Agreeing to Disagree After All
(Extended Abstract)

Yoram Moses

Department of Applied Math and CS
The Weizmann Institute of Science

Gal Nachum

Department of Computer Science
Tel Aviv University

Abstract

Bacharach and Cave independently generalized Aumann's celebrated agreement theorem to the case of *decision functions.* Roughly speaking, they showed that once two like-minded agents reach common knowledge of the actions each of them intends to perform, they will perform identical actions. This theorem is proved for decision functions that satisfy a condition that Bacharach calls the *sure thing condition*, which is closely related to Savage's *sure thing principle.* The assumption that any reasonable decision function should satisfy the sure thing condition seems to have been widely accepted as being natural and intuitive.

By taking a closer look at the meaning of the sure thing condition in this context, we argue that the technical definition of the sure thing condition does not capture the intuition behind Savage's sure thing principle very well. It seems to involve nontrivial hidden assumptions, whose appropriateness in the case of non-probabilistic decision functions is questionable. Similar trouble is found with the technical definition of the like-mindedness of two agents. Alternative definitions of the sure thing principle and like-mindedness are suggested, and it is shown that the agreement theorem does not hold with respect to these definitions. In particular, it is shown that the agreement theorem does not apply to a particularly appealing example attributed to Bacharach. Conditions that do guarantee the agreement theorem for decision functions are presented. Finally, we consider similar issues that arise in the case of communication among more than two agents, as studied by Parikh and Krasucki.

1 Introduction

In his seminal paper [Aum76], Aumann proved that agents that have the same prior probability distribution over the states of the world cannot agree to disagree. More precisely, once their posteriors for a certain event are common knowledge, these posteriors must coincide, despite the fact that they may be based on different information. In [Bac85,Cav83] Bacharach and Cave independently generalized Aumann's result from posterior probabilities to decision functions: Roughly speaking, they showed that once two like-minded agents reach common knowledge of the actions each of them intends to perform, they will perform identical actions. This theorem is proved for decision functions that satisfy a condition Bacharach calls the *sure thing condition.* The following story, due to Bacharach, is intended to vividly capture the essence of the generalized agreement theorem. We present here the version found in [Aum89].

> *A murder has been committed. To increase the chances of conviction, the chief of police puts two detectives on the case, with strict instructions to work independently, to exchange no information. The two, Alice and Bob, went to the same police school; so given the same clues, they would reach the same conclusions. But as they will work independently, they will, presumably, not get the same clues.*
>
> *At the end of thirty days, each is to decide whom to arrest (possibly nobody). On the night before the thirtieth day, they happen to meet in the locker room at headquarters, and get to talking about the case. True to their instructions, they exchange no substantive information, no clues; but both are self-confident individuals, and feel that there is no harm in telling each other whom they plan to arrest. Thus, when they leave the locker room, it is common knowledge between them whom Alice will arrest, and it is common knowledge between them whom Bob will arrest.*
>
> *Conclusion: They arrest the same people; and this, in spite of knowing nothing about each other's clues.*

In the case of this example, the theorem assumes that the rules by which each detective decides whom to arrest satisfy the sure thing condition. While the precise definition of this condition will be given in a later section, it is intuitively explained in [Aum89] as follows: "*Suppose that if you knew which of the mutually exclusive events J happened, you would do b (which is the same for all J). Then you will take the same action b if you only know that some J happened, without knowing which one. Thus, if Alice would arrest the butler if a certain blood stain is typed A, B, AB, or O, (perhaps for different reasons in each case), then she can arrest the butler without bothering to send the stain to the police laboratory.*" Throughout this paper, we will think of this explanation as corresponding to Savage's *sure thing principle* [Sav54].

By taking a close look at the proof of the agreement theorem, we will see that Bacharach's technical definition of the sure thing principle is considerably stronger than is implied by the intuitive explanation given above. Roughly speaking, it requires the decision function to be defined in a manner satisfying certain consistency properties at what amount to impossible situations. This is analogous to requiring a chess player's strategy to be defined, say, for a board with three kings. In addition, it turns out that the notion of like-mindedness implicitly used in [Bac85,Cav83,Aum89] doesn't quite correspond to the idea that given the same information the agents will reach the same decisions. The agreement theorem for decision functions is thus less applicable than initially thought. In particular, it does not directly capture the above murder story example.

It is, nevertheless, conceivable that once we make the appropriate definitions capturing the sure thing principle and like-mindedness, the agreement theorem will still hold. We show that this is not the case. In fact, we present a scenario consistent with the above murder story, in which the detectives are like-minded and satisfy the sure thing principle as described above, and yet Alice and Bob *agree to disagree after all.* Given these negative results, we turn to study whether there are other natural conditions one could make that do ensure the agreement theorem. It is shown that a strengthening of the sure thing principle, which we call the *inclusive sure thing principle*, does work.

Finally, we consider Parikh and Krasucki's extensions of Bacharach and Cave's agreement theorem to the case of communication among more than two agents [PK87]. In extending the previous work, Parikh and Krasucki's analysis suffers from similar shortcomings in terms of its applicability to deterministic decision functions. They present conditions that guarantee the agreement theorem for $n \geq 3$ agents. We show that their condition for $n = 3$ can be extended to an appropriate inclusive condition as above, while their condition for $n \geq 4$ cannot be extended in a nontrivial way.

This paper is organized as follows. In the next section we present the model of knowledge along the lines of [Aum89]. The agreement theorem is stated, proven, and discussed in Section 3. In Section 4 we reformulate the formal definitions of like-mindedness and the sure thing principle. Section 5 shows that the agreement theorem fails for the reformulated definitions, by presenting a "counterexample" scenario that is an instance of the murder story example. The agreement theorem is shown to hold with respect to an inclusive variant of the sure thing principle. Finally, the case of communication between more than two agents is discussed. Section 6 concludes with a discussion of the meaning of all this.

2 Knowledge and Common Knowledge

We begin by reviewing the model of knowledge used for the standard proof of the (generalized) agreement theorem. Our exposition will closely follow that of [Aum89]. We start with a set Ω whose members are called *states of the world.* An *event* e is defined as a subset of Ω. The family of all events is denoted by $\mathcal{E}$. An event in the ordinary sense of the word, such as "a crime has been committed", will be identified in the formalism with the set of all states of the world at which a crime has been committed. Inclusion of events corresponds to implication, union to disjunction, intersection to conjunction and complementation to negation. In addition, we have a set $\Pi = \{1 \ldots n\}$ of *agents*. For each agent $j \in \Pi$ we will assume that we are given a function $I_j : \Omega \to 2^{\Omega}$ satisfying:

1. $\omega \in I_j(\omega)$
2. $I_j(\omega)$ and $I_j(\omega')$ are either identical or disjoint.

I_j thus induces a partition of the states of Ω. I_j is called the *information function* of agent j. Intuitively $I_j(\omega)$ is the set of all those states of the world that are for agent j *indistinguishable* from ω. Thus, if ω is the true state of the world, j will usually not know this; he will know only that the true state is a member of $I_j(\omega)$. Put another way, $I_j(\omega)$ consists of all the states ω' that are considered by j to be possible when ω is the true state of the world. An event e is called a *possible state of knowledge* for agent j if $e = I_j(\omega)$ for some $\omega \in \Omega$.

A *knowledge operator* $K_j : \mathcal{E} \to \mathcal{E}$ is now defined as follows: for every event e,

$$K_j(e) = \{\omega \in \Omega \;:\; I_j(\omega) \subseteq e\}$$

In words, $K_j(e)$ is the event that j knows that e obtains. (Notice that when we interpret events as formulas, such a definition of knowledge has the properties of the modal system S5 [HM85].) The event that *everyone* in a set $N \subseteq \Pi$ of agents *knows* that e obtains is captured by an operator $E_N : \mathcal{E} \to \mathcal{E}$ defined as follows:

$$E_N(e) = \cap_{j \in N} K_j(e).$$

It is easy to verify that all members of N know e iff $E_N(e)$ holds. We can iterate the E_N operator, and define $E_N^1(e) = E_N(e)$, and $E_N^{m+1}(e) = E_N(E_N^m(e))$ for $m \geq 1$.

An event e is called *common knowledge* among a group N of agents, denoted $C_N(e)$, if they all know e, all know that all know it, all know that all know that all know it, and so on ad infinitum. Formally, we define

$$C_N(e) = \cap_{m=1}^{\infty} E_N^m(e).$$

We will generally omit the subscript N when either the set in question is the set Π of all agents or when its identity is clear from context.

An equivalent definition of common knowledge uses the notion of a greatest fixed point. An event e is a fixed point of a function $f : \mathcal{E} \to \mathcal{E}$ if $f(e) = e$; that is, if f maps the event e to itself. An event e is the *greatest fixed point* of f, if it is a fixed point of f and $e' \subseteq e$ for any other fixed point e' of f. An equivalent definition of $C_N(e)$ is as the greatest fixed point of the function

$$f(X) = E_N(e \cap X).$$

This definition seems to better reflect the way common knowledge actually arises. Common knowledge does not result from an infinite iterative process in which higher and higher levels of $E_N^m(e)$ are attained, but rather from a situation that is a fixed point of the E_N operator [HM89]. For other frequently used definitions of common knowledge, see [Aum76,HM89].

3 The Agreement Theorem

In order to prove the agreement theorem we need a few definitions. An *action function* d_j specifies agent j's actions at a given state, as a function of the agent's state of knowledge (the set of states of the world she considers possible). It follows that the value of $d_j(\omega)$ is completely determined by the set $I_j(\omega)$. We are therefore led to think of such functions as functions from nonempty subsets of Ω. Define a *decision function* to be a function D from nonempty subsets of Ω to a set $\mathcal{A}$ of actions. We say that the agent j using the action function d_j *follows the decision function* D if for all states ω it is the case that $d_j(\omega) = D(I_j(\omega))$. According to Bacharach, a decision function D is said to satisfy the *sure thing condition* (STC),[1] if whenever a nonempty event e is a disjoint union of a family of events $\{J\}$, on each of which $D(J) = b$, then also $D(e) = b$. (We will refer to this technical definition as the sure thing *condition*, as opposed to the sure thing *principle*, by which we refer to Savage's intuitive notion.) Finally, Bacharach defines two agents to be *like-minded* if they both follow the same decision function.

We are now in a position to state and prove the agreement theorem in the form found in [Bac85,Cav83,Aum89]. We will denote by "$d_1 = a$" the event consisting of all worlds $\omega \in \Omega$ for which $d_1(\omega) = a$. In words, the agreement theorem says that if two like-minded agents following a decision function that satisfies STC agree (attain common knowledge of the fact) that one of them intends to perform action a and the other intends to perform b, then $a = b$. This captures the fact that the agents cannot agree to disagree on what the appropriate course of action is.

Theorem 3.1: *[Agreement Theorem] Let* $\Pi = \{1,2\}$. *If the agents are like-minded, and follow a decision function D satisfying the sure thing condition, then*

$$[C(d_1 = a) \cap C(d_2 = b) \neq \emptyset] \quad \Rightarrow \quad a = b.$$

Proof: Set $e = C(d_1 = a) \cap C(d_2 = b)$. Because e is common knowledge, $K_1(e) = e$ and therefore e is a disjoint union of elements of the form $I = I_1(\omega)$. By definition of e, at every $\omega \in e$ we have $d_1(\omega) = a$, and hence $D(I_1(\omega)) = a$. Now, since e is a (nonempty) disjoint union of such sets I, the fact that D satisfies the sure thing condition implies that $D(e) = a$. Similar reasoning with respect to agent 2 shows that $D(e) = b$, and as a result we obtain $a = b$. ∎

3.1 Discussion

The agreement theorem is based on two assumptions: The like-mindedness of the agents, and the fact that their decision function satisfies the sure thing condition. The theorem immediately applies to the case in which the decision function yields the agent's posterior (conditional) probability of a particular event, based on a common prior probability distribution over the

[1] Cave uses the term *union consistent* to denote essentially the same property.

states of Ω, which is the case in Aumann's original agreement theorem [Aum76]. It has also been shown to apply in other situations originating from a common prior distribution, such as expectations or functions that maximize conditional expectations [Bac85,Cav83]. However, as exemplified in the murder story example, Bacharach intuitively argues that these assumptions apply to a very wide range of situations. He claims that the sure thing condition applies to "*just about any plausible theory of rational decision*", and interprets the agreement theorem as showing that, in general, "*differences in information alone cannot account for differences in behavior of rational persons ... any more than they can account for differences of opinion*" [Bac85]. We now attempt a critical assessment of this point of view.

3.1.1 Decision functions and STC

The proof of agreement theorem makes strong use of the fact that the decision function D is defined, and in a way that satisfies the sure thing condition, on the event e, which is a union of states of knowledge. But notice that an agent j's actions are completely determined once D is defined for all the events in $\mathcal{I} = \{I_j(\omega) : \omega \in \Omega\}$, which are all the possible states of knowledge of j. Extending D beyond $\mathcal{I}$ is intuitively analogous to extending a strategy for white in the game of chess to impossible board positions. For cardinality reasons it is obvious that $\mathcal{I}$ cannot include all the nonempty subsets of Ω. The following lemma characterizes a particularly interesting class of events that are definitely not possible states of knowledge for j:

Lemma 3.2: *If e' is a union of two or more different states of knowledge for j, then e' is not a possible state of knowledge for j.*

Proof: Let $e' = I_j(\omega)$ for some ω, and let ω' be a state such that $e' \supseteq I_j(\omega')$. Given that $\omega' \in I_j(\omega')$, we have $I_j(\omega) \cap I_j(\omega') \neq \emptyset$ and therefore $I_j(\omega') = I_j(\omega) = e'$. It follows that e' cannot be the union of two or more *different* states of knowledge for j. ∎

It follows that taking the union of states of knowledge in which an agent has differing knowledge does not result in a state of knowledge in which the agent is more ignorant. It simply does not result in a state of knowledge at all! If this is the case, however, then the technical definition of the sure thing condition does not quite capture the intuition given in the introduction for Savage's sure thing principle. It does not formalize the intuition that "if Alice would arrest the butler if a certain blood stain is typed A, B, AB, or O, (perhaps for different reasons in each case), then she can arrest the butler without bothering to send the stain to the police laboratory." (We will consider the question of how this intuition *can* be captured formally in a later section.) It is easy to check that for a decision function that is defined only on the agent's possible states of knowledge, the STC is never violated, since the decision function is not defined on nontrivial unions. Once we require a decision function to be extendible to such unions, however, the STC does appear to be a nontrivial consistency condition. Except that it no longer seems to capture our original intuition, and its actual meaning is quite unclear.

Returning to the proof of the agreement theorem, we see that it is based in a crucial way on the fact that the decision function D is defined on the event $e = C(d_1 = a) \cap C(d_2 = b)$. Since this event will generally be a nontrivial union of states of knowledge, Lemma 3.2 thus implies that e will in general *not* be a possible state of knowledge for either agent. As we have argued above, the assumption that the agents' decision functions are defined on e and that they satisfy the sure thing condition with respect to e and its subsets cannot be accepted without additional motivation. We conclude that it is no longer obvious that the agreement theorem, as stated, applies to cases such as Bacharach's murder story example.

3.1.2 Like-mindedness

In our story the two detectives are described as "like-minded". By this it is meant that given the same information they would behave in the same way. Technically, this property has been formalized by having both agents follow the same decision function. Indeed, this seems to capture the idea that given the same information, the agents should perform the same action. However, a closer inspection shows that here again there are some unexpected subtleties. Consider for example a situation where for some state ω we have $I_A(\omega) \subseteq K_A(e) \cap K_A \neg K_B(e)$. Let us denote $I_A(\omega)$ by I_a. In I_a, Alice knows that e obtains and at the same time knows that Bob does not know this. Clearly, I_a is not a possible state of knowledge for Bob, because this would require him to know that e obtains and at the same time to know that he (Bob) does not know this. So there is no sense in demanding that when Bob's state of knowledge is I_a he should take the same action that Alice does when she is in I_a. To be more precise, I_a is not a possible state of knowledge for Bob, and hence d_B — Bob's action function — need not be defined on I_a at all. Indeed, it is not hard to check that events that are possible states of knowledge for more than one agent are of a very special form:

Lemma 3.3: *If some event e' is a possible state of knowledge for both Alice and Bob, then* $e' = C_{\{A,B\}}(e')$.

It thus turns out that the vast majority of Alice's states of knowledge are not possible states of knowledge for Bob, and vice versa. As a result, the technical requirement that has been taken to capture the notion of like-mindedness does not capture the idea that Alice and Bob went to the same school and will act similarly given the same evidence.

Let us digress for a moment and consider an argument that has been brought up against our attack on the sure thing condition in the previous subsection. We claimed in Lemma 3.2 that a nontrivial union of an agent's state of knowledge does not result in a state of knowledge that is possible for that agent. One could, nevertheless, imagine a situation where a third agent is added, for whom the union is a possible state of knowledge. The point, however, is that there does not seem to be a good reason to relate what Alice and Bob do when their knowledge state is an element of the union, with what the third agent should do when in the union. In addition, of course, it is possible to prove an analogue of Lemma 3.3 that shows that a third agent's state of knowledge will coincide with the union only under rather particular circumstances.

4 Reformulating the conditions

The intuitive notions of like-mindedness and the sure thing principle are perfectly sensible. Even if, as we have argued, these notions are not formalized properly in [Bac85,Aum89], they definitely deserve to be given alternative formal definitions. This is what we will attempt to do in this section.

Let us start by reconsidering the notion of like-mindedness. When, for example, we say that the detectives in the murder story example are like-minded, we intuitively mean that given the same information they would act in the same way. The question is what we mean by "given the same information". Clearly, an agent's knowledge state determines the information it has. What Lemma 3.3 shows us is that the agents' having the same information is not the same as their having the same knowledge state. Indeed, decision functions usually depend only on certain aspects of the knowledge state. The toothpaste a detective regularly uses may be part of the description of a state ω, but the detective's decision will presumably not depend on this aspect of the world. We will thus want to consider agents as having the same information when their knowledge about the facts relevant to their decision is the same. What these facts are will in general depend on the model and the decision function. We remark that when a decision function truly depends on all aspects of the knowledge state, the lemma shows that the notion of different agents having the "same information" is not well defined in general. We will therefore consider reductions of the complete knowledge state to a state of knowledge about facts that are relevant to the decision; these will be called *states of relevant knowledge.* We will then be able to compare what different agents do when they are in the same state of relevant knowledge. However, to avoid the type of problems that arose in the previous section, a definition of like-mindedness seems to truly capture the notion only when every state of relevant knowledge that is possible for one agent is possible for the other agent as well. The concept of a state of relevant knowledge will also prove useful for reformulating the sure thing condition. In this case, however, we will be able to overcome the problem raised in Lemma 3.2 once we require that the union of states of relevant knowledge is always a possible state of relevant knowledge.

We now formalize the above discussion. A *projection* of Ω is a function $\rho : \Omega \rightarrow Q$, where Q is an arbitrary set. We extend such a projection ρ to events over Ω by defining $\rho(e) = \{\rho(\omega) : \omega \in e\}$. Given a model $(\Omega, I_1, I_2, \ldots)$, a decision function D is said to be *admissible* with respect to a set N of agents if there exists an associated projection ρ_D satisfying the following three conditions:

1. Whenever $\rho_D(e) = \rho_D(e')$, then $D(e) = D(e')$.
2. For all agents $j, j' \in N$ and states ω there exists a state ω' such that $\rho_D(I_j(\omega)) = \rho_D(I_{j'}(\omega'))$.
3. For every nonempty set $e \subseteq \Omega$ and agent $j \in N$ there exists a state ω' such that $\rho_D(I_j(\omega')) = \bigcup_{\omega \in e} \rho_D(I_j(\omega))$.

The first condition demands that the decisions taken according to D depend only on the knowledge about the relevant aspects of the world, as they are projected out by ρ_D. This guarantees that the projection ρ_D does not discard information that is relevant to the decision taken. The second condition says that every state of relevant knowledge that is possible for one agent of N is possible for any other. (N will often be a set of two agents; we do not intend for N to generally be the set of all agents.) Finally, the third condition guarantees that the union of states of relevant knowledge is itself a state of relevant knowledge.[2] A projection ρ_D satisfying the above conditions with respect to D and N we call a *relevance projection* for D and N. We remark that an admissible decision function D may have many different relevance projections (with respect to the same N; as usual, we will not mention the set N when it is clear from context). Notice that for any decision function, the identity projection mapping every state $\omega \in \Omega$ to itself and every event to itself will clearly satisfy the first condition. However, Lemmas 3.2 and 3.3 show that it does not satisfy the other two. The second and third conditions restrict the decision function so that the problems raised in the previous section do not apply.

We can now redefine the notion of like-mindedness as follows:

Definition 4.1: The agents in a set N are said to be *like-minded* if (i) they all follow the same decision function D, and (ii) D is admissible with respect to N.

Recall that, by the first condition, an agent's actions are a function of its state of relevant knowledge. As a consequence, our definition says precisely that like-minded agents act in the same way given the same (relevant) information.

The third condition requires that a union of states of relevant knowledge must itself be a state of relevant knowledge. Recall that our problem with Bacharach's definition of the sure thing condition was caused by the fact that the union of states of knowledge is in general not a state of knowledge (see Lemma 3.2). Thus, whereas unions of states of knowledge did not correspond to a state of knowledge of the disjunction of the states, unions of states of *relevant* knowledge do correspond to disjunctions of the relevant facts known in the states of knowledge being united. We can thus define the sure thing principle with respect to an admissible decision function as follows:

Definition 4.2: An admissible decision function D is said to satisfy the *projected sure thing condition* (PSTC) if there exists a relevance projection ρ_D corresponding to D such that whenever a set $\rho_D(e)$ is a disjoint union of the sets $\rho_D(J)$ for a family of events $\{J\}$ on each of which $D(J) = b$, then also $D(e) = b$.

Having introduced admissible decision functions, and redefined like-mindedness and the sure thing principle with respect to them, we are faced with a number of questions. First of all, we

[2] We make this requirement for ease of exposition. Technically, a slightly weaker condition would suffice. While our choice has the effect of making fewer decision functions admissible, this condition holds in many cases of interest.

should see whether such admissible decision functions arise in common scenarios. This will be done by providing a natural interpretation of the murder story example in which the detectives are like-minded, and follow an admissible decision function satisfying the projected sure thing condition. Once we do this, we will face the question of whether the agreement theorem holds with respect to the new definitions.

5 Agreement Theorem Revisited

We are now in a position to show the following:

Proposition 5.1: *[Disagreement Theorem] Like-minded agents following an admissible decision function that satisfies the projected sure thing condition (PSTC) may agree to disagree.*

In order to prove this proposition, we will construct an example in which the agreement theorem fails. Roughly speaking, the example will be an instance of the murder story; we will first give an informal presentation of the scenario, and later complete the formal details.

> There are initially three main suspects, called α, β, and γ. It is assumed that γ is the toughest of the three, followed by β, while α is the weakest. Therefore, α is more likely than β to cooperate in an interrogation, while β in turn is more likely than γ. The police school in Macho Macho maintains that at most one arrest should be performed in any investigation, in order to keep up the public's trust in the competence and integrity of its police. As a consequence, when it is known that a particular suspect is guilty of a crime, he is immediately brought to trial. When there is a doubt between two suspects, the weaker of the two is brought in and competently interrogated until the police discovers whether he is innocent. Finally, as long as there are more than two suspects, the investigation proceeds without any arrests being performed.
>
> In the particular instance being described, γ is the culprit. By the time Alice and Bob report to the chief of police, Alice has discovered that β is innocent, and is still in doubt whether it was α or γ that was at fault. Bob, on the other hand, was less successfull, and still considers each of the three a suspect. Alice therefore initially intends to suggest that α be arrested, while Bob would suggest that no arrests be made at this stage. Throughout their conversation, Alice insists on suggesting that α be arrested for interrogation, while Bob, not being able to use this information to rule out any of the suspects, insists on suggesting that no arrests be made. After a while it becomes common knowledge that Alice and Bob will stick to these conclusions, and they enter the chief's room having agreed to disagree.

It is obvious that Bob's declaration does not provide Alice with any useful information. Let us consider why Bob gains no relevant information from Alice's declaration. Alice's declaration implies that she has discovered about one of β and γ that he is innocent, and still considers the other one of them and α suspect. However, Bob cannot determine from Alice's declaration whom she has discovered as being innocent. As a consequence, he still considers each of α, β, and γ suspect after her declaration.[3] As a consequence, in their second round of declarations both Alice and Bob will stick to their original declarations. The exact same reasoning and the same declarations repeat themselves in each of the following rounds of Alice and Bob's conversation, until at some point they attain common knowledge of their respective suggestions. We now formalize this example.

5.1 Modeling the murder story scenario

Before we turn to the task of modeling the particular example described above, we wish to enrich Aumann's model of knowledge to one in which the internal structure of the states of the world is made more explicit. A state of the world consists of local states of a distinguished set of agents, and the state of the rest of the world, which we call the *environment*. (The environment will often contain additional agents; the agents that are singled out are typically the ones in whose states and knowledge we are most interested.) More formally, following work on reasoning about knowledge in distributed systems [HM89,Hal87], we will associate with every state of the world ω an *internal structure* $int(\omega)$ of the form $(\ell_1, \ldots, \ell_n, \ell_{env})$, where ℓ_j represents the local state of agent j, and ℓ_{env} represents the state of the environment. We will denote agent j's local state at $int(\omega)$ by ω_j, and the environment's state by ω_{env}. The actual states of the agent and of the environment will in general themselves have additional structure, depending on the particulars of the scenario being modeled.

Once we ascribe to each state of the world such internal structure, we can derive the information functions I_j directly from the internal structure, based on the agents' local states. Two worlds will be considered indistinguishable by agent j exactly if she has the same local state in both. Formally, we define

$$I_j(\omega) = \{\omega' : \omega_j = \omega'_j\}.$$

In effect, local states replace information functions as primitives of the model. We remark that adding such internal structure to states of the world and deriving the information functions in this manner can be done without loss of generality to any model of knowledge of the type described in Section 2. Specifically, assuming there are distinct names for all the states of the world, and for all the possible states of knowledge, we can simply take ω_j to be the name of $I_j(\omega)$, and take ω_{env} to be the name of ω. This construction yields a model isomorphic (with respect to knowledge) to the original one.

In the murder story application, the agents we wish to study carefully are Alice and Bob. We will thus associate with every state of the world an internal structure of the form $(\ell_A, \ell_B, \ell_{env})$. All relevant facts other than Alice and Bob's states will be modeled by the environment. Of

[3] The intuitive reason why Bob cannot suggest to interrogate α at this point is that if Bob would now interrogate α and discover that α is innocent, he could not say who should be indicted without a second arrest.

course, this environment may very well contain other agents related to the murder case, may involve their knowledge, etc. (The choice is up to us. We could just as well have chosen to model these additional agents as first rate citizens with their own local states. There are many ways to slice a cake.) Thinking in terms of agents' local states has the advantage that we can often define these states in a natural manner based on the scenario we are considering. This allows a careful study of what knowledge each agent obtains at any given point, and of how knowledge evolves over time.

The murder story scenario seems to involve three essential steps. The first stage involves the murder and all events up to the point at which the chief of police puts Alice and Bob on the case. The second stage is the detectives' inquiry stage, in which they each investigate various aspects of the murder case. Finally, the third case consists of their conversation in the locker room, resulting in their final report to the chief. Each of the detectives enters the second stage with a particular local state, which includes the wisdom gained in police school, and may also include personal tendencies as to how to perform the investigation. By the end of the second stage, the local state of each detective also contains the evidence s/he has obtained. In the third stage, Alice and Bob compare notes. They alternate turns in each announcing the action it would take given its current knowledge. Following each announcement, the contents of this announcement are appended to both agents' local states. We are thus assuming that Alice and Bob do not forget any of their discussion.[4] Notice that as a consequence of our defining an agent's knowledge based on its local state, the agents' relevant knowledge is updated automatically. Our model, taken from the distributed systems methodology, thus incorporates the evolution of agents' knowledge in a manner that is directly related to the change in the information they maintain. We remark that equivalent methods of updating agents' knowledge in the course of such a conversation are defined explicitly in [GP82,Mak83,Cav83,Bac85,PK87].

5.2 The counterexample

What remains to be done in order to complete the formal proof of Proposition 5.1 is to (i) describe the detectives' decision function, (ii) show that it is admissible and satisfies PSTC, and (iii) show that Alice and Bob agree to disagree in the particular scenario presented.

The detectives' decision function D can easily be described by specifying its relevance projection ρ_D, and by specifying the associated function D^r, whose domain consists of the sets $\rho_D(e)$. Formally, we will have $D(e) = D^r(\rho_D(e))$. Given a state ω of the world, we define $\rho_D(\omega)$ to be the actual murderer in the state ω, if the murder has already taken place in ω's

[4] This is purely an assumption we make for simplicity in this particular example. As is well-known, we could model forgetting in various ways too. For example, we could capture the idea that an agent only remembers the outcome of its test and the three latest announcements by having the local states contain exactly that information.

history. The definition of D^r is as follows:

$$\begin{array}{lll} D^r(\{\alpha\}) = \mathit{indict}(\alpha) & D^r(\{\beta,\gamma\}) = \mathit{arrest}(\beta) & \\ D^r(\{\beta\}) = \mathit{indict}(\beta) & D^r(\{\alpha,\gamma\}) = \mathit{arrest}(\alpha) & D^r(\{\alpha,\beta,\gamma\}) = \emptyset \\ D^r(\{\gamma\}) = \mathit{indict}(\gamma) & D^r(\{\alpha,\beta\}) = \mathit{arrest}(\alpha) & \end{array}$$

The set Q is thus $\{\alpha,\beta,\gamma\}$. An immediate corollary of a theorem of Geanakoplos and Polemarchakis in [GP82] states that if Q is finite, then Alice and Bob are guaranteed to reach common knowledge of the actions each one of them intends to perform, after a finite number of rounds.

Let us check that D is an admissible decision function. By defining D based on ρ_D, we have guaranteed that it satisfy the first admissibility condition. The second admissibility condition corresponds to the assumption, which we are implicitly making, that any subset of $\{\alpha,\beta,\gamma\}$ is a possible state of relevant knowledge for each of the detectives. Finally, the third condition is satisfied because D^r is defined on all nonempty subsets of Q and hence on all unions of states of relevant knowledge. It follows that D is admissible. The fact that D satisfies the projected sure thing condition is trivially satisfied, as there are no disjoint states of relevant knowledge on which the same decision is taken. (The only two state of relevant knowledge on which the same action is taken are $\{\alpha,\beta\}$ and $\{\alpha,\gamma\}$, but these are not disjoint.)

The actual scenario in the counterexample is one in which Alice finishes her inquiry (and enters the locker room) with $\{\alpha,\gamma\}$ as her state of relevant knowledge, while Bob has $\{\alpha,\beta,\gamma\}$. A simple proof by induction on the number of rounds now shows that in this particular example, Alice will forever have $\{\alpha,\gamma\}$ as her state of relevant knowledge, and hence will insist on deciding to arrest α, while Bob will have $\{\alpha,\beta,\gamma\}$ as his state, and will insist on continuing the investigation with no arrests. Now, since their decisions will become common knowledge after a finite number of steps, they will in fact end up *agreeing to disagree*!

5.3 Sufficient conditions for agreement

We now turn to study sufficient conditions for the agreement theorem. In order to do so, let us start by reconsidering why the theorem fails in the counterexample presented above. The point seems to be that whereas different states of knowledge of a given agent are guaranteed to be disjoint, different states of relevant knowledge are not guaranteed to be disjoint. However, in the spirit of STP and STC, union consistency is required by PSTC only for disjoint unions. Looking at the details of the example, Bob's state of relevant knowledge when they agree to disagree is $\{\alpha,\beta,\gamma\}$, and he remains uncertain whether Alice's state of knowledge is $\{\alpha,\beta\}$ or $\{\alpha,\gamma\}$. Observe that $D^r(\{\alpha,\beta\}) = D^r(\{\alpha,\gamma\}) = \mathit{arrest}(\alpha)$, while $D^r(\{\alpha,\beta\} \cup \{\alpha,\gamma\}) = D^r(\{\alpha,\beta,\gamma\}) = \emptyset$. We are thus led to the question of whether modifying PSTC to apply to arbitrary unions can salvage the agreement theorem for scenarios such as the murder story. Define an admissible decision function D to satisfy the *inclusive projected sure thing condition* (IPSTC) if whenever a set $X \subseteq Q$ is a union of a family of (not necessarily disjoint) subsets $\{Y\}$, on each of which $D^r(Y) = b$, then also $D^r(X) = b$. We can show:

Theorem 5.2: *Like-minded agents following an admissible decision function satisfying the inclusive projected sure thing condition (IPSTC) cannot agree to disagree.*

Proof: As in the proof of Theorem 3.1, set $e = C(d_1 = a) \cap C(d_2 = b)$. Again, because e is common knowledge, $K_1(e) = e$ and therefore e is a (disjoint) union of elements of the form $I = I_1(\omega)$. By definition of e, in every $\omega \in e$ we have $d_1(\omega) = a$, and hence $D^r(\rho_D(I_1(\omega))) = a$. Now, since e is a (nonempty) union of such sets I, we have by definition of ρ_D that $\rho_D(e)$ is the union of the corresponding sets $\rho_D(I)$. The fact that D satisfies the inclusive projected sure thing condition now implies that $D(e) = D^r(\rho_D(e)) = a$. Similar reasoning with respect to agent 2 shows that $D(e) = b$, and as a result we obtain $a = b$. ∎

The inclusive projected sure thing condition is clearly stronger than PSTC. However, for decision functions that do not originate from probabilistic reasoning, it does not seem to be much harder to accept the inclusive condition than it is to accept the original (exclusive) PSTC. Indeed, on second thought, it is not clear why disjointness appears in the blood stains motivation for STP quoted in the introduction. In non-probabilistic settings, restricting attention to disjoint unions is often quite unnatural. (In cases when a decision function is based on posterior probabilities or expectations, the disjointness requirement arises in a natural way.) Indeed, we consider the inclusive projected sure thing condition to be the more natural condition in the non-probabilistic case. We therefore conclude that, once the definitions are formulated appropriately and STP is replaced by inclusive STP, the agreement theorem does hold in the context of the murder story. In this case the statement that if the detectives use a reasonable decision function they cannot agree to disagree is justified. This also explains why our counterexample may seem somewhat contrived or unnatural: Any non-probabilistic decision function that satisfies PSTC but not IPSTC is likely to be somewhat unnatural.

It is interesting to note that PSTC does suffice when the projections of disjoint states of knowledge are guaranteed to be disjoint. We can thus state the following proposition:

Proposition 5.3: *Let $\Pi = \{1,2\}$, and let $e = C(d_1 = a) \cap C(d_2 = b)$. Assume that (i) the agents are like-minded, (ii) they follow an admissible decision function D satisfying PSTC, and (iii) for $j \in \Pi$ and all $\omega, \omega' \in e$ it is the case that whenever $I_j(\omega) \cap I_j(\omega') = \emptyset$, then also $\rho_D(I_j(\omega)) \cap \rho_D(I_j(\omega')) = \emptyset$. Then*

$$[C(d_1 = a) \cap C(d_2 = b) \neq \emptyset] \quad \Rightarrow \quad a = b.$$

We can think of the requirement here as a combined condition, applying both to the decision function *and* to the protocol the agents are using for their interaction. It would be interesting to study other combined conditions that yield the agreement theorem.

5.4 More than two agents

Our final topic involves the question of how the analysis presented above applies to the case of more than two agents. It turns out that there are at least a couple of interesting types of scenarios that are worth considering in this case. One of them is the direct generalization of the agreement theorem to agreement among $n > 2$ agents. This corresponds to the case in which, say, three detectives, Alice, Bob, and Charlie are assigned to the case, and all three meet in the locker room and carry out a joint discussion of their intended actions. Everything any one of them says in this discussion immediately becomes common knowledge to all three. It is not hard to show that the results of [Aum76,Bac85,Cav83] and our above results all extend immediately to more than two agents in this case. However, in [PK87] Parikh and Krasucki considered a somewhat more practical situation in which the agents communicate in pairs, and the contents of a message are known only to the sender and the reciever of the message. They assume, however, that the communication protocol being used is *fair*,[5] and that the decision functions are real valued.

Definition 5.4: A decision function D is said to be *weakly convex* if for all $X, Y \subseteq \Omega$ we have

$$X \cap Y = \emptyset \Rightarrow \quad D(X \cup Y) = a \cdot D(X) + b \cdot D(Y),$$

for some $a, b \geq 0$ such that $a + b = 1$. D is *strongly convex* if it guarantees that $a, b > 0$.

Parikh and Krasucki show that the agreement holds for fair protocols with decision functions satisfying STC when there are $n = 2$ agents, and that this is not the case for $n > 2$. Weak convexity of D is a sufficient condition for the agreement theorem in the case of $n = 3$ agents, but not for $n > 3$. Finally, the agreement theorem holds in the context of fair protocols for $n \geq 4$ when D is strongly convex. We now review these results in light of our analysis.

The weak and strong convexity conditions, like the STC, are applied to disjoint unions of subsets of Ω. They therefore suffer from the same shortcomings as STC with respect to their intuitive meaning for deterministic or generally non-probabilistic decision functions. For the case of $n = 3$ we can obtain a positive result similar to Theorem 5.2. We define a decision function D to satisfy *inclusive weak convexity* if for all sets $X, Y \subseteq \Omega$ (not necessarily disjoint) it is the case that $D(X \cup Y) = a \cdot D(X) + b \cdot D(Y)$, for some $a, b \geq 0$ such that $a + b = 1$. An appropriate modification of the proof of [PK87] yields:

Proposition 5.5: *Three like-minded agents using a fair protocol and a continuous admissible decision function satisfying inclusive weak convexity cannot agree to disagree.*

It is possible to attempt to do the same for strong convexity and $n \geq 4$. Again, we would relax the disjointness requirement and obtain the notion of inclusive strong convexity. This, however, yields a rather meaningless result, as the following lemma shows:

[5] By a *protocol* they mean an a priori fixed list of sender-receiver pairs, specifying the order in which messages are sent. Given such a protocol, consider a graph whose vertices are the agents, where there is a directed edge from vertex i to vertex j exactly if i sends j an infinite number of messages according to the protocol. They call a protocol *fair* if its corresponding graph thus constructed is strongly connected.

Lemma 5.6: *A decision function D satisfying inclusive strong convexity is necessarily a constant function.*

Proof: For any $X \subseteq \Omega$ we have

$$D(\Omega) = D(X \cup \Omega) = a \cdot D(X) + b \cdot D(\Omega).$$

Since $a > 0$ and $a + b = 1$, we get $D(X) = D(\Omega)$. ∎

We are therefore unable to suggest any nontrivial positive result for fair protocols in the case $n \geq 4$.

6 Conclusions

We argued that formal definitions of the sure thing condition and the notion of like-mindedness used in [Aum89,Bac85,Cav83] do not capture their intended intuitive meaning. As a result, the agreement theorem of [Bac85,Cav83] is perhaps less widely applicable than was believed. It is important to note, however, that this theorem is both technically correct and of considerable practical significance. While not quite applying to arbitrary decision functions, it extends Aumann's original agreement theorem [Aum76] from agreeing on posteriors to many other interesting agreements, such as expectations, actions that maximize conditional expectations, etc. It seems not to directly apply in the presence of (non-probabilistic) nondeterminism in the agents' actions. In particular, it does not apply to the murder story example.

The reason that the technical definitions of like-mindedness and STC fail to capture their intended meaning seems to be that states of knowledge are a special subclass of the set of all events. This set is not closed under operations such as taking union, nor is it guaranteed to be symmetric with respect to the agents. Some of the structure of this set is more transparent once we model the internal structure of states of the world explicitly, as is done in the distributed systems literature, rather than implicitly, as in the game theoretic literature. The explicit modeling has the additional advantage that it provides useful machinery for the formalization of particular examples, and makes the evolution of knowledge over time more apparent.

Our definition of admissible decision functions and PSTC made it possible for us to consider the agreement theorem in the context of scenarios such as the murder story example. It is interesting to note, however, that it is not the case that posterior probability is a special case of an admissible decision function satisfying PSTC. (Recall that the STC is a proper generalization of posterior probability.) The reason for this is that our relevant knowledge states are constructed in a way that does not maintain information about the probabilities of the events known about. Indeed, the interaction of relevant knowledge and probability in this context is as yet somewhat unclear to us. We believe it deserves to be studied further. Perhaps the framework of Halpern and Tuttle in [HT89] could prove useful here.

We do not consider the results reported on in this abstract as showing that the claim made in the murder story is philosophically wrong. On the contrary, once the definitions are reformulated, and we consider an *inclusive* version of the sure thing principle, the agreement theorem is maintained. Our results mainly raise issues regarding how one should formalize concepts related to, and in the presence of, knowledge. They imply that Bacharach and Cave's original definitions and theorem do not apply to the murder story. It remains an interesting open problem whether there is a natural formulation of the agreement theorem that generalizes both our results and those of Bacharach and Cave. Such a theorem would at once apply both to probabilistic and to non-probabilistic situations, and would justify Bacharach's claim that "*differences in information alone cannot account for differences in behavior of rational persons ... any more than they can account for differences of opinion*" [Bac85].

Acknowledgments

We'd like to thank Robert Aumann for useful discussions and for comments on an earlier version of this abstract. His comments improved the presentation of the material in a substantial way. We also thank Yishai Feldman, Joe Halpern, and Moshe Vardi for useful discussions on the topic of this paper. Special thanks to Gadi Taubenfeld for drawing our attention to [Aum89], and thereby starting us off on this research project.

References

[Aum76] R. J. Aumann, Agreeing to Disagree, *The Annals of Statistics*, Vol 4 No 6, 1976, pp. 1236-1239.

[Aum89] R. J. Aumann, Notes on Interactive Epistemology, unpublished manuscript, version 89.04.06, 1989.

[Bac85] M. Bacharach, Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge, *Journal of Economic Theory*, Vol 37, 1985 pp. 167-190.

[Cav83] J. A. K. Cave, Learning to Agree, *Economics Letters*, Vol 12, 1983, pp. 147-152.

[GP82] J. D. Geanakoplos and H. M. Polemarchakis, We Can't Disagree Forever, *Journal of Economic Theory*, Vol 28, 1982, pp. 192-200.

[Hal87] J. Y. Halpern, Using reasoning about knowledge to analyze distributed systems, *Annual Review of Computer Science*, Vol 2, J. Traub et al. eds., 1987, pp. 37-68.

[HM85] J. Y. Halpern and Y. Moses, A Guide to the Modal Logics of Knowledge and Belief, *Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, 1985, pp. 480-490.

[HM89] J. Y. Halpern and Y. Moses, Knowledge and Common Knowledge in a Distributed Environment, *IBM RJ 4421*, 4th revision, September 1989. An early version appeared in *Proceedings of the 3rd ACM Symposium on Principles of Distributed Computing*, 1984, pp. 50-61.

[HT89] J. Y. Halpern and M. R. Tuttle, Knowledge, Probability, and Adversaries, *Proceedings of the 8th ACM Symposium on Principles of Distributed Computing*, 1989, pp. 119-128.

[Mak83] L. Makowski, Common Knowledge and Common Learning, unpublished manuscript, UC Davis, June 1983.

[PK87] R. Parikh and P. Krasucki, Communication, Consensus and Knowledge, unpublished manuscript, 1987.

[Sav54] L. J. Savage, *The Foundations of Statistics*, John Wiley and sons, 1954. 2nd revised edition by Dover Publications, 1972.