistemic type – the kind of information that an external theorist normally possesses. However our puzzle suggests that our **S4** agent simply cannot consume such information *on pain of incoherence*.

The lesson we learn from this puzzle seems to be that there is something distinctive about the perspective of first-person agents and that of a third-person theorist/onlooker in epistemic models – such first/third-person asymmetry may create a divide that is unbridgeable even for someone as capable as a **S4** agent, let alone for those who are less capable than **S4**, epistemically speaking.

4. Thanks are due to Haim Gaifman for suggesting to me this way of formulating the puzzle.

Now, to return to our critic's question as to why the players themselves cannot be the theorist who comes to theorise the game situation at hand with a detailed modelling of their own epistemic scope, we would like to follow up with a further question: What does it even mean for an agent to acquire this type of meta-theoretic information about their own epistemic structure at all, and, more importantly, *how*?

3.2. All-inclusive states

Luckily, the players modelled in Aumann's original agreeing-to-disagree argument are all **S5** agents, so his system is – at least on the face of it – not plagued by the puzzle we just described where the protagonist is a **S4** agent. However, this does not mean that the question raised above about the contrast between the perspective of the players and that of a theorist can be automatically dispensed with in such models. Besides, what is so special about **S5** agents after all? What enables them to lift themselves to the level of an external theorist?

In a later work, Aumann (1987) described the asymmetry of different epistemic viewpoints in terms of the tension between the "Bayesian" (first-person) and the "game-theoretic" (third-person) views of the world. He maintains that these two perspectives can be coherently integrated in an analytic model. At the core of Aumann's proposal lies the assumption of so-called "all-inclusive" states of the world. This concept has since then been widely adopted in game-theoretic analyses.[5] Here is an explicit characterisation of the nature of an all-inclusive state given by Geanakoplos (1992):

> A "state of the world" is very detailed. It specifies the physical universe, past, present, and *future*; it describes what every agent knows, and what every agent knows about what every agent knows, and so on; it specifies what every agent does, and what every agent thinks about what every agent does, and what every agent thinks about what every agent thinks about what every agent does, and so on; it specifies the utility to every agent of every action, not only of those that are taken in that state of nature, but also those that *hypothetically* might have been taken, and it specifies what everybody thinks about the utility to everybody else of every possible action, and so on; it specifies not only what agents know, but what probability they assign to every event, and what probability they assign to every other agent assigning some probability to each event, and so on. (p. 57, emphasis added)

5. Aumann and Brandenburger (1995) adopted a more refined approach to the problem where each player's belief about other players' actions and beliefs are explicitly represented by the notions of *conjectures* and *theories* respectively. Both concepts are constituent components of the player's *types*, an idea originated by Harsanyi (1968), which essentially plays the same role as all-inclusive states with perhaps less informative contents about the physical world. But, at any rate, each type profile (a state) includes a description of actions of all players and it is further assumed that there is a common prior defined over all states. This implies that the players have prior probabilistic judgments over their own actions. This, however, is the recipe for another heated debate in the foundations of Bayesian probability/decision theory on action credences. For a discussion see Liu and Price (2019).

In other words, a state is taken to be a *complete* description of the world, which includes not only the information about the actions each player carries out and their mutual beliefs about what the others believe and are likely to do, which are usually direct targets of game-theoretic modelling; it contains also meta-theoretic information such as players' probabilities judgements over *all* the states, their criteria for decision making (including to what extend the other players are being rational), as well as their information structures over the states. The slogan is 'conditioning on one particular state, everybody knows everything!'

Alas, dear readers, it seems that we have now reached an impasse of having to untangle yet another creature of so-called 'all-inclusive states,' which, by definition, is all-knowing, all-wise, and all-powerful. What can we say about *that*?

3.3. Shadows remain

In this essay we told two tales about epistemic models with the goal of gaining insight of how certain ways of modelling knowledge/beliefs are possible, if at all. Both stories uncovered an important feature of first/third-person asymmetry in epistemic modelling. The lesson we learned seems to be that, unless we take a leap of faith in believing that for finite beings like us who stand the chance of accessing and comprehending 'all-inclusive states,' then, from agents' first-person perspectives there is always something that remains as mystery to themselves. It is reminiscent that under the sun everything leaves a shadow one way or another, except, of course, for the almighty sun itself.

Acknowledgement

I am grateful to Jessica Collins and Haim Gaifman for helpful feedback on early versions of this paper, and to three anonymous referees for many astute comments. This publication was made possible through the support of the Leverhulme Trust (ECF-2018-305) and the Isaac Newton Trust.

References

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55(1):1–18.

Aumann, R. J. and Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63(5):1161–1180.

Geanakoplos, J. (1992). Common knowledge. *The Journal of Economic Perspectives*, 6(4):53–82.

Harsanyi, J. (1967-1968). Games with incomplete information played by "Bayesian" players, I-III. *Management science*, 14(3):159–182, 320–334, 486–502.

Harsanyi, J. C. (1982a). Rejoinder to professors Kadane and Larkey. *Management Science*, 28(2):124–125.

Harsanyi, J. C. (1982b). Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Science*, 28(2):120–124.

Hintikka, J. (1962). *Knowledge and Belief: an introduction to the logic of the two notions*. Cornell University Press, Ithaca, New York.

Kadane, J. and Larkey, P. (1982a). Subjective probability and the theory of games. *Management Science*, 28(2):113–120.

Kadane, J. B. and Larkey, P. D. (1982b). Reply to professor Harsanyi. *Management Science*, 28(2):124.

Liu, Y. (2017). The sure-thing principle and P2. *Economics Letters*, 159:221–223.

Liu, Y. and Price, H. (2019). Heart of DARCness. *Australasian Journal of Philosophy*, 97(1):136–150.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press, Cambridge, Mass.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc., New York, second revised edition.