Games and Economic Behavior 132 (2022) 390–399

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

The impossibility of agreeing to disagree: An extension of the sure-thing principle ☆

Dov Samet

Coller School of Management, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Received 15 June 2021
Available online xxxx

JEL classification:
C70
D82

Keywords:
Sure thing principle
Agreeing to disagree
Common knowledge

ABSTRACT

The *impossibility of agreeing to disagree* in the non-probabilistic setup means that agents cannot commonly know their decisions unless they are all the same. We study the relation of this property to the *sure thing principle* when it is expressed in epistemic terms. We show that it can be presented in two equivalent ways: one is in terms of knowledge operators, which we call the principle of *follow the knowledgeable*, the other is in terms of *kens*—bodies of agents' knowledge—which we call *independence of irrelevant knowledge*. The latter can be easily extended to a property which is equivalent to the impossibility of agreeing to disagree.

1. Introduction

The *impossibility of agreeing to disagree* (IAD) for a group of agents was introduced by Aumann (1976) in a probabilistic setting. This property means that the agents cannot have common knowledge (agreeing) of their posteriors of a given event when these posteriors are not the same (disagree). Aumann showed that IAD holds when the agents have a common prior. Here we study IAD in a non-probabilistic setting where the probability ascribed by an agent to an event is replaced by a decision made by the agent. We formulate IAD for two agents, Adam and Eve, and discuss the obvious extension for any group of agents at the end of the introduction.

IAD: If Adam's and Eve's decisions are common knowledge between them, then these decisions are the same.

We will show that the non-probabilistic IAD is an extension of the *sure-thing principle* (STP) as described in epistemic terms in an example in Savage (1954). For this we define formally the STP in epistemic terms in two different ways — one as an embodiment of the maxim *follow the knowledgeable* (FTK), and the other as reflecting the idea of *independence of irrelevant knowledge* (IIK) for the purpose of making one's decision. We show that the two definitions of the STP are equivalent. The STP concerns an agent in two periods, today and tomorrow, who can be formally considered as two distinct agents. Obviously, these two agents are not symmetric in the sense that the agent of tomorrow is more knowledgable than the agent of today. However, the second formulation of the STP, as IIK, can be easily extended to cover symmetric situations as well. This extension is equivalent to IAD.

☆ Financial support of ISF Grant #722/18 is acknowledged.

E-mail address: dovsamet@gmail.com.

https://doi.org/10.1016/j.geb.2022.01.016
0899-8256/© 2022 Published by Elsevier Inc.

1.1. Savage's example of the sure-thing principle

The *sure-thing principle* (STP) was introduced by Savage (1954) using the following story.

> 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* that the Republican candidate were going to win, and again 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. It is all too seldom that a decision can be arrived at on the basis of this principle, but except possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance. (emphasis added)

This story is told in terms of knowledge and decisions. However, when it came to describing the STP formally, Savage explains why he had to give up the epistemic nature of the story:

> The sure-thing principle cannot appropriately be accepted as a postulate in the sense that P1 is, because *it would introduce new undefined technical terms referring to knowledge and possibility that would render it mathematically useless without still more postulates governing these terms.* It will be preferable to regard the principle as a loose one that suggests certain formal postulates well articulated with P1. (emphasis added)

Thus, Savage did not consider his second postulate, P2, to be a formalization of the STP. Such formalization requires, as he says, a specification of the postulates governing the terms knowledge and possibility, that is, a formal model of knowledge which was not available at the time. Almost a decade after Savage's *The foundations of statistics*, Hintikka (1962) introduced formal modeling of knowledge, syntactic and semantic, in his *Knowledge and belief*, while a semantic multi-agent model of knowledge was introduced more than a decade later in Aumann (1976).

Here, we use Aumann's formal model of multi-agent knowledge, which became a standard, for the formulation of the sure-thing principle that reflects its original intended meaning. We first provide two formalizations of the STP and show their equivalence.

1.2. A temporal formulation

The simplest way to read Savage's story concerns an agent in two periods. In the second, the agent will be more knowledgeable, as he will know whether the Democrat candidate won the election ("D") or the Republican ("R"). As Savage puts it, "he considers the outcome of the next presidential election relevant" to the question of buying a certain property. Since the agent knows in the first period that being more knowledgeable in the second he will buy the property in either case, he can buy it even in the first period at a time that he does not yet know who the winner is.[1] The following is a rough draft of the STP that we call *follow the knowledgeable* (FTK):

> **FTK**: If an agent knows today her decision of tomorrow, when she is more knowledgeable, then this should be her decision today.[2]

The following formulation of the STP reflects the idea that the agent's decision is determined by her *body of knowledge*, that is, all the things the agent knows. At this stage we use these notions intuitively and do not specify formally what the 'things' that the agent knows are. For the sake of brevity we call the body of knowledge of an agent *ken*. In the toy example used by Savage to describe STP, tomorrow the agent may have one of two kens. If the Democrat wins then the agent's ken will include this fact, denoted "D" for short, and if the Republican wins tomorrow's ken will include "R". Both kens include the fact "D or R", of course. The agent's ken of today includes only the *intersection* of tomorrow's two kens, namely, "D or R". Now, the businessman in Savage's story "considers the outcome of the next presidential election relevant" to his decision. But to be sure, the very essence of the STP demonstrated in this story is that neither "R" nor "D" are relevant to the agent's decision to buy the property. What remains relevant is just the knowledge common to both kens, namely "D or R", which is what the agent knows today.

This suggests the following version of the STP, which we call *independence of irrelevant knowledge* (IIK).

[1] One should not confuse the STP with the good advice "never put off until tomorrow what you can do today". The STP does not assume any gains or losses from the timing of the decision itself.

[2] The principle of *reflection*, introduced by van Fraassen (1984), is a probabilistic version of FTK. It says that an agent's present belief in a proposition p, given the proposition that in the future her belief in p is α, is α. An extreme case of this principle is when the future is the present. In this case the requirement is that an agent's belief in a proposition p, given that her belief in p is α, is α. Samet (2000) called this principle *the conditioning axiom*, and showed that it is equivalent to requiring that the agent is certain that she is introspective.

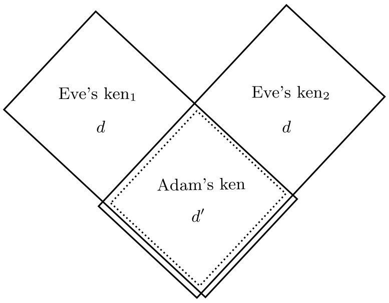

Fig. 1. Independence of irrelevant knowledge.

> **IIK**: If the agent's decision is the same for all of her tomorrow's kens, then what is relevant for her decision is the intersection of all these kens, namely her ken today, and therefore this should be her decision today.[3]

In the temporal formulation the agent is split into two agents: the agent today and the agent tomorrow. The STP is based on two implicit, natural assumptions concerning the relation between these two agents. First, we assume that the two agents share the same interests. In other words, the agent's preferences do not change over time. This assumption is essential because without it knowledge of the decision made tomorrow should not imply that this decision is appropriate today. Second, we assume that the agent tomorrow is more knowledgeable than the agent today, and, moreover, that the agent today knows that.[4]

1.3. A two-agent formulation

We now give FTK and IIK an *atemporal* formulation with two agents, Adam and Eve. The two assumptions made in the temporal formulation should be made in the atemporal formulation. First, it is essential to assume that Adam and Eve share the same interests in order to be able to infer from the decision of one agent what the other agent's decision should be. This would be the case, for example, if they are doctors who are called to make a decision concerning a patient whose interest is also theirs. As for the second assumption, we spell out explicitly the requirement that Eve is more knowledgeable than Adam, and that Adam knows this.

The first formulation of the STP in the atemporal setup is:

> **FTK** (two-agent version): If Adam knows that everything that he knows Eve knows, and he also knows Eve's decision, then this must be his decision too.

The second atemporal formulation of STP is:

> **IIK** (two-agent version): If Adam's ken is the intersection of some of Eve's kens, and Eve's decision is the same for all these kens, then what is relevant for her decision is the intersection of all these kens of hers, namely, Adam's ken, and therefore this should also be Adam's decision.

An example of the application of IIK is graphically depicted in Fig. 1 by a Venn diagram.

Our first result states:

> FTK *and* IIK *are equivalent.*

1.4. Extending the STP

Adam and Eve play asymmetric roles in IIK. While there is a *single* ken for Adam, there is a *family* of Eve's kens. But the same reasoning can be applied to cases where the agents have a symmetric role and each of them is endowed with a family of kens. We assume that if within a family of kens of an agent the decision is the same for all kens, then what is

[3] IIK has features in common with various versions of the independence of irrelevant alternatives. It also resembles an ancient legal syllogism from the mishna called "binyan av" (prototype) or "hatsad hashaveh" (the common characteristic): If X and Y are two sets of circumstances, and in each the ruling is R, then the circumstances relevant to this ruling are those in $X \cap Y$, and the ruling R also applies to any set of circumstances that includes this intersection. (See an instance of this syllogism in the Babylonian Talmud, Tractate Sanhedrin 64b.) IIK is more restricted. It applies only to a ken which is the intersection of other kens, and not to kens that contain this ken.

[4] When we say that the agent is *more* knowledgeable in later periods, we are being a little bit sloppy for the sake of brevity. We mean that in later periods he is *at least* as knowledgable.

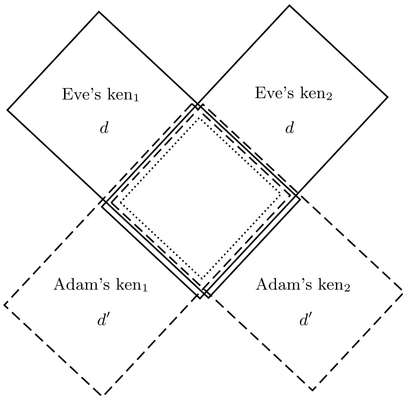

Fig. 2. Extended IIK.

relevant for making this decision is the intersection of kens in the family even when this intersection is not itself a ken of any of the agents. This implies the following principle — extended independence of irrelevant knowledge (EIIK):

> **EIIK**: If within a family of Adam's kens Adam's decision is the same for all kens, and within a family of Eve's kens Eve's decision is the same for all kens, and if moreover, the intersection of the kens in Adam's family of kens is the same as the intersection of kens in Eve's family of kens, then Adam and Eve make the same decision.

An example of the application of extended IIK is graphically depicted in Fig. 2 by a Ven diagram.

We note that despite the similarity between extended IIK and IIK, there is a crucial difference between them. The two agents in IIK can be viewed as the same agent in different periods. The asymmetry between the two agents in the IIK is essential for this interpretation and reflects the asymmetry of time with respect to knowledge. No such interpretation of the two agent is possible for EIIK.

Our second result states:

> EIIK *is equivalent to* IAD.

1.5. Formalization

We now describe the choices we make in formalizing knowledge. We first choose the objects that are subject to the agents' knowledge in the standard way knowledge is modeled in economic theory. These objects are subsets, to which we refer as *events*, of a given set of states.[5] The knowledge of each agent varies with states. There are two ways to describe the knowledge of an agent:

- *Kens*: Specify for each state and agent the set of all the events that the agent knows at the state. Such a set of events is the formal rendering of the ken mentioned before.
- *Knowledge operators*: Define for each agent i, a knowledge operator $\mathrm{K_i}$. For each event E, the event $\mathrm{K_i}E$, which reads 'i knows E', consists of all the states at which i knows E.

In such a model we are able to present formally the four notions: IIK, extended IIK, FTK, and IAD. The first two notions are described in terms of kens which are now defined rigourously. The last two notions are described in terms of knowledge operators which are analogous to the words "Adam knows" and "Eve knows" in the informal description of these notions.

To prove the equivalences stated above it is required that kens and knowledge operators have certain properties and are not arbitrarily defined. For this we adopt the most familiar and most commonly used model, the partition model of multi-agent knowledge of Aumann (1976).[6] In this model each agent i is equipped with a partition Π_i of the state space. A *possibility function* π_i assigns to each state ω an event $\pi_i(\omega)$ which is the element of Π_i that contains ω. The ken of i at a state ω consists of all the supersets of $\pi_i(\omega)$. Thus, the event $\mathrm{K_i}E$ consists of the all the states ω such that $\pi_i(\omega) \subseteq E$.

Note that the informal, intuitive discussion of FTK and IIK in subsections 1.2–1.4 is presented in a way that makes it accessible to an audience, like Savage himself, not acquainted with formal theories of knowledge. That is why this discussion

[5] Alternatively, we could choose the objects that are known to be sentences in a formal language (Aumann, 1999). Such a formalization is closer to the informal reasoning about knowledge, but would require longer exposition and proofs. Students of economic theory are also less familiar with it. In other formalizations the objects that are known are elements of a Boolean algebra (Blackburn et al., 2001; Halpern et al., 2009), or the elements of a set of proposition on which one operation, called negation, is defined (Samet, 1990).

[6] The results also hold for non-partitional models of knowledge that are discussed in subsection 5.3.

$$\begin{tikzcd}
 & \underline{\text{kens}} & & \underline{\text{operators}} \\
\text{STP:} & \textbf{IIK} \arrow[rr, Leftrightarrow] & & \textbf{FTK} \\
\text{extended STP:} & \textbf{extended IIK} \arrow[u, Rightarrow] \arrow[rr, Leftrightarrow] & & \textbf{IAD} \arrow[u, Rightarrow]
\end{tikzcd}$$

Fig. 3. Four versions of STP and their relations.

does not involve states, partitions, or possibility functions. They are used, though, in our formal model to define kens and knowledge operators. But, once these are defined we can formally express all four versions of STP in a way that resembles the informal description of these conditions without mentioning states and partitions. Nevertheless, we formulate the STP also in terms of states and possibility functions and show that this version of the STP is also equivalent to FTK and IIK. We also formulate an extension of STP in terms of states and possibility functions, and it is equivalent to EIIK and IAD.

The diagram in Fig. 3 summarizes the relations between the four versions of the STP. The asymmetric role of the agents in IIK, which is expressed in terms of kens, is easily and naturally symmetrized and generalized by extended IIK. In contrast, FTK, which is expressed in terms of knowledge operators, is not easily symmetrized. It turns out that the symmetric version of FTK, in terms of knowledge operators, is IAD. The partitional versions of the STP are equivalent to the notions defined in terms of kens and knowledge operators.

1.6. The multi-agent case

Extended IIK is formulated for two agents, but it can be formulated for a group of agents of any size. Suppose a family of kens is given for each agent in the group and within each such family the decision is the same for all the kens in the family. We can require in this case, in the spirit of IIK, that if the intersection of the kens in each family is the same for all agents, then the same decision is made in each family. However, if we require that extended IIK holds for each pair of agents in the group, then it must hold for the whole group, as any pair must make the same decision in their family of kens.

Similarly, IAD is formulated for a pair of agents, but can be formulated for any group of agents by requiring that if their decisions are common knowledge between them, then the decisions must be the same. Again, if we require IAD for any pair in the group, then it holds for the whole group. Indeed, if the decisions of the agents are common knowledge between them, then the decisions of any pair of agents are common knowledge between this pair of agents, and hence are the same. For these reasons, we formulate our results in a model of two agents only.

2. Models of knowledge and decisions

Since the various definitions of the STP and their extensions involve only two agents we consider models of knowledge for two agents denoted 1 and 2. A *knowledge model* (Ω, Π_1, Π_2) consists of a set of *states* Ω, and for each agent $i \in \{1, 2\}$, a partition Π_i of Ω. We define for each i a function, called a *possibility function*, $\pi_i : \Omega \to 2^{\Omega}$ such that $\pi_i(\omega)$ is the element of the partition Π_i that contains ω. The name "possibility function" expresses the interpretation of $\pi_i(\omega)$ as the set of states that i considers possible at ω.

Subsets of Ω are called *events*. We say that i *knows event* E at ω if $\pi_i(\omega) \subseteq E$. The event that *i knows* E, denoted $\mathrm{K}_{\mathrm{i}}E$ is the set of all states in which i knows E, that is, $\mathrm{K}_{\mathrm{i}}E = \{\omega \mid \pi_i(\omega) \subseteq E\}$. Thus, $\mathrm{K}_{\mathrm{i}}E$ is the union of all the elements of Π_i that are contained in E. The function $\mathrm{K}_{\mathrm{i}} : 2^{\Omega} \to 2^{\Omega}$, thus defined, is called i's *knowledge operator*.

Denote for each E, $\mathrm{K}E = \mathrm{K}_1 E \cap \mathrm{K}_2 E$, that is, $\mathrm{K}E$ is the event that both agents know E. The event that all know E and all know that all know E and so on is denoted by $\mathrm{C}E$, that is, $\mathrm{C}E = \cap_{m=1}^{\infty} \mathrm{K}^m E$. The operator C is called the *common knowledge* operator. The knowledge operators enable us to describe formally the property FTK without direct reference to states, partitions, or possibility functions, exactly as in the informal discussion of FTK in the introduction.

We can now describe formally the notion of body of knowledge, or in its short name ken, which was introduced informally in the introduction and was used to describe IIK. Since the knowledge of an agent varies with the state, so does the agent's ken. The *ken* of agent i at ω is the set of all the events that i knows at ω, that is, $\mathrm{ken}_i(\omega) = \{E \mid \pi_i(\omega) \subseteq E\}$. Alternatively, $\mathrm{ken}_i(\omega) = \{E \mid \omega \in K_i(E)\}$.[7] We denote by Ken_i the family of all of i's kens, that is, $\mathrm{Ken}_i = \{\mathrm{ken}_i(\omega) \mid \omega \in \Omega\}$.

There is a one-to-one correspondence between the set of i's kens and the elements of the partition Π_i, as each element of Π_i defines a ken. Moreover, any relation between, or operation on, kens can be easily translated into a relation between, or operation on, the elements of the partition that define these kens. For example, consider the intersection of two kens of

[7] Kens were introduced in Samet (1990) for more abstract models of knowledge. Using formal epistemic language, a ken of agent i is defined as a maximal set of sentences, Φ, such that the set of sentences $\{i \text{ knows } f \mid f \in \Phi\}$ is consistent.

i, $\mathbb{K}_i^1 \cap \mathbb{K}_i^2$, where $\pi_i(\omega_1)$ and $\pi_i(\omega_2)$ are the elements of Π_i that define $\mathbb{K}_i^1$ and $\mathbb{K}_i^2$ respectively. Now, $E \in \mathbb{K}_i^1 \cap \mathbb{K}_i^2$ if and only if E is a superset of both $\pi_i(\omega_1)$ and $\pi_i(\omega_2)$. This is the case if and only if E is a superset of $\pi_i(\omega_1) \cup \pi_i(\omega_2)$. Thus, the operation of intersection of kens is translated into the union of the corresponding elements of the partition. This is stated in Claim 1 in Section 6.

Let D be a nonempty set of *decisions*. A *decision function* $\mathbf{d}_i$ for agent i associates a decision with each of i's kens. That is, $\mathbf{d}_i$ is a function $\mathbf{d}_i\colon \mathrm{Ken}_i \to D$. A pair of decision functions for the two agents $\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2)$ is called a *decision function profile*. With some abuse of notation we write $\mathbf{d}_i(\omega)$ for $\mathbf{d}_i(\mathrm{ken}_i(\omega))$. We denote by $[\mathbf{d}_i = d]$ the event that i's decision is d, namely $[\mathbf{d}_i = d] = \{\omega \mid \mathbf{d}_i(\omega) = d\}$.

3. The equivalence of FTK and IIK

The principle of "follow the knowledgeable", which we now formally define, spells out in precise terms the verbal description of this principle as given in subsection 1.3. We note that $\neg \mathrm{K_i} E \cup \mathrm{K_i K_j} E$ is the event that if i knows E then i knows that j knows E. Thus, $\bigcap_{E \subseteq \Omega} \neg \mathrm{K_i} E \cup \mathrm{K_i K_j} E$ is the event that i knows that j is at least as knowledgeable. Our first formulation of the STP requires that if i knows that j is at least as knowledgable, and happens to know j's decision, then this should also be i's decision.

Follow the knowledgeable (FTK):
FTK *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *if for all agents* i, j, *and decision* d,

$$\bigcap_{E \subseteq \Omega} \left(\neg \mathrm{K_i} E \cup \mathrm{K_i K_j} E\right) \cap \mathrm{K_i}[\mathbf{d}_j = d] \subseteq [\mathbf{d}_i = d]. \tag{1}$$

The independence of irrelevant knowledge below, IIK, is a formal rendering of the description given to it in subsection 1.3. Like the informal description the formal definition does not make direct use of partitions or states. IIK says that if j's decision is d at each ken in a family of her kens $\mathcal{K}_j$, and if i's ken is the intersection of j's kens in $\mathcal{K}_j$, then i's decision is d. More formally,

Independence of irrelevant knowledge (IIK):
IIK *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *when, for all agents* i, j, *decisions* d_i, d_j, $\mathbb{K}_i \in \mathrm{Ken}_i$, *and* $\mathcal{K}_j \subseteq \mathrm{Ken}_j$,

if
(1) $\mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$, *and*
(2) *for each* $\mathbb{K}_j \in \mathcal{K}_j$, $\mathbf{d}_j(\mathbb{K}_j) = d_j$, *and*
(3) $\mathbf{d}_i(\mathbb{K}_i) = d_i$
then $d_i = d_j$.

The STP has been formulated so far in terms of knowledge operators (FTK) and kens (IIK) without mentioning explicitly states, partitions, or possibility functions. We now introduce another version of STP purely in terms of states and possibility functions.

Possibility STP (PSTP):
PSTP *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *when for all agents* i, j, *decisions* d_i, d_j, *and state* ω,

if
(1) for each $\omega' \in \pi_i(\omega)$, $\pi_j(\omega') \subseteq \pi_i(\omega)$ *and*
(2) $\mathbf{d}_j(\omega') = d_j$, *and*
(3) $\mathbf{d}_i(\omega) = d_i$
then $d_i = d_j$.

We can now state the equivalence of all forms of the STP.

Theorem 1. *For a knowledge model with a decision function profile, the three conditions, follow the knowledgeable* (FTK), *independence of irrelevant knowledge* (IIK), *and the possibility sure thing principle* (PSTP)) *are equivalent.*

4. Extensions of the STP

The IIK version of the STP is extended below based on the following observation. Although the whole body of knowledge of an agent, namely her ken, determines her decision, it may be determined, and most probably is determined by only part of her ken which is relevant to her decision. Thus, if the same decision is made by i in all of her kens in some family of kens $\mathcal{K}_i \subseteq \text{Ken}_i$ we may assume that it is the intersection of the kens in $\mathcal{K}_i$ that determines this decision.

In the following extension of IIK 1 and 2 play a symmetric role. Both are endowed with a family of kens: $\mathcal{K}_1 \subseteq \text{Ken}_1$ and $\mathcal{K}_2 \subseteq \text{Ken}_2$. If in each family the same decision is associated with each ken, then the relevant knowledge for making this decision is the intersection of the kens in each family. Thus, if the intersection of the kens in $\mathcal{K}_1$ coincides with the intersection of kens in $\mathcal{K}_2$ the same decision should be made by both agents.

Extended independence of irrelevant knowledge (EIIK):
EIIK *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *when for all decisions* d_1, d_1, *and non-empty sets of kens,* $\mathcal{K}_1 \subseteq \text{Ken}_1$, *and* $\mathcal{K}_2 \subseteq \text{Ken}_2$,

if
(1) $\cap_{\mathbb{K}_1 \in \mathcal{K}_1} \mathbb{K}_1 = \cap_{\mathbb{K}_2 \in \mathcal{K}_2} \mathbb{K}_2$, *and*
(2) *for each* $\mathbb{K}_1 \in \mathcal{K}_1$, $\mathbf{d}_1(\mathbb{K}_1) = d_1$, *and*
(3) *for each* $\mathbb{K}_2 \in \mathcal{K}_2$, $\mathbf{d}_2(\mathbb{K}_2) = d_2$,
then, $d_1 = d_2$.

Obviously, IIK is a special case of EIIK, where one of the families of kens $\mathcal{K}_1$, $\mathcal{K}_2$ is a singleton.

We next formulate the condition of *impossibility of agreeing to disagree* (IAD), which is equivalent to EIIK, but is expressed, like FTK, in terms of knowledge operators rather than kens. The condition requires that the agents cannot have common knowledge of their decisions when the decisions are not the same. It is analogous to a condition for models of knowledge and probabilistic beliefs, that agents cannot have common knowledge of their posterior probability of a given event when the two posterior probabilities differ. Here we show that IAD is equivalent to EIIK, which is an extended version of the STP.

Impossibility of agreeing to disagree (IAD):
IAD *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *if for all* d_1 *and* d_2,

$$\mathsf{C}\big([\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]\big) \subseteq [\mathbf{d}_1 = \mathbf{d}_2].$$

Equivalently, IAD holds if when $C\big([\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]\big)$ *is not empty, then* $d_1 = d_2$.

Extended IIK is formulated in terms of kens, and IAD in terms of knowledge operators, both without explicitly mentioning states, partitions or possibility functions. The next condition is formulated in terms of states and possibility functions.

Extended possibility STP (EPSTP):
EPSTP *holds for the knowledge model and the decision function profile* $\mathbf{d}$ *when for all decisions* d_1, d_2, *and event* E,

if
(1) *for each* $\omega \in E$, $\pi_1(\omega) \subseteq E$, *and*
(2) $\mathbf{d}_1(\omega) = d_1$, *and*
(3) *for each* $\omega \in E$, $\pi_2(\omega) \subseteq E$ *and*
(4) $\mathbf{d}_2(\omega) = d_2$,
then $d_1 = d_2$.

Obviously, PSTP is a special case of EPSTP where E is either $\pi_1(\omega)$ or $\pi_2(\omega)$.

Theorem 2. *For a knowledge model with a decision function profile the three conditions, extended independence of irrelevant knowledge* (EIIK)*, the impossibility of agreeing to disagree* (IAD)*, and the extended possibility sure thing principle* (EPSTP) *are equivalent.*

5. Discussion

5.1. Early attempts

The first attempt to use the sure-thing principle in an epistemic setup was made, independently, by Cave (1983) and Bacharach (1985), although it was the latter who used the term STP in this context. Both papers proposed a sufficient condition for IAD, generalizing the probabilistic agreement theorem of Aumann (1976) to the non-probabilistic cases, where at each state of the world a *decision* of each agent is specified, rather than a *posterior probability*.

Both papers use a partition model with a *virtual decision function* δ (a term suggested in Samet (2008)) from which individual decisions are derived. Such a function assigns a decision to *each* event. The interpretation is that the decision $\delta(E)$ associated with an event E is the decision made when knowledge is given by E. This is very much in the spirit of

Savage's example and the approach adopted here. The sure-thing principle in this setup says that for two disjoint events E and F for which $\delta(E)=\delta(F)$, it is the case that $\delta(E\cup F)=\delta(E)$.

Virtual decision functions are hard to interpret properly. Considering events which are not elements of the partition as describing knowledge is incongruent with the knowledge structure given by the partition. Moreover, by its very essence the STP cannot be applied to a single knower. The union $E\cup F$ purports to represent a body of knowledge—a ken—which is the intersection of the kens given by E and by F. But this idea is inconsistent with partition models: it is impossible for an agent in a partition model to have kens with an intersection that is also a ken of the *same agent*, except for the trivial case that the intersecting kens are identical. The only way to express the STP is through either the knowledge of an agent in *two* periods, or alternatively, as is the case here, the knowledge of *different* agents. Moses and Nachum (1990) were the first to study conceptual difficulties regarding virtual decision functions. In contrast, here, the decision function $\mathbf{d}_i$ of an agent i is defined only on the kens of i, or equivalently, on elements of i's partition. It is not defined on other events not even unions of the partition's elements.

5.2. A sufficient condition for IAD

Samet (2008) found a sufficient condition for IAD in terms of the STP, avoiding the conceptual pitfalls in Cave (1983) and Bacharach (1985). For this purpose, Samet (2008) formulated the STP in terms of knowledge operators under the name *interpersonal sure-thing principle* (ISTP). This condition requires that

$$\bigcap_{E\subseteq\Omega} \mathrm{K_i}\big(\neg\mathrm{K_i}E\cup\mathrm{K_j}E\big)\cap\mathrm{K_i}[\mathbf{d}_j=d]\subseteq[\mathbf{d}_i=d].$$

The description in FTK of the event that i knows that j is more knowledgeable is a little bit simpler than it is in ISTP, but the two events are the same.[8] Thus, FTK and ISTP are equivalent. However, FTK (or equivalently, ISTP) does not imply IAD, as is shown here, because FTK is equivalent to IIK, which is weaker than extended IIK, and hence weaker than IAD. In order to find a sufficient condition for IAD in terms of FTK, Samet (2008) had to fortify ISTP. The condition that implies IAD is ISTP-*expandability*, which requires that FTK also holds when another agent is added to the model who is less knowledgeable than all the other agents. ISTP-expandability is a strong condition that *implies* IAD but is not implied by it. In contrast, extended IIK is *equivalent* to IAD. Thus, IAD is not only an *implication* of some condition that involves the sure-thing principle, but rather a generalized version of the sure-thing principle itself.

5.3. Non-partitional models of knowledge

In the partition model we adopted here, knowledge has the following properties[9]:

(1) *Truth*: Whatever is known is true: for each E, $\mathrm{K_i}E\subseteq E$.
(2) *Positive introspection*: If the agent knows an event she knows that she knows it: for each E, $\mathrm{K_i}E\subseteq\mathrm{K_i}\mathrm{K_i}E$.
(3) *Negative introspection*: If the agent does not know an event, then she knows that she does not know it: for each E, $\neg\mathrm{K_i}E\subseteq\mathrm{K_i}\neg\mathrm{K_i}E$.

We now address the question which of these properties are needed to establish the equivalences in Theorems 1 and 2. For this we need to consider non-partitional models of knowledge.

Kens and knowledge operators were defined in Section 2 in terms of the possibility functions π_i which in turn were defined in terms of the partitions Π_i. In order to construct non-partition models of knowledge we start with possibility functions as primitive.[10] The three properties of knowledge listed above are equivalent to the following requirement from the possibility functions.[11]

(1) *Truth*: For each ω, $\omega\in\pi_i(\omega)$.[12]
(2) *Positive introspection*: For each ω, if $\omega'\in\pi_i(\omega)$ then $\pi_i(\omega')\subseteq\pi_i(\omega)$.[13]

[8] In ISTP the requirement is that for each E, "i knows that if she knows E then j knows E", while in FTK the requirement is that for each E, "if i knows E then she knows that j knows E." For partition models these two events are the same. Indeed, $\mathrm{K_i}\big(\neg\mathrm{K_i}E\cup\mathrm{K_j}E\big)$ is the union of all the elements of Π_i contained in $\neg\mathrm{K_i}E\cup\mathrm{K_j}E$. But $\neg\mathrm{K_i}E$ is a union of elements of Π_i, and the union of element of Π_i in $\mathrm{K_j}E$ is $\mathrm{K_i}\mathrm{K_j}E$. Thus, $\mathrm{K_i}\big(\neg\mathrm{K_i}E\cup\mathrm{K_j}E\big)=\neg\mathrm{K_i}E\cup\mathrm{K_i}\mathrm{K_j}E$. In non-partitional models the two events are not necessarily the same, and $\mathrm{K_i}\big(\neg\mathrm{K_i}E\cup\mathrm{K_j}E\big)\subseteq\neg\mathrm{K_i}E\cup\mathrm{K_i}\mathrm{K_j}E$.

[9] See, for example, Aumann (1999).

[10] Another, equivalent way to define the model is by using a binary relation on Ω for each i, called *accessibility*. When ω' is accessible from ω for i, then i considers ω' a possible state at state ω. The accessibility relation can be equivalently described by the possibility function $\pi_i(\omega)=\{\omega'\mid\omega'\text{ is accessible from }\omega\text{ for }i\}$.

[11] See, for example, (Halpern, 2003, p. 190).

[12] In terms of the accessibility relation of i this is equivalent to requiring that the relation is reflexive.

[13] In terms of the accessibility relation of i this is equivalent to requiring that the relation is transitive.

(3) *Negative introspection*: For each ω, if $\omega' \in \pi_i(\omega)$, then $\pi_i(\omega) \subseteq \pi_i(\omega')$.[14]

It turns out that Theorems 1 and 2 hold true also for models that satisfy only truth and positive introspection. Indeed, the proof of Theorem 1 makes use only of the first two properties of the possibility functions. The proof of Theorem 2 makes use of the partition structure in describing the common knowledge operator C as the knowledge operator defined by the meet partition. However, even in models that satisfy only truth and positive introspection, C can be shown to be a knowledge operator. We say that an event E is *evident* if for each $\omega \in E$, $\pi_i(\omega) \subseteq E$ for each i. We now define $\pi(\omega)$ to be the intersection of all evident events that contain ω. This possibility function satisfies truth and positive introspection and it defines the operator C. With this definition Theorem 2 can be proved for knowledge models that satisfy just truth and positive introspection.

6. Proofs

A ken $\mathbb{K}_j$ of j is the set of all supersets of some element in π_j. Denote this element by $\pi_j(\mathbb{K}_j)$.

Claim 1. *For any j and $\mathcal{K}_j \subseteq \mathrm{Ken}_j$, $\cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$ consists of all the supersets of $\cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j)$.*

Proof. By the definition of a ken, $E \in \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$ if and only if $\pi_j(\mathbb{K}_j) \subseteq E$ for each $\mathbb{K}_j \in \mathcal{K}_j$. This holds if and only if $\cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j) \subseteq E$. □

Corollary 1. *For each ω, i, j and $\mathcal{K}_j \subseteq \mathrm{Ken}_j$, $\mathrm{ken}_i(\omega) = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$ if and only if $\pi_i(\omega) = \cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j)$.*

Proof. The second equality holds if and only if the events in the two sides of the equation have the same family of supersets. By the definition of ken and Claim 1 this is the first equality. □

Claim 2. *For each i, j, and E, $E = \cup_{\omega \in E} \pi_j(\omega)$ if and only if for each $\omega \in E$, $\pi_j(\omega) \subseteq E$.*

Proof. Obviously if the equality holds, then for each $\omega \in E$, $\pi_j(\omega) \subseteq E$. Conversely, if this condition holds, then $\cup_{\omega \in E} \pi_j(\omega) \subseteq E$, and for the opposite inclusion we note that for each $\omega \in E$, $\omega \in \pi_j(\omega) \subseteq \cup_{\omega \in E} \pi_j(\omega)$.[15] □

Lemma 1. *For each i, j, and ω, the following conditions are equivalent.*

(a) $\omega \in \cap_{E \subseteq \Omega} \neg \mathrm{K_i} E \cup \mathrm{K_i K_j} E$.
(b) *For each $\omega' \in \pi_i(\omega)$, $\pi_j(\omega') \subseteq \pi_i(\omega)$.*
(c) $\pi_i(\omega) = \cup_{\omega' \in \pi_i(\omega)} \pi_j(\omega')$.
(d) *For some $\mathcal{K}_j \subseteq \mathrm{Ken}_i$, $\mathrm{ken}_i(\omega) = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$.*

Proof. To see that (a) implies (b), note that for $E = \pi_i(\omega)$, (a) implies $\omega \in \neg \mathrm{K_i} \pi_i(\omega) \cup \mathrm{K_i K_j} \pi_i(\omega)$. But $\omega \in \pi_i(\omega) = \mathrm{K_i} \pi_i(\omega)$, and therefore $\omega \in \mathrm{K_i K_j} \pi_i(\omega)$ Thus, $\pi_i(\omega) \subseteq \mathrm{K_j} \pi_i(\omega)$. Hence for each $\omega' \in \pi_i(\omega)$, $\omega' \in \mathrm{K_j} \pi_i(\omega)$, which means that $\pi_j(\omega') \subseteq \pi_i(\omega)$.

Now, (b) implies (c) by Claim 2, for $E = \pi_i(\omega)$.

Define $\mathcal{K}_j = \{\mathrm{ken}_j(\omega') \mid \omega' \in \pi_i(\omega)\}$. If (c) holds, then $\pi_i(\omega) = \cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j)$. Thus (d) holds by Corollary 1.

When (d) holds then by Corollary 1,

$$\pi_i(\omega) = \cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j). \tag{2}$$

To show that (d) implies (a) we need to show that if equation (2) holds and $\omega \in \mathrm{K_i} E$, then $\omega \in \mathrm{K_i K_j} E$. Indeed, if $\omega \in \mathrm{K_i} E$, then $\pi_i(\omega) \subseteq E$. Now, by equation (2), for each $\omega' \in \pi_i(\omega)$, $\omega' \in \pi_j(\mathbb{K}_j)$ for some $\mathbb{K}_j \in \mathcal{K}_j$ and thus, $\pi_j(\omega') \subseteq \pi_j(\mathbb{K}_j)$.[16] But $\pi_j(\mathbb{K}_j) \subseteq \pi_i(\omega) \subseteq E$, and hence, $\pi_j(\omega') \subseteq E$. This means that $\omega' \in \mathrm{K_j} E$. This is true for all $\omega' \in \pi_i(\omega)$, and therefore $\pi_i(\omega) \subseteq \mathrm{K_j} E$. Thus $\omega \in \mathrm{K_i K_j} E$. □

[14] In terms of the accessibility relation of i this is equivalent to requiring that the relation is Euclidean, namely if ω' and ω'' are accessible from ω, then ω'' is accessible from ω'.

[15] Here we used the truth property of π_i mentioned in subsection 5.3.

[16] In the partition model $\pi_j(\omega') = \pi_j(\mathbb{K}_j)$. But all we need for the proof is just the inclusion, which is satisfied when positive introspection holds, as discussed in subsection 5.3.

Proof of Theorem 1.

FTK ⇒ IIK: Suppose that the knowledge space with $(\mathbf{d}_1, \mathbf{d}_2)$ satisfies FTK, and that the antecedents (1)-(3) in the definition of IIK hold. Let ω be the state for which $\mathbb{K}_i$ in (1) is $\text{ken}_i(\omega)$. As (d) implies (a), in Lemma 1, $\omega \in \cap_{E\subseteq\Omega} \neg \mathrm{K_i} E \cup \mathrm{K_i K_j} E$. By Corollary 1, (1) implies $\pi_i(\omega) = \cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j)$. By (2) it follows that for each $\omega' \in \pi_i(\omega)$, $\mathbf{d}_j(\omega') = d_j$. Thus, $\pi(\omega) \subseteq [\mathbf{d}_j = d_j]$ and therefore $\omega \in K_i([\mathbf{d}_j = d_j])$. By FTK, $\omega \in [\mathbf{d}_i = \mathbf{d}_j]$, which implies together with (3) that $d_i = \mathbf{d}_i(\text{ken}_i(\omega)) = \mathbf{d}_i(\omega) = d_j$.

IIK ⇒ PSTP: Suppose that IIK holds, and the antecedents (1)-(3) in the definition of PSTP are satisfied. Let $\mathbb{K}_i$ be the ken defined by $\pi_i(\omega)$ and $\mathcal{K}_j = \{\text{ken}_j(\omega') \mid \omega' \in \pi_i(\omega)\}$. By (1) in the definition of PSTP and Claim 2, $\pi_i(\mathbb{K}_i) = \cup_{\mathbb{K}_j \in \mathcal{K}_j} \pi_j(\mathbb{K}_j)$, and thus by Corollary 1, $\mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$. Items (2) and (3) in the definition of PSTP follow immediately from (2)and (3) in the definition of IIK. Thus, we conclude, by IIK, that $d_i = d_j$.

PSTP ⇒ FTK: Suppose that PSTP holds. We need to show that if ω is in the left hand side of equation (1) then it is in the right hand side of this equation. If ω is in the left hand side, then since (a) implies (b), in Lemma 1, (1) holds. Since $\omega \in \mathrm{K_i}[\mathbf{d}_j = d]$, $\pi_i(\omega) \subseteq [\mathbf{d}_j = d]$, and (2) holds for $d = d_j$. By PSTP, $d_j = d_i = \mathbf{d}_i(\omega)$. Therefore, $\omega \in [\mathbf{d}_i = d_j] = [\mathbf{d}_i = d]$, which is the right hand side of equation (1).

This completes the proof of Theorem 1. □

Proof of Theorem 2. We recall that the *meet* of the partitions Π_1 and Π_2 is the finest partition which is coarser than Π_1 and Π_1. An event E is the union of elements of Π_1 and of elements of Π_2, if and only if it is the union of elements of the meet. The knowledge operator defined by the meet is the common knowledge operator C. Hence, $\mathrm{C}E = E$ if and only if E is the union of elements of the meet, or equivalently, if and only if it is the union of elements of Π_i for $i = 1, 2$.

IAD ⇒ EIIK: Suppose that IAD holds and also the antecedents (1)-(3) in the definition of EIIK are satisfied. By (1) and Claim 1, $\cup_{\mathbb{K}_1 \in \mathcal{K}_1} \pi_1(\mathbb{K}_1) = \cup_{\mathbb{K}_2 \in \mathcal{K}_2} \pi_2(\mathbb{K}_2)$. Denote this set by F. If $\omega \in F$, then, for $i = 1, 2$, $\omega \in \pi_i(\mathbb{K}_i)$ for some $\mathbb{K}_i \in \mathcal{K}_i$, and thus, $\pi_i(\omega) = \pi_i(\mathbb{K}_i)$. Hence, $\pi_i(\omega) \subseteq F$ and by Claim 2, $F = \cup_{\omega' \in F} \pi_i(\omega')$, for $i = 1, 2$. We conclude that F is a union of elements of the meet and therefore, $\mathrm{C}F = F$. By (2) and (3) $F \subseteq [\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]$. Since C is a knowledge operator it is monotonic with respect to inclusion, and with IAD we conclude, $F = \mathrm{C}F \subseteq \mathrm{C}([\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]) \subseteq [\mathbf{d}_1 = \mathbf{d}_2]$. Therefore, $d_1 = d_2$.

EIIK ⇒ EPSTP: Suppose that EIIK holds and the antecedents (1)-(4) in the definition of EPSTP are satisfied. By (1), (3), and Claim 2, E is a union of elements of the meet. Let $\mathcal{K}_i = \{\text{ken}_i(\omega) \mid \omega \in E\}$, for $i = 1, 2$. Then, $E = \cup_{\mathbb{K}_1 \in \mathcal{K}_1} \pi_1(\mathbb{K}_1) = \cup_{\mathbb{K}_2 \in \mathcal{K}_2} \pi_2(\mathbb{K}_2)$. By Claim 1, $\cap_{\mathbb{K}_1 \in \mathcal{K}_1} \mathbb{K}_1 = \cap_{\mathbb{K}_2 \in \mathcal{K}_2} \mathbb{K}_2$, which is (1) in the definition of EIIK. Also, (3) and (4) in the definition of EPSTP imply (2) and (3) in the definition of EIIK. Thus, by EIIK, $d_1 = d_2$.

EPSTP ⇒ IAD: Suppose that EPSTP holds. Let $E = \mathrm{C}([\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2])$. Then E is a union of elements of the meet and therefore (1) and (3) hold. Also, since C is a knowledge operator, then for each event X, $\mathrm{C}X \subseteq X$. Thus $E \subseteq [\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]$. This implies (2) and (4) and hence, by EPSTP, $d_1 = d_2$. Therefore, $\mathrm{C}([\mathbf{d}_1 = d_1] \cap [\mathbf{d}_2 = d_2]) = E \subseteq [\mathbf{d}_1 = \mathbf{d}_2]$. □

References

Aumann, R.J., 1976. Agreeing to disagree. Ann. Stat. 4 (6), 1236–1239.

Aumann, R.J., 1999. Interactive epistemology I: knowledge. Int. J. Game Theory 28, 263–300.

Bacharach, M., 1985. Some extensions of a claim of Aumann in an axiomatic model of knowledge. J. Econ. Theory 37 (1), 167–190.

Blackburn, P., de Rijke, M., Venema, Y., 2001. Modal Logic. Cambridge University Press.

Cave, J., 1983. Learning to agree. Econ. Lett. 12, 147–152.

Halpern, J.Y., 2003. Reasoning About Uncertainty. MIT Press.

Halpern, J.Y., Samet, D., Segev, E., 2009. Defining knowledge in terms of belief: the modal logic perspective. Rev. Symb. Log. 2, 469–487.

Hintikka, J., 1962. Knowledge and Belief. Cornell University Press, Ithaca, NY.

Moses, Y., Nachum, G., 1990. Agreeing to disagree after all (extended abstract). In: Theoretical Aspects of Reasoning About Knowledge. Pacific Grove, CA,. In: Morgan Kaufmann Ser. Represent. Reason. Morgan Kaufmann, San Mateo, CA, pp. 151–168.

Samet, D., 1990. Ignoring ignorance and agreeing to disagree. J. Econ. Theory 52, 190–207.

Samet, D., 2000. Quantified beliefs and believed quantities. J. Econ. Theory 95, 169–185.

Samet, D., 2008. Agreeing to disagree: the non-probabilistic case. Games Econ. Behav. 69, 169–174.

Savage, L.J., 1954. The Foundations of Statistics. John Wiley & Sons Inc., New York.

van Fraassen, C., 1984. Belief and the Will. J. Philos. 5, 190–207.