The sure-thing principle and independence of irrelevant knowledge

Dov Samet
The Faculty of Management, Tel Aviv University[*]

July 16, 2008

Abstract

Savage (1954) introduced the sure-thing principle in terms of the dependence of decisions on knowledge, but gave up on formalizing it in epistemic terms for lack of a formal definition of knowledge. Using simple models of knowledge, we examine the sure-thing principle, presenting two ways to capture it. One is in terms of the *union* of future events, for which we reserve the original name—*the sure-thing principle*; the other is in terms of the *intersection* of *kens*—bodies of agents' knowledge—which we call *independence of irrelevant knowledge*. We show that the two principles are equivalent and that the only property of knowledge required for this equivalence is the axiom of truth—the requirement that whatever is known is true. We present a symmetric version of the independence of irrelevant knowledge which is equivalent to the impossibility of agreeing to disagree on the decision made by agents, namely the impossibility of agents making different decisions being common knowledge

1 Introduction

1.1 An 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

[*] Currently on leave at ILLC, University of Amsterdam and CWI, Amsterdam, The Netherlands.

> *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)

Savage clearly considered STP as an epistemic principle describing certain relations between knowledge and decisions. He even stated the epistemic nature of this principle explicitly in the following paragraph (emphasis added):

> 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.

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. 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 two decades later in Aumann (1976).

Here we use a standard model of multi-agent knowledge for the formulation of two versions of the sure-thing principle. We show that under minimal assumptions about the nature of knowledge these two principles are equivalent.

1.2 A temporal model

The simplest way to read Savage's story concerns an agent in two periods. In the second, the agent will know whether the Democrat candidate won the election ("D") or the Republican ("R"). Since in either case he will buy the property he should buy it in the first period at a time that he does not yet know who will win. Thus, the following is a rough draft of the STP:

> **STP** (long version): If an agent knows that tomorrow she will know either E_1 or E_2 ... and in either case her decision will be the same, then this should be her decision today.[1]

[1] In Savage's story the two cases "D" and "R" appear to be exhaustive, but this assumption seems to be irrelevant. Suppose that there is a third candidate - a spoiler. As long as our businessman knows that the spoiler will not win, the situation is similar to the one where there are only two candidates.

Indeed, a more succinct but no less compelling formulation leaves the possible information implicitly assumed:

> **STP** (short version): If an agent knows her decision of tomorrow, then this should be her decision today.[2]

These versions formulate the STP in terms of the union, or disjunction, of what will be known tomorrow. But the story can also be formulated in terms of intersection. If tomorrow the Democrat wins then our agent will know "D", and a fortiori "D or R". If tomorrow the Republican wins, the agent will know "R", and obviously "D or R". However, today, she knows only the intersection of of these two bodies of knowledge, 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, neither "D" nor "R" is relevant to his decision. Indeed, the very essence of the STP demonstrated in this story is that "R" and "D" prove to be *irrelevant* to his decision to buy the property. What remains relevant is just the knowledge common to both cases, namely "D or R".

In order to suggest a draft of this formulation of the STP we use the term *ken* to describe the body of knowledge of an agent. We further assume, along with Savage's assumption, that the decision of an agent depends on her knowledge, and more precisely on her ken. We call this version of the STP the *independence of irrelevant knowledge* (IIK).

> **IIK**: If the agent's decision is the same for all of her kens tomorrow, 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]

1.3 A multi-agent model

In the temporal model the agent is split into several knowers, one for each period. Certain implicit assumptions are naturally made in this set up concerning the relation between the different knowers associated with the same agent. If we examine closely the reasoning that led to the STP we see that we assume that the agent is more knowledgeable in later periods, and, moreover, that the agent knows that.[4]

[2] One should not confuse this principle 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.

[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 a little bit sloppy for the sake of brevity. We mean that in later periods he is at least as knowledgable.

Instead of using a temporal model, we reformulate the STP and IIK in a multi-agent atemporal model. One can always think of the different agents as being the different knowers associated with a single agent at different times. The multi-agent model forces us to spell out explicitly the relations implicitly assumed between the different knowers of the same agent. The assumption made before, concerning the agent in the temporal model, is translated in the multi-agent model into the assumption that one agent is more knowledgeable than the other, and that the less knowledgeable agent knows it.

We can now state the multi-agent version of the STP where this assumption is made explicitly. One can think of Adam, in the following formulation, as being the agent today and Eve as the same agent tomorrow, or, alternatively, we can think of them as two different individuals who share the same interests.

> **IIK** (multi-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, namely, Adam's ken, and therefore this should also be Adam's decision.

The STP in the multi-agent version is:

> **STP** (multi-agent version): If Adam knows that Eve is more knowledgeable, and he also knows Eve's decision, then this must be his decision too.

Our first result states:

For partition models of knowledge, IIK *and the* STP *are equivalent.*

The STP and IIK seem to convey somewhat different stories. The STP has the air of proverbial advice: If Adam knows that Eve is wiser, that is, more knowledgeable than him, and he also knows her decision, he should follow her. IIK tells us an almost opposite story. The reason Adam's decision should be the same as Eve's is because all her extra knowledge is irrelevant to the decision. What *is* relevant is exactly what *Adam* knows. The "should" in this principle is not good advice but a logical necessity: If the decision is based on knowledge (and is independent of the knower) then the fact that Adam's decision should be the same as Eve's is because this is the right decision given *Adam's* knowledge.[5]

1.4 Symmetrization

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 symmetric cases where each of them is endowed with a family of kens.

[5] The reasoning behind IIK seems more compelling. But in richer models, for example ones that also have beliefs, it is hard to avoid the story told by the STP.

> **Symmetric IIK**: If within a family of kens of each agent the decision is the same for all kens, then what is relevant for making this decision is the intersection of kens in the family. Therefore, if 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.

While symmetric IIK bears a strong resemblance to IIK, the STP does not lend itself to straightforward symmetrization. The "right" formulation, the one that is equivalent to symmetric IIK, requires the introduction of common knowledge. The condition in this case is well known from the agreement theorem and its generalziation to the non-probabilistic case (Aumann (1976), Samet (2007)): *the impossibility to agree to disagree* (IAD).

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

We show that

> *For partition models of knowledge, symmetric* IIK *is equivalent to* IAD.

We note that despite the similarity between symmetric 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 of 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 the symmetric IIK.

1.5 Partition models and beyond

Partition models are the commonly used models of knowledge in economics and game theory. More general models of knowledge are useful to explore the implications of various properties of knowledge. Here, we study the equivalence between the STP and IIK using such general models. The versions of the STP and IIK presented so far are equivalent only in partition models. We show, however, that slightly different versions are equivalent for any knowledge model which satisfies the truth axiom, according to which whatever is known is true.

For this purpose we consider a weak version of IIK in which the family of Eve's kens is restricted to all those kens that are permitted by Adam's ken, that is, those of Eve's kens that Adam's knowledge about Eve's knowledge does not exclude. We also consider a strong version of STP in which the event that Adam knows that Eve is more knowledgeable is more loosely defined.

With this formulation we state:

> *For all models of knowledge that satisfy the truth axiom, the weak version of* IIK *is equivalent to the strong version of the* STP.

Thus, positive and negative reflection axioms are not required either for the the formulation of the STP and the IIK or for the proof of their equivalence. However, reflection across time in the single agent model, or across agents in the multi-agent formulation plays a central role in these principles.

2 Models of knowledge and decisions

2.1 The epistemic model

A *knowledge model* $(I, \Omega, (\pi_i)_{i \in I})$ consists of a set of *agents* I, a set of *states* Ω, and for each agent $i \in I$, a *possibility function* $\pi_i\colon \Omega \to 2^{\Omega}$.[6] Subsets of Ω are called *events*. The event $\pi_i(\omega)$ is the set of states that are considered possible for i at ω, while all other states are excluded by i at ω. We say that i *knows event* E at ω if $\pi_i(\omega) \subseteq E$. The event that i *knows* E, denoted $K_i(E)$ is the set of all states in which i knows E. Thus, $K_i(E) = \{\omega \mid \pi_i(\omega) \subseteq E\}$. The function $K_i : 2^{\Omega} \to 2^{\Omega}$, thus defined is called i's *knowledge operator.*

In the sequel we use common knowledge for two agents i and j. Denote for each E, $K_{ij}(E) = K_i(E) \cap K_j(E)$. The operator C_{ij} defined by $C_{ij}(E) = \cap_{m=1}^{\infty} K_{ij}^m(E)$ is the *common knowledge* (between i and j) operator.

Although we do not introduce the whole apparatus of epistemic logic here, we note that knowledge models serve for interpretation of sentences in epistemic language, where each sentence corresponds to an event, the propositional connectors correspond to set theoretic operations, and the language operator "i knows" corresponds to the operator K_i.[7] Thus, the conjunction fg corresponds to the intersection, $F \cap G$, of the events that correspond to the sentences f and g. Similarly, the disjunction $f \vee g$ corresponds to the union $F \cup G$, and the negation $\sim f$ to the complement event $\neg F$. Inspired by the logical equivalence between $\neg f \vee g$ and the implication $f \to g$, we denote the event $\neg F \cup G$ by $F \to G$. Obviously, $\omega \in F \to G$ iff the following holds: if $\omega \in F$, then $\omega \in G$.

2.2 Possibility functions

Properties of the possibility functions correspond to properties of knowledge. The claims made in this subsection are well known in the literature of modal and epistemic logic and therefore we omit the proofs (see, for example Aumann (1999), Chellas (1980), Fagin et al. (1995)).

First, by virtue of being defined in terms of a possibility function, each K_i distributes over intersection. That is, for each i, E and F, $K_i(E \cap F) =$

[6] In epistemic logic, and more generally in modal logic, models are equipped with binary relations on Ω, called the *accessibility relations*, rather than possibility functions. The two approaches are equivalent by the identification of the event assigned to a state by the possibility function with the set of all states accessible from it.

[7] Our terminology here conforms with the usual use of "model of knowledge" in the social sciences. In the terminology of modal logic, what we call here a model is called a *frame.* A model in modal logic is a frame with the interpretation of atomic sentences.

$K_i(E) \cap K_i(F)$.[8] We call this condition the *distribution axiom.* It implies that K_i is *monotonic.* That is, if $E \subseteq F$ then $K_i(E) \subseteq K_i(F)$.

If for each E the event $K_i(E) \to E$ is the whole state space, or equivalently, if $K_i(E) \subseteq E$, we say that K_i satisfies the *truth axiom.* We can rephrase it by saying that whatever i knows is true. When this holds for each agent i we say that the model satisfies the truth axiom. The operator K_i satisfies the truth axiom iff for each ω, $\omega \in \pi_i(\omega)$.[9]

If for each E, $\neg K_i(E) \to K_i(\neg K_i(E))$ is the whole state space, or equivalently, if $\neg K_i(E) \subseteq K_i(\neg K_i(E))$, we say that K_i satisfies the *negative introspection axiom.* When this holds for each i we say that the model satisfies this axiom.[10]

A model of knowledge satisfies both the truth axiom and the negative introspection axiom iff it is a *partition model*, that is, iff for each i there exists a partition Π_i of Ω, such that for each ω, $\pi_i(\omega)$ is the element of Π_i that contains ω.

The *positive introspection axiom* holds for K_i if for each E, $K_i(E) \to K_i(K_i(E))$, or equivalently, if $K_i(E) \subseteq K_i(K_i(E))$.[11] Note that in partition models the positive introspection axiom holds for each K_i.

2.3 Kens

The set of all the events E that i knows at ω, that is, the set $\{E \mid \omega \in K_i(E)\}$, is called i's *ken* at ω and is denoted by $\text{ken}_i(\omega)$. Obviously, $\text{ken}_i(\omega)$ consists of all the supersets of $\pi_i(\omega)$. We denote by Ken_i the family of all of i's kens, that is, $\text{Ken}_i = \{\text{ken}_i(\omega) \mid \omega \in \Omega\}$.

For a given ken of i, $\mathbb{K}_i$, we now describe those kens of j, $\mathbb{K}_j$, for which i's knowledge, in $\mathbb{K}_i$, concerning j's knowledge, is not contradicted by j's knowledge given by $\mathbb{K}_j$. Formally, a ken $\mathbb{K}_i \in \text{Ken}_i$ *permits* a ken $\mathbb{K}_j \in \text{Ken}_j$ if for each $E \in \mathbb{K}_j$, $\neg K_j(E) \notin \mathbb{K}_i$ and for each $E \notin \mathbb{K}_j$, $K_j(E) \notin \mathbb{K}_i$. The set of all the kens in Ken_j that are permitted by $\mathbb{K}_i$ is denoted by $\text{Permit}_j(\mathbb{K}_i)$.[12]

2.4 Decisions

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 $\mathbf{d}_i\colon \text{Ken}_i \to D$. A vector of decision functions $\mathbf{d} = (\mathbf{d}_i)_{i \in I}$ is called a *decision function profile.*

[8] For finite state spaces, the opposite also holds. That is, if an operator $K_i\colon 2^\Omega \to 2^\Omega$ satisfies this equality, then it is defined by a possibility function.

[9] This property of π_i corresponds to the reflexivity of the accessability relation associated with π_i.

[10] The negative introspection axiom holds for K_i iff for each ω and ω', if $\pi_i(\omega) \cap \pi_i(\omega') \neq \emptyset$ then $\omega \in \pi_i(\omega')$. The accessibility relation associated with such π_i is said to be Euclidian.

[11] The axiom of positive introspection holds for K_i iff for each ω and $\omega' \in \pi_i(\omega)$, $\pi_i(\omega') \subseteq \pi_i(\omega)$. This property of π_i corresponds to the transitivity of the accessability relation associated with it.

[12] Kens were introduced in Samet (1990) for more abstract models of knowledge. Using formal epistemic language, a ken can be defined as a maximal set of sentences, Φ, such that the set of sentences $\{k_i f | f \in \Phi\}$ is consistent.

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 main equivalence

The independence of irrelevant knowledge below is a formal rendering of the description given to it in subsection 1.3.

Independence of irrelevant knowledge (IIK):

The decision function profile $\mathbf{d}$ *satisfies* IIK *when, for each pair of agents i, j, a decision d,* $\mathbb{K}_i \in \mathrm{Ken}_i$*, 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* $\mathbb{K}_j \in \mathcal{K}_j$*,* $\mathbf{d}_j(\mathbb{K}_j) = d$*,*

then $\mathbf{d}_i(\mathbb{K}_i) = d$*.*

The sure-thing principle, presented next, spells out in precise terms of models of knowledge and decisions the verbal description of this principle as given in subsection 1.3. We note that $\bigcap_{E \subseteq \Omega} K_i\big(K_i(E) \rightarrow K_j(E)\big)$ is the event that i knows that j is at least as knowledgeable. More discussion of this event follows in the next section.

The sure-thing principle (STP):

The decision function profile $\mathbf{d}$ *satisfies the* STP *if for each pair of agents i, j, and decision d,*

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

Theorem 1 *For partition models of knowledge, a decision function profile satisfies independence of irrelevant knowledge if and only if it satisfies the sure-thing principle.*

4 Symmetrization

The following is an extension of IIK in which i and j play a symmetric role. Both are endowed with a family of kens: $\mathcal{K}_i \subseteq \mathrm{Ken}_i$ and $\mathcal{K}_j \subseteq \mathrm{Ken}_j$. 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 $\mathcal{K}_i$ coincides with the intersection of kens in $\mathcal{K}_j$ the same decision should be made by the two agents.

Symmetric independence of irrelevant knowledge (SIIK):

The decision function profile $\mathbf{d}$ *satisfies* SIIK *when, for each i, j, d_i, d_j,* $\mathcal{K}_i \subseteq \mathrm{Ken}_i$*, and* $\mathcal{K}_j \subseteq \mathrm{Ken}_j$*,* *if*

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

then, $d_i = d_j$.

We next formulate the symmetric "union" version that is equivalent to the symmetric "intersection" rule SIIK. For this we use common knowledge between two agents.

Impossibility of agreeing to disagree (IAD):
The decision function profile $\mathbf{d}$ *satisfies* IAD *if for each* i, j, *and* $d_i \neq d_j$,

$$C_{ij}\big([\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j]\big) = \emptyset.$$

Theorem 2 *For partition models of knowledge, a decision function profile satisfies symmetric independence of irrelevant knowledge if and only if it satisfies the impossibility of agreeing to disagree.*

5 Models that satisfy the truth axiom

In this section we show that the truth axiom is the only property of knowledge required to establish the equivalence between the "union" and the "intersection" aspects of the sure-thing principle. The two equivalence theorems of the previous sections indeed hold only for partition models. However, we now formulate a weak version of IIK and a strong version of the STP, which are equivalent for any knowledge model that just satisfies the truth axiom. For partition models the difference between the previous versions and the new ones is cosmetic. For such models, weak IIK is equivalent to IIK and strong STP is equivalent to STP.

In the weak version of IIK the set of kens $\mathcal{K}_j$ is restricted to be the set of j's kens that are permitted by $\mathbb{K}_i$.

Weak independence of irrelevant knowledge:
The decision functions profile $\mathbf{d}$ *satisfies weak* IIK *when, for each pair of agents* i, j, *a decision* d, $\mathbb{K}_i \in \mathrm{Ken}_i$, *and* $\mathcal{K}_j = \mathrm{Permit}_j(\mathbb{K}_i)$,

if

1. $\mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$,
2. *for each* $\mathbb{K}_j \in \mathcal{K}_j$, $\mathbf{d}_j(\mathbb{K}_j) = d$,

then $\mathbf{d}_i(\mathbb{K}_i) = d$.

The strong version of the STP is obtained by replacing the event

$$\bigcap_{E\subseteq\Omega} K_i\big(K_i(E) \to K_j(E)\big) \tag{1}$$

by the event

$$\bigcap_{E\subseteq\Omega} K_i(E) \to K_i\big(K_j(E)\big) \tag{2}$$

Although (1) and (2) seem to say the same thing, namely that i knows that j is at least as knowledgeable, there is a slight difference between them. In (1) it is said explicitly. The event $\cap_{E\subseteq\Omega} K_i(E) \to K_j(E)$ is the event that whatever i knows j knows, that is, the event that j is at least as knowledgeable as i. Thus, the event $K_i\big(\cap_{E\subseteq\Omega} K_i(E) \to K_j(E)\big)$ is the event that i knows that j is at least as knowledgeable. But this last event is the event in (1).

In (2) the event that j is at least as knowledgeable as i, and i's knowledge of this are interwind. Indeed, for models that satisfy the truth axiom, (2) implies that j is at least as knowledgeable as i, since $K_i\big(K_j(E)\big) \subseteq K_j(E)$, and therefore, $K_i(E) \to K_i\big(K_j(E)\big) \subseteq K_i(E) \to K_j(E)$. But (2) does not imply that i knows this event unless negative introspection is assumed.

We consider the next version of STP to be a strengthening of the STP since, as we show in the proof of Proposition 2, for models that satisfy positive introspection the event in (1) is a subset of the event in (2).

The strong sure-thing principle:
The decision function profile **d** *satisfies the strong* STP *when, for each pair of agents i, j, and decision d,*

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

The following two propositions examine the changes introduced in this section to IIK and the STP from the perspective of partition models.

Proposition 1 *In partition models of knowledge, if $\mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$, for $\mathbb{K}_i \in \mathrm{Ken}_i$ and $\mathcal{K}_j \subseteq \mathrm{Ken}_j$, then $\mathcal{K}_j = \mathrm{Permit}_j(\mathbb{K}_i)$.*

Proposition 2 *In partition models of knowledge,*

$$K_i\big(K_i(E) \to K_j(E)\big) = K_i(E) \to K_i\big(K_j(E)\big).$$

In light of these two propositions the difference between the two versions of IIK and the STP are insignificant as far as partition models are concerned.

Corollary 1 *For partition models of knowledge, a decision function profile satisfies weak* IIK *if and only if it satisfies* IIK*, and it satisfies the strong* STP *if and only if it satisfies* STP*.*

The reason for the formulation of weak IIK and the strong STP is the following theorem.

Theorem 3 *For models of knowledge that satisfy the axiom of truth, a decision function profile satisfies weak independence of irrelevant knowledge if and only if it satisfies the strong sure-thing principle.*

The symmetric IIK is also too strong for Theorem 2 to hold for non-partition models. We weaken it by requiring that all the permitted kens of an agent's ken are found in the family of the other agent's kens.

Weak symmetric independence of irrelevant knowledge:
The decision function profile $\mathbf{d}$ *satisfies weak* SIIK *when, for all* i, j, d_i, d_j, $\mathcal{K}_i \subseteq \mathrm{Ken}_i$, *and* $\mathcal{K}_j \subseteq \mathrm{Ken}_j$,

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

Note, that the requirement that the two intersections of the players' kens coincide, which is part of the definition of IIK, is not found here. The explanation is in the next proposition.

Proposition 3 *Let* $\mathcal{K}_i \subseteq \mathrm{Ken}_i$ *and* $\mathcal{K}_j \subseteq \mathrm{Ken}_j$, *and consider the following two properties:*

1. *for each* $\mathbb{K}_i \in \mathcal{K}_i$, $\mathrm{Permit}_j(\mathbb{K}_i) \subseteq \mathcal{K}_j$ *and for each* $\mathbb{K}_j \in \mathcal{K}_j$, $\mathrm{Permit}_i(\mathbb{K}_j) \subseteq \mathcal{K}_i$,
2. $\cap_{\mathbb{K}_i \in \mathcal{K}_i} \mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$.

Then for models that satisfy the truth axiom, (1) *implies* (2). *For partition models,* (1) *and* (2) *are equivalent.*

Corollary 2 *For partition knowledge models, weak* SIIK *is equivalent to* SIIK.

Theorem 4 *For models that satisfy the truth axiom, a decision function profile satisfies weak symmetric independence of irrelevant knowledge if and only if it satisfies the impossibility of agreeing to disagree.*

6 Discussion

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 generalization of 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 from which individual decisions are derived.[13] 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.

The STP as presented here was suggested in Samet (2007) under the name *interpersonal* sure-thing principle. Here, we retract the attribute interpersonal, because there is simply no STP for one knower only. The STP is used in Samet (2007) as a condition for an agreement theorem. We see here that the impossibility of agreeing to disagree on decisions is *equivalent* to symmetric IIK. The asymmetric version of IIK does not imply the IDA. This is why in Samet (2007) it is required that the model satisfies not only STP (which is equivalent to asymmetric IIK) but also STP-expandability. That is, we require that STP also holds when another agent is added to the model. In view of our result, adding an agent enables us to convert a symmetric situation as described in the definition of SIIK, into an asymmetric one to which the STP can be applied.

7 Proofs

Lemma 1 *In any model of knowledge, for each ω, $\omega' \in \pi_i(\omega)$ and j,* $\mathrm{ken}_j(\omega') \in \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$.

[13] The term virtual decision function was suggested in Samet (2007).

Proof: Suppose $\omega' \in \pi_i(\omega)$ and let $E \in \mathrm{ken}_j(\omega')$. Then $\omega' \in K_j(E)$. If $\neg K_j(E) \in \mathrm{ken}_i(\omega)$, then $\omega \in K_i(\neg K_j(E))$ and thus, $\omega' \in \pi_i(\omega) \subseteq \neg K_j(E)$, which is a contradiction. If $E \notin \mathrm{ken}_j(\omega')$ then $\omega' \notin K_j(E)$. If $K_j(E) \in \mathrm{ken}_i(\omega)$, then $\omega \in K_i(K_j(E))$ and thus, $\omega' \in \pi_i(\omega) \subseteq K_j(E)$, which is a contradiction. ∎

Proof of Proposition 1: Suppose that $\mathcal{K}_j \subseteq \mathrm{Ken}_j$, and $\mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$, where $\mathbb{K}_i = \mathrm{ken}_i(\omega)$. For each $\mathrm{ken}_j(\omega') \in \mathcal{K}_j$, $\mathrm{ken}_i(\omega) \subseteq \mathrm{ken}_j(\omega')$, and hence, $\pi_j(\omega') \subseteq \pi_i(\omega)$. Thus, by the truth axiom, $\omega' \in \pi_i(\omega)$ and by Lemma 1, $\mathrm{ken}_j(\omega') \in \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$.

Conversely, let $\mathrm{ken}_j(\omega') \in \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$, and suppose that for all $\mathrm{ken}_j(\omega'') \in \mathcal{K}_j$, $\mathrm{ken}_j(\omega') \neq \mathrm{ken}_j(\omega'')$. Then for each such ω'', $\pi_j(\omega'') \neq \pi_j(\omega')$, and as the π_j is derived from a partition, $\pi_j(\omega'') \cap \pi_j(\omega') = \emptyset$. Hence, $\neg\pi_j(\omega') \in \mathrm{ken}_j(\omega'')$. This implies that $\neg\pi_j(\omega') \in \mathrm{ken}_i(\omega)$. By the definition of knowledge and the truth axiom, $\pi_j(\omega') = K_j(\pi_j(\omega'))$. Substituting in the previous inclusion we conclude that $\neg K_j(\pi_j(\omega')) \in \mathrm{ken}_i(\omega)$. But obviously, $\pi_j(\omega') \in \mathrm{ken}_j(\omega')$ which contradicts our assumption that $\mathrm{ken}_j(\omega')$ is permitted by $\mathrm{ken}_i(\omega)$. ∎

Proof of Proposition 2: By the distribution axiom and monotonicity, for any E and F, $K_i(E \to F) \cap K_i(E) = K_i((\neg F \cup E) \cap E) = K_i(E \cap F) \subseteq K_i(F)$. Thus $K_i(E \to F) \subseteq \neg K_i(E) \cup K_i(F) = K_i(E) \to K_i(F)$. Using this inclusion and the fact that positive introspection and the truth axiom imply $K_i(K_i(E)) = K_i(E)$ we conclude that $K_i\big(K_i(E) \to K_j(E)\big) \subseteq K_i(E) \to K_i\big(K_j(E)\big)$.

For the opposite inclusion, it is enough to show that $\big(K_i(E) \to K_i\big(K_j(E)\big)\big) \cap K_i\big(K_i(E) \to K_j(E)\big) = K_i(E) \to K_i\big(K_j(E)\big)$. By the truth axiom and negative introspection, $K_i(\neg K_i(E)) = \neg K_i(E)$ and thus the lefthand side of this equation is $\big(K_i(\neg K_i(E)) \cup K_i(K_j(E))\big) \cap \big(K_i(\neg K_i(E) \cup K_j(E))\big)$. By distribution this set is $K_i(\neg K_i(E)) \cup K_i(K_j(E))$ which is the righthand side of the equation. ∎

In view of Propositions 1 and 2, Theorem 1 follows from Theorem 3, which we prove next.

Proof of Theorem 3: Suppose that the strong STP holds. Assume that $\mathrm{ken}_i(\omega)$ is the intersection of the kens in $\mathrm{Permit}_j(\mathrm{ken}_i(\omega))$ and for each $\mathbb{K}_j$ in this set, $\mathbf{d}_j(\mathbb{K}_j) = d$.

As $\mathrm{ken}_i(\omega)$ is the intersection of all kens in $\mathrm{Permit}_j(\mathrm{ken}_i(\omega))$, it follows by Lemma 1 that for each $\omega' \in \pi_i(\omega)$, $\mathrm{ken}_i(\omega) \subseteq \mathrm{ken}_j(\omega')$ and therefore $\pi_j(\omega') \subseteq \pi_i(\omega)$. We show that for each E, $\omega \in K_i(E) \to K_i(K_j(E))$. Indeed, if $\omega \in K_i(E)$, then $\pi_i(\omega) \subseteq E$. Thus, for each $\omega' \in \pi_i(\omega)$, $\pi_j(\omega') \subseteq E$, i.e., $\omega' \in K_j(E)$. Hence, $\pi_i(\omega) \subseteq K_j(E)$, which means that $\omega \in K_i(K_j(E)$.

As $\mathbf{d}_j(\omega') = \mathbf{d}_j(\mathrm{ken}_j(\omega')) = d$ for each $\omega' \in \pi_i(\omega)$, it follows that $\pi_i(\omega) \subseteq [\mathbf{d}_j = d]$. Thus, $\omega \in K_i([\mathbf{d}_j = d])$ and since we assumed that the STP holds, $\mathbf{d}_i(\omega) = d$, i.e., $\mathbf{d}_i(\mathrm{ken}_i(\omega)) = d$ as required.

Suppose now that weak IIK holds and assume that for each E, $\omega \in (K_i(E) \to K_i(K_j(E)))$, and $\omega \in K_i([\mathbf{d}_j = d])$. We claim that $\mathrm{ken}_i(\omega)$ is the intersection of the kens in $\mathrm{Permit}_j(\mathrm{ken}_i(\omega))$.

First we show that $\mathrm{ken}_i(\omega) = \cap_{\omega' \in \mathrm{ken}_i(\omega)} \mathrm{ken}_j(\omega')$. Indeed, if $E \in \mathrm{ken}_i(\omega)$ then $\omega \in K_i(E)$. Therefore $\omega \in K_i(K_j(E))$ and thus $\pi_i(\omega) \subseteq K_j(E)$. Hence for each $\omega' \in \pi_i(\omega)$, $E \in \mathrm{ken}_j(\omega')$. Conversely, if $E \in \mathrm{ken}_j(\omega')$ for each $\omega' \in \pi_i(\omega)$, then for each such ω', $\omega' \in K_j(E)$. Hence, $\pi_j(\omega') \subseteq E$, and since by the truth axiom $\omega' \in \pi_j(\omega')$, it follows that $\omega' \in E$. Therefore $\pi_i(\omega) \subseteq E$, and hence $E \in \mathrm{ken}_i(\omega)$.

To prove the claim it is enough to show that $\mathrm{ken}_j(\omega') \in \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$ if and only if $\omega' \in \pi_i(\omega)$. By Lemma 1 it suffices to show that if $\omega' \notin \pi_i(\omega)$ then $\mathrm{ken}_j(\omega') \notin \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$. Let $E = \pi_i(\omega)$. By the truth axiom, $K_j(E) \subseteq E$, and as $\omega' \notin E$ it follows that $\omega' \notin K_j(E)$, i.e., $E \notin \mathrm{ken}_j(\omega')$. But, obviously, $\omega \in K_i(E)$ and therefore $\omega \in K_i(K_j(E))$, per our assumption, which means that $K_j(E) \in \mathrm{ken}_i(\omega)$.

To complete the proof, we note that since $\omega \in K_i([\mathbf{d}_j = d])$, $\pi_i(\omega) \subseteq [\mathbf{d}_j = d]$. Thus, for each $\omega' \in \pi_i(\omega)$, $\mathbf{d}_j(\omega') = d$, i.e., $\mathbf{d}_j(\mathrm{ken}_j(\omega')) = d$. By what we have shown, this means that $\mathbf{d}_j(\mathbb{K}) = d$ for all kens in $\mathrm{Permit}_j(\mathrm{ken}_i(\omega))$. By IIK this says that $\mathbf{d}_i(\mathrm{ken}_i(\omega)) = d$. ∎

Proof of Proposition 3: To prove the first part, assume that for each $\mathbb{K}_i \in \mathcal{K}_i$, $\mathrm{Permit}_j(\mathbb{K}_i) \subseteq \mathcal{K}_j$ and for each $\mathbb{K}_j \in \mathcal{K}_j$, $\mathrm{Permit}_i(\mathbb{K}_j) \subseteq \mathcal{K}_i$. We show that

$$\{\omega \mid \mathrm{ken}_i(\omega) \in \mathcal{K}_i\} = \{\omega \mid \mathrm{ken}_j(\omega) \in \mathcal{K}_j\} \tag{3}$$

Indeed, suppose that $\mathrm{ken}_i(\omega) \in \mathcal{K}_i$. As $\omega \in \pi_i(\omega)$, it follows from Lemma 1 and our assumption that $\mathrm{ken}_j(\omega) \in \mathrm{Permit}_j(\mathrm{ken}_i(\omega))$ and hence that $\mathrm{ken}_j(\omega) \in \mathcal{K}_j$. The inverse inclusion is similarly proved.

Denote the set in (3) by F. Again, by Lemma 1 and our assumption it follows that for each $\omega \in F$, $\pi_i(\omega) \subseteq F$ and $\pi_j(\omega) \subseteq F$.

We now show that that $E \in \cap_{\mathbb{K}_i \in \mathcal{K}_i} \mathbb{K}_i$ iff $F \subseteq E$. Since the same equivalence holds for $\cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_i$ this proves the equality of these two intersections. To prove this equivalence note that $E \in \cap_{\mathbb{K}_i \in \mathcal{K}_i} \mathbb{K}_i$ iff for each $\omega \in F$, $E \in \mathrm{ken}_i(\omega)$ which holds iff $\pi_i(\omega) \subseteq E$. Since we have shown that for each such ω, $\pi_i(\omega) \subseteq F$, and $\omega \in \pi_i(\omega)$, the required equivalence follows.

To prove the second part of the propostion, suppose that $\cap_{\mathbb{K}_i \in \mathcal{K}_i} \mathbb{K}_i = \cap_{\mathbb{K}_j \in \mathcal{K}_j} \mathbb{K}_j$ and let $\mathrm{ken}_i(\omega) \in \mathcal{K}_i$. We first note that if $\pi_j(\omega') \cap \pi_i(\omega) = \emptyset$, then $\pi_j(\omega') \notin \mathrm{Permit}_j(\mathbb{K}_i)$. Indeed, by the truth axiom, for each $\omega'' \in \pi_i(\omega)$, $\omega'' \in \pi_j(\omega'') \cap \pi_i(\omega)$ and thus, $\pi_j(\omega'') \neq \pi_j(\omega')$. As π_j is derived from a partition, $\pi_j(\omega'') \cap \pi_j(\omega') = \emptyset$. Therefore, for $E = \neg\pi_j(\omega')$, $\omega'' \in K_j(E)$. Hence, $\pi_i(\omega) \subseteq K_j(E)$, which means that $K_j(E) \in \mathrm{ken}_i(\omega)$. But, obviously, $E \notin \mathrm{ken}_j(\omega')$ which shows that $\mathrm{ken}_j(\omega') \notin \mathrm{Permit}_j(\mathbb{K}_i)$.

We have shown now that all j's kens that are permitted by $\mathrm{ken}_i(\omega)$ must intersect it. Thus, it is now enough to show that if $\pi_j(\omega') \cap \pi_i(\omega) \neq \emptyset$, $\mathrm{ken}_j(\omega') \in \mathcal{K}_j$. Suppose $\mathrm{ken}_j(\omega') \notin \mathcal{K}_j$. Then, as we argued before, for $E = \neg\pi_j(\omega')$ and for each $\mathrm{ken}_j(\omega'') \in \mathcal{K}_j$, $E \in \mathrm{ken}_j(\omega'')$. Thus, $E \in \cap_{\mathbb{K}_i \in \mathcal{K}_i} \mathbb{K}_i$, and in particular, $E \in \mathrm{ken}_i(\omega)$. But this is impossible, since $\pi_i(\omega) \not\subseteq E$. ∎

In view of Propositions 3, Theorem 2 follows from Theorem 4, which we prove next. To do so we express the common knowledge operator C_{ij} in terms of the possibility functions π_i and π_j. Omitting henceforth the subscripts ij, we define a function $\pi\colon \Omega \to 2^{\Omega}$ by $\pi(\omega) = \pi_i(\omega) \cup \pi_j(\omega)$. For an event F, let $\pi(F) = \cup_{\omega \in F} \pi(\omega)$. Note, that $F \subseteq K(E)$ iff $\pi(F) \subseteq E$. Finally, define $\pi_c \colon \Omega \to 2^\Omega$ by $\pi_c(\omega) = \cup_{m \geq 1} \pi^m(\omega)$, where π^m are powers of π.

Proposition 4 *The common knowledge operator C is derived from the possibility function π_c. That is, for each E, $C(E) = \{\omega \mid \pi_c(\omega) \subseteq E\}$.*[14]

Proof: It is enough to show that for each $m \geq 1$, $\omega \in K^m(E)$ iff $\pi_c^m(\omega) \subseteq E$, which we prove by induction on m. For $m = 1$ this holds by the definition of the knowledge operators K_i. Suppose we prove for m and assume that $\omega \in K^{m+1}(E)$, i.e., $\omega \in K^m(K(E))$. By the induction hypothesis, this is equivalent to $\pi_c^m(\omega) \subseteq K(E)$, which in turn is equivalent to $\pi_c(\pi_c^m(\omega)) \subseteq E$. ∎

Proof of Theorem 4: Suppose that weak SIIK holds and assume that $\omega \in C_{ij}([\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j])$. For $m \in \{i, j\}$, define $\mathcal{K}_m = \{\text{ken}_m(\omega') \mid \omega' \in \pi_c(\omega)\}$. We show that (1) and (2) in the definition of weak SIIK hold for $\mathcal{K}_i$ and $\mathcal{K}_j$.

To prove (1) it is enough to show that for any $\omega' \in \pi_c(\omega)$ and $\omega'' \notin \pi_c(\omega)$, $\text{ken}_j(\omega'') \notin \text{Permit}_j(\text{ken}_i(\omega'))$. Indeed, let $E = \pi_c(\omega)$. Then, since for each $\bar{\omega} \in E$, $\pi_j(\bar{\omega}) \subseteq E$ it follows that $E \subseteq K_j(E)$. Since $\pi_i(\omega') \subseteq E$ it follows that $\omega' \in K_i(K_j(E))$, i.e., $K_j(E) \in \text{ken}_i(\omega')$. But obviously, $E \notin \text{ken}_j(\omega'')$ because $\pi_j(\omega'') \not\subseteq F$ as ω'' in $\pi_j(\omega'')$ but not in E.

To show (2), we note that by Proposition 4, $\pi_c(\omega) \subseteq C_{ij}([\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j])$, and by the truth axiom the latter event is a subset of $[\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j]$. Thus for every $\omega' \in \pi_c(\omega)$, $\mathbf{d}_i(\omega') = d_i$ and $\mathbf{d}_j(\omega') = d_j$. Thus, for each $\mathbb{K}_i \in \mathcal{K}_i$, $\mathbf{d}_i(\mathcal{K}_i) = d_i$ and a similar equality holds for j. By weak SIIK it follow that $d_i = d_j$, as required.

Suppose that IAD holds and assume that (1), (2) in the definition of weak SIIK hold. For $\text{ken}_i(\omega) \in \mathcal{K}_i$, we show that $\omega \in C_{ij}([\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j])$. Consider the set F constructed in the proof of Proposition 3. By (2) it follows that $F \subseteq [\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j]$. We have shown that for each $\omega' \in F$, $\pi(\omega') \subseteq F$. By the definition of F, $\omega \in F$. It follows by induction that $\pi_c(\omega) \subseteq F$. By Proposition 4 this shows that $\omega \in C_{ij}([\mathbf{d}_i = d_i] \cap [\mathbf{d}_j = d_j])$. By IAD $d_i = d_j$. ∎

References

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

Aumann, R. J. (1999), Interactive epistemology I: Knowledge, *Internat. J. Game Theory*, 28(3), 263–300.

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

Cave, J. (1983), Learning to agree, *Economics Letters*, 12, 147–152.

Chellas, B. F (1980) *Modal logic, and introduction.* Cambridge University Press.

[14] The reasoning behind this proposition is the motivation for the definition of common knowledge for partition models in Aumann (1976). The observation that the proposition holds for any model of knowledge based on possibility functions is due to Fagin et al. (1995).

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995), *Reasoning about knowledge.* MIT Press, Cambridge, MA.

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

Moses, Y. and G. Nachum (1990), Agreeing to disagree after all (extended abstract), In *Theoretical aspects of reasoning about knowledge (Pacific Grove, CA,)*, Morgan Kaufmann Ser. Represent. Reason., 151–168. Morgan Kaufmann, San Mateo, CA.

Samet, D. (1990), Ignoring ignorance and agreeing to disagree, *J. of Economic Theory*, 52, 190–207.

Samet, D. (2007), Agreeing to disagree: The non-probabilistic case, forthcoming *GEB.*

Savage, L. J. (1954), *The foundations of statistics.* John Wiley & Sons Inc., New York.