Counterfactuals in "agreeing to disagree" type results[☆]

Bassel Tarbush

Merton College, University of Oxford
Merton Street, Oxford, OX1 4JD, UK

Abstract

Moses and Nachum (1990) identified conceptual flaws (later echoed by Samet, 2010) in Bacharach's (1985) generalization of Aumann's (1976) seminal "agreeing to disagree" result by demonstrating that the crucial assumptions of like-mindedness and the Sure-Thing Principle are not meaningfully expressible in standard partitional information structures. This paper presents a new agreement theorem couched in "counterfactual information structures" that resolves these conceptual flaws. The new version of the Sure-Thing Principle introduced here, which accounts for beliefs at counterfactual states, is also shown to sit well with the intuition of the original version proposed by Savage (1972).

Keywords: Agreeing to disagree, counterfactuals, knowledge, belief
JEL: C70, D83

1. Introduction

Aumann (1976) showed that if agents' posterior beliefs over some event, which are obtained from updating over private information, are commonly

[☆]A preliminary version of this paper appeared as an extended abstract under the title "Agreeing on decisions: an analysis with counterfactuals" in the *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge*.

Email address: bassel.tarbush@economics.ox.ac.uk (Bassel Tarbush)

Final accepted version: November 2016.

known, then these beliefs must be the same. Bacharach (1985) and Cave (1983), independently, were the first to generalize this seminal "agreeing to disagree" result to the non-probabilistic case.[1] Specifically, Bacharach isolated the relevant properties that hold both of conditional probabilities and of the common prior assumption – which drive the original result – and imposed them as independent conditions on general decision functions in partitional information structures. As such, he was able to isolate and interpret the assumptions underlying Aumann's original result as (i) an assumption of "like-mindedness" and, (ii) an assumption that he claimed is analogous to requiring the agents' decision functions to satisfy Savage's Sure-Thing Principle (Savage, 1972).

Moses and Nachum (1990) and Samet (2010) pointed out that Bacharach's result is conceptually problematic because like-mindedness and the Sure-Thing Principle do not have a well-defined informational content in partitional information structures. Indeed, in partitional information structures, the partition elements are the informational primitives and only *they* have a well-defined informational content. Now, like-mindedness requires agents with the same information to take the same action. Technically, Bacharach's condition requires defining the decision function of an agent i over the partition elements of an agent j, but the informational content that j's partition element has for i is not well-defined since different agents will typically have different information partitions. An agent's decision function is said to satisfy the Sure-Thing Principle if whenever the decision over each element of a set of disjoint events is x, the decision over the union of all those events is also x. Since the union of events is intended to capture the notion of being "more ignorant", the principle is understood as capturing the intuition that *If an agent i takes the same action in every case in which i is better informed, i takes the same action in the case in which i is more ignorant.* However, there is no partition element that corresponds to a union of partition elements for any agent, so the informational content of the union (over which an agent's decision function is defined) is not well-defined.

This paper provides a simple solution to the conceptual flaws outlined above by developing a method that adds "counterfactual states" to any partitional information structure. New versions of like-mindedness and of the Sure-Thing Principle are defined and are shown to be meaningfully express-

[1] Bonanno and Nehring (1997) and Ménager (2007) survey the "agreeing to disagree" literature. Bach and Cabessa (2011), Dégremont and Roy (2012), Bach and Perea (2013), Dominiak and Lefort (2013), Heifetz et al. (2013), Demey (2014), and Bach and Cabessa (2016) are recent examples of results in this literature.

ible within the resulting counterfactual structure. Notably, there is no requirement to define an agent's decision function over another agent's partition elements, and the notion of being "more ignorant" (as captured by the union of partition elements) has a well-defined informational content within the structure. A new agreement theorem is proved within counterfactual structures that therefore avoids the conceptual flaws found in Bacharach (1985).

Furthermore the new version of the Sure-Thing Principle introduced here is shown to capture the intuition that: *If an agent i takes the same action in every case in which i is better informed, i takes the same action in the case in which i is secretly "just" more ignorant.* It is argued that this version sits well with the intuition of the original version proposed by Savage (1972).

Finally, the solution to the conceptual flaws presented in this paper is compared with the solutions of Moses and Nachum (1990), of Aumann and Hart (2006), and of Samet (2010).

Section 2 presents the basic notation and Section 3 outlines Bacharach's original result and its conceptual flaws formally. Section 4 introduces the counterfactual structures and analyzes their properties. Section 5 provides new definitions for the Sure-Thing Principle and for like-mindedness, proves a new agreement theorem within counterfactual structures, and shows how the approach resolves the conceptual flaws. Section 6 relates the approach presented here with other solutions proposed in the literature.

2. Information structures: notation

Let Ω denote a finite set of *states* and N a finite set of agents. A subset $e \subseteq \Omega$ is called an *event*. For every agent $i \in N$, define a binary relation $R_i \subseteq \Omega \times \Omega$, called a *reachability* relation. The state $\omega \in \Omega$ is said to *reach* the state $\omega' \in \Omega$ if $\omega R_i \omega'$. In terms of interpretation, if $\omega R_i \omega'$, then at ω, agent i considers the state ω' *possible*. An *information structure* $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$ is entirely determined by the state space, the set of agents, and the reachability relations. A *possibility set* at state ω for agent $i \in N$ is defined by

$$b_i(\omega) = \{\omega' \in \Omega | \omega R_i \omega'\} \tag{1}$$

A possibility set $b_i(\omega)$ is therefore the set of all states that i considers possible at ω. In terms of notation, let $\mathcal{B}_i = \{b_i(\omega) | \omega \in \Omega\}$. For any $e \subseteq \Omega$, define the *simple epistemic operator* by

$$B_i(e) = \{\omega \in \Omega | b_i(\omega) \subseteq e\} \tag{2}$$

For any $e \subseteq \Omega$, and any $G \subseteq N$, define the *common epistemic operator* by,

$$C_G(e) = \cap_{m=1}^{\infty} M_G^m(e) \tag{3}$$

where $M_G^1(e) = \cap_{i \in G} B_i(e)$ and $M_G^{m+1}(e) = M_G(M_G^m(e))$ for $m \geq 1$. A state $\omega' \in \Omega$ is said to be *reachable* among the agents in G from a state $\omega \in \Omega$ if there exists a sequence of states $\omega \equiv \omega_0, \omega_1, \omega_2, ..., \omega_n \equiv \omega'$ such that for each $k \in \{0, 1, ..., n-1\}$, there exists an agent $i \in G$ such that $\omega_k R_i \omega_{k+1}$. The *component* $T_G(\omega)$ of the state ω is the set of all states that are reachable among the agents in G from ω. The common epistemic operator can be given an alternative characterization (which is standard and for example follows Fagin et al., 1995 or Hellman, 2013, p. 12),

$$C_G(e) = \{\omega \in \Omega | T_G(\omega) \subseteq e\} \tag{4}$$

No particular restrictions have yet been imposed on the reachability relations. But it is precisely the restrictions on these relations that determine the properties that the epistemic operators satisfy and that therefore determine the proper interpretation of the operators. Information structures $\mathcal{S}$ in which the reachability relations are equivalence relations (i.e. reflexive and Euclidean) are the standard "knowledge" or *partitional* structures.[2] (Information structures, as defined above, are therefore more general than standard partitional structures.) In partitional structures the set $\mathcal{B}_i$ partitions the state space, the possibility sets are the partition elements of $\mathcal{B}_i$, and the simple epistemic operator $B_i(.)$ is interpreted as a *knowledge* operator because it satisfies the well-known properties that conventionally characterize knowledge: Kripke, Consistency, Truth, Positive Introspection, and Negative Introspection (Aumann, 1976; Bacharach, 1985; Aumann, 1999).[3] Similarly, $C_G(.)$ is then interpreted as the familiar *common knowledge* operator.

[2]The reachability relations $\{R_i\}_{i \in N}$ are said to be serial if $\forall i \in N$, $\forall \omega \in \Omega$, $\exists \omega' \in \Omega$, $\omega R_i \omega'$, reflexive if $\forall i \in N$, $\forall \omega \in \Omega$, $\omega R_i \omega$, transitive if $\forall i \in N$, $\forall \omega, \omega', \omega'' \in \Omega$, if $\omega R_i \omega' \,\&\, \omega' R_i \omega''$, then $\omega R_i \omega''$, and Euclidean if $\forall i \in N$, $\forall \omega, \omega', \omega'' \in \Omega$, if $\omega R_i \omega' \,\&\, \omega R_i \omega''$, then $\omega' R_i \omega''$.

[3]The properties are given textbook treatments, for example in Fagin et al. (1995); Chellas (1980), or van Benthem (2010), but are listed here for convenience (noting that for any $e \subseteq \Omega$, $\neg e$ denotes the set $\Omega \backslash e$): The Kripke property states that if an agent i knows that e and knows that e implies f, then i must also know that f $(B_i(\neg e \cup f) \cap B_i(e) \subseteq B_i(f))$. The Consistency property states that if an agent i knows that e, then i cannot also know that not e $(B_i(e) \subseteq \neg B_i(\neg e))$. The Truth property states that if an agent i knows that e, then e must be true $(B_i(e) \subseteq e)$. The Positive Introspection property states that if an agent i knows that e, then i knows that i knows that e $(B_i(e) \subseteq B_i(B_i(e)))$, and the Negative Introspection property states that if an agent i does not know that e, then i knows that i does not know that e $(\neg B_i(e) \subseteq B_i(\neg B_i(e)))$.

The epistemic operators were introduced above without particular restrictions on the reachability relations to allow for the same formal definitions to be used in the context of non-partitional structures. Notably, the *counterfactual* structures presented in Section 4 are non-partitional and the operators $B_i(.)$ and $C_G(.)$ will then be interpreted as *belief* and *common belief* operators respectively.

3. The original result and its conceptual flaws

The original result of Bacharach (1985) was derived in a partitional information structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$. For every agent $i \in N$, an *action function* δ_i maps from the state space Ω to a set $\mathcal{A}$ of actions. The function $\delta_i : \Omega \to \mathcal{A}$ therefore specifies agent i's action at any given state. A *decision function* D_i for agent i, maps from a field $\mathcal{F}$ of subsets of Ω into the set $\mathcal{A}$ of actions. That is,

$$D_i : \mathcal{F} \to \mathcal{A} \tag{5}$$

Following the terminology of Moses and Nachum (1990), an agent i using the action function δ_i is said to *follow* the decision function D_i if for all states $\omega \in \Omega$, $\delta_i(\omega) = D_i(b_i(\omega))$.

Definition 1. The decision function D_i of agent i satisfies the *Sure-Thing Principle* if whenever for all $e \in \mathcal{E}$, $D_i(e) = x$ then $D_i(\cup_{e \in \mathcal{E}} e) = x$, where $\mathcal{E} \subseteq \mathcal{F}$ is a non-empty set of disjoint events.

One can think of an event as representing some information and of a decision over that event as determining the action that is taken as a function of that information. The union of events "coarsens" the information and is intended to capture the idea of being "more ignorant". Bacharach's Sure-Thing Principle is therefore intended to capture the intuition that *If an agent i takes the same action in every case in which i is better informed, i takes the same action in the case in which i is more ignorant.*

Definition 2. Agents are said to be *like-minded* if they have the same decision function. That is, for any any agents i and j, and any $e \in \mathcal{F}$, $D_i(e) = D_j(e)$.

That is, over the same subsets of states, the agents take the same action if they are like-minded. This is intended to capture the intuition that given the same information, the agents would take the same action.

Bacharach's agreement theorem can be stated as follows: Let $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$ be a partitional structure. If the agents $i \in N$ are like-minded (Definition 2) and follow the decision functions $\{D_i\}_{i\in N}$ (as defined in (5)) that satisfy the Sure-Thing Principle (Definition 1), then for any $G \subseteq N$, if $C_G(\cap_{i\in G}\{\omega' \in \Omega | \delta_i(\omega') = x_i\}) \neq \emptyset$, then $x_i = x_j$ for all $i, j \in G$. That is, if the action taken by each member of a group of like-minded agents who follow decision functions that satisfy the Sure-Thing Principle is common knowledge among that group, then the members of the group must all take the same action.

3.1. The conceptual flaws

Definition 3. Let $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$ be some arbitrary information structure. An event e is a *possible belief* for agent i in $\mathcal{S}$ if there exists a state $\omega \in \Omega$ such that $e = b_i(\omega)$.

When $\mathcal{S}$ is a partitional structure, this definition corresponds exactly to e being what Moses and Nachum (1990) call a "possible state of knowledge". The authors identify conceptual flaws in Bacharach's result by showing that (1) the Sure-Thing Principle forces decisions to be defined over unions of possibility sets, but no union of possibility sets can be a possible belief for any agent (see Moses and Nachum, 1990, Lemma 3.2), and (2) the assumption of like-mindedness forces the decision function of an agent i to be defined over the possibility sets of agents $j \neq i$, but – other than the case when the sets correspond trivially – these are not possible beliefs for agent i (see Moses and Nachum, 1990, Lemma 3.3).[4] In other words, possible beliefs are the informational primitives in any information structure and only *they* have a well-defined informational content, but Bacharach's framework requires the decision function of an agent to be defined over events (such as the union of partition elements or partition elements of other agents) that are not possible beliefs for that agent.

[4] The proofs of Lemmas 3.2 and 3.3 in Moses and Nachum (1990) are reproduced here for convenience. In each case, suppose that the information structure is partitional. Lemma 3.2: Suppose an event $e \subseteq \Omega$ is a possible belief for agent i, so $e = b_i(\omega)$ for some $\omega \in \Omega$, and suppose $b_i(\omega') \subseteq e$ for some $\omega' \in \Omega$. The set $b_i(\omega')$ is not empty (indeed $\omega' \in b_i(\omega')$ since the structure is partitional) so $b_i(\omega) \cap b_i(\omega') \neq \emptyset$. Since the structure is partitional, is follows that $b_i(\omega) = b_i(\omega')$, so e cannot be the union of two or more distinct possible beliefs. Lemma 3.3: The only events that are possible beliefs for multiple agents must also be commonly known. Indeed, if $e = b_i(\omega) = b_j(\omega')$ then $e = C_{\{i,j\}}(e)$.

4. Counterfactual structures

This section presents a method which augments any partitional structure with counterfactual states. Section 4.1 presents a motivating example that anticipates the formal definition of the resulting counterfactual structures (Section 4.2) and the discussions relating to their properties (Sections 4.3 to 4.6).

4.1. Motivating example

Consider a partitional information structure $\mathcal{S} = (\Omega, \{a\}, R_a)$ in which there is a single agent a, the state space is given by $\Omega = \{\omega_1, \omega_2\}$, and the reachability relation for a is given by $R_a = \{(\omega_1, \omega_1), (\omega_2, \omega_2)\}$. This is a structure in which agent a knows that the state is ω_1 whenever the state is ω_1 and knows that it is ω_2 whenever the state is ω_2. For sake of argument, suppose that ω_1 is a state in which it is very sunny and ω_2 is a state in which it is raining heavily, and furthermore suppose that in each case, agent a's action is to take an umbrella – to shade herself from the sun in one case and to protect herself from the rain in the other. Since a takes an umbrella when she knows that it is very sunny and takes an umbrella when she knows that it is raining heavily, she must, according to the Sure-Thing Principle, take an umbrella when she believes that it is very sunny or raining heavily but does not know which. But as discussed in Section 3.1, Moses and Nachum (1990) rightly point out that since there is no possible belief for a that corresponds to the union of ω_1 and ω_2, there is no possible belief for a within the partitional structure $\mathcal{S}$ that captures the information "a believes that it is very sunny or raining heavily but does not know which". However, such a possible belief *can* exist in an extended structure in which states are added to $\mathcal{S}$ at which a no longer distinguishes between ω_1 and ω_2.

Specifically, let us consider an extended structure $\mathcal{S}' = (\Sigma, \{a\}, R'_a)$, referred to as the *counterfactual structure* of $\mathcal{S}$, in which the extended state space $\Sigma = \Omega \cup \Lambda$ consists of the the original states Ω plus a set of counterfactual states $\Lambda = \left\{\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}, \lambda_{a,\omega_2}^{\{\omega_1,\omega_2\}}\right\}$, and the new reachability relation R'_a consists of the original relation R_a plus a set of links from the counterfactual states to the original states. The formal definition of counterfactual structures is given in Section 4.2, but applied to this case, the new reachability relation is given by

$$R'_a = R_a \cup \left\{\left(\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}, \omega_1\right), \left(\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}, \omega_2\right), \left(\lambda_{a,\omega_2}^{\{\omega_1,\omega_2\}}, \omega_1\right), \left(\lambda_{a,\omega_2}^{\{\omega_1,\omega_2\}}, \omega_2\right)\right\}$$

Given the construction of R'_a, the counterfactual state $\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}$ is interpreted as the state in which a is made "more ignorant" relative to ω_1 (if the actual

state were ω_1) by not distinguishing between the states ω_1 and ω_2.[5] Indeed, since this counterfactual state reaches both ω_1 and ω_2, a considers both of the original states to be possible. In fact, the possibility set at the counterfactual state is precisely the union of ω_1 and ω_2. Therefore, there *is* a possible belief in the counterfactual structure $\mathcal{S}'$ that captures the information "a believes that it is very sunny or raining heavily but does not know which". The formal properties that are satisfied by epistemic operators in counterfactual structures are analyzed in Section 4.3 and the precise formal meaning of "more ignorance" is analyzed in Section 4.4.

Crucially, agent a always excludes the existence of counterfactual states. Indeed, no original state in Ω reaches a counterfactual state and no counterfactual state reaches itself via the reachability relation R'_a. This implies that a cannot be interpreted as somehow *imagining* herself to be more ignorant at a counterfactual state. Rather, if ω_1 were the actual state, then a simply *is* more ignorant at the counterfactual state $\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}$ than at ω_1. Counterfactual states are therefore auxiliary states that allow for the formal inclusion of a type of information ("more ignorance") that is missing in the standard partitional structure, and counterfactual structures allow for a formal characterization of what is meant by "more ignorance". Since counterfactual states are excluded by the agents themselves, they can be considered to be "counterfactual" from the point of view of the modeler. The Sure-Thing Principle within counterfactual structures is later shown to be interpretable, in line with Savage (1972) or Bacharach (1985), broadly as: "If an agent i takes the same action in every case when i is better informed, i takes the same action in the case when i *is* more ignorant".[6] If counterfactual states were not excluded by the agents, then the principle may have to be interpreted as "If an agent i takes the same action in every case when i is better informed, i takes the same action in the case when i *imagines* him/herself to be more ignorant". In our example, it seems convincing to think that a should take an umbrella when she does not know whether it is raining heavily or very sunny (but knows it is one or the other). It is relatively less convincing to think that a should take an umbrella when she *imagines* herself not to know whether it is raining heavily or very sunny.

The definition of counterfactual structures given in Section 4.2 allows for structures with multiple agents. If a second agent b were added to the counterfactual structure considered in the example above, b's reachability

[5] The other counterfactual state has an analogous interpretation.

[6] A more refined version of this statement is given in Section 5.

relation R'_b would link the counterfactual state $\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}$ of agent a to the original states, and will do so in a manner that guarantees that b's information at that state is exactly the same as the information that b has at the corresponding original state ω_1. The consequence is that when a is "more ignorant" at the counterfactual state $\lambda_{a,\omega_1}^{\{\omega_1,\omega_2\}}$, b believes, at that state, that a's information is unchanged relative to the original situation. It is formally shown in Section 4.5 that a is therefore "secretly" more ignorant at a counterfactual state. This shows that the construction of counterfactual structures guarantees that the counterfactual "more ignorance" scenario that is required by the Sure-Thing Principle applies to agents individually without distorting the beliefs of others.

4.2. The definition of counterfactual structures

Suppose $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$ is a partitional structure. For every agent $i \in N$, define $I_i(\omega) = \{\omega' \in \Omega | \omega R_i \omega'\}$. Trivially, for each $i \in N$, $\mathcal{I}_i = \{I_i(\omega)|\omega \in \Omega\}$ is a partition of the state space Ω, and $I_i(\omega)$ is the equivalence class of the state ω for agent i. Now define,

$$\Gamma_i = \{\cup_{e\in\mathcal{E}} e | \mathcal{E} \subseteq \mathcal{I}_i, \mathcal{E} \neq \emptyset\} \tag{6}$$

That is, Γ_i consists of all the partition elements of i, and of all the possible unions across those partition elements.[7] The set $\Gamma_i \setminus \mathcal{I}_i$ therefore consists of all unions of partition elements for agent i.

Definition 4. Given a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$, the corresponding *counterfactual structure* $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i\in N})$ has a state space $\Sigma = \Omega \cup \Lambda$ and reachability relations $R'_i \subseteq \Sigma \times \Omega$ such that,

1. Λ is a set of *counterfactual states* which contains a distinct element $\lambda_{i,\omega}^e$ for each agent $i \in N$, each event $e \in \Gamma_i \setminus \mathcal{I}_i$, and each state $\omega \in \Omega$.
2. For any $i \in N$, R'_i is identical to R_i over $\Omega \times \Omega$, and
   (i) every $\lambda_{i,\omega}^e \in \Lambda$ reaches exactly every state in e via R'_i.
   (ii) every $\lambda_{j,\omega}^e \in \Lambda$ (with $j \neq i$) reaches exactly every state in $I_i(\omega)$ via R'_i.

For notational purposes, ωs are used to denote original states in Ω, λs are used to denote counterfactual states in Λ, and σs are used to denote states belonging to $\Sigma = \Omega \cup \Lambda$.

[7] If $\rho(\mathcal{I}_i)$ denotes the sigma-algebra generated by the partition $\mathcal{I}_i$, then Γ_i can equivalently be defined as $\Gamma_i = \rho(\mathcal{I}_i) \setminus \emptyset$.

Definition 4 shows that the counterfactual structure $\mathcal{S}'$ of a partitional structure $\mathcal{S}$ consists of the original state space Ω and of the original reachability relations over $\Omega \times \Omega$, and adds to these a set of counterfactual states Λ such that each element $\lambda \in \Lambda$ reaches only states in Ω. In particular, starting from any state σ in Σ only states in Ω are reached via R'_i (for any $i \in N$). When it is restricted to the original states, R'_i is the original relation R_i.

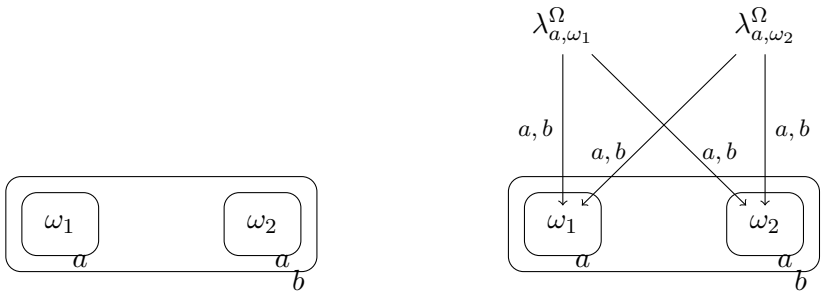
For any state $\sigma \in \Sigma$ in a counterfactual structure, the possibility set $b_i(\sigma)$ is the informational primitive in terms of which the operators $B_i(.)$ and ultimately $C_G(.)$ are defined. The partition elements $I_i(\omega)$ of the partition $\mathcal{I}_i$ are defined only over the original state space $\Omega \subseteq \Sigma$, and were introduced only for notational convenience. In fact, for any $\omega \in \Omega$, $b_i(\omega) = I_i(\omega)$.

As discussed in Section 4.4, $\lambda^e_{i,\omega}$ is interpreted as the state in which agent i is made "more ignorant" relative to ω by no longer distinguishing between the states in the partition elements the union of which constitutes the event e. As implied by point 2(ii) in Definition 4, the counterfactual state $\lambda^e_{j,\omega}$ of agent j (at which j is being made "more ignorant") reaches every state in $I_i(\omega)$ via the reachability relation R'_i of every agent $i \neq j$, which has the consequence of leaving agent i's information at the counterfactual state of j unchanged relative to i's information at the original state ω (Section 4.5).

The motivating example of Section 4.1 is revisited below in a more formal manner to illustrate Definition 4.

Example 1. Consider a partitional structure $\mathcal{S}$ with $\Omega = \{\omega_1, \omega_2\}$, $N = \{a, b\}$, and partitions $\mathcal{I}_a = \{\{\omega_1\}, \{\omega_2\}\}$ and $\mathcal{I}_b = \{\Omega\}$ which are represented as cells in Figure 1a. Note that $\Gamma_a = \{\{\omega_1\}, \{\omega_2\}, \Omega\}$ and $\Gamma_b = \{\Omega\}$. Since $\Gamma_b \setminus \mathcal{I}_b$ is empty, there are no counterfactual states to consider for agent b. On the other hand, $\Gamma_a \setminus \mathcal{I}_a = \{\Omega\}$ so the set of counterfactual states is given by $\Lambda = \{\lambda^\Omega_{a,\omega_1}, \lambda^\Omega_{a,\omega_2}\}$. For agent a each counterfactual state reaches every state in Ω, and for agent b each counterfactual state $\lambda^\Omega_{a,\omega}$ reaches every state in the corresponding partition element $I_b(\omega)$, which in this case corresponds to Ω. The complete counterfactual structure $\mathcal{S}'$ of $\mathcal{S}$ is illustrated in Figure 1b. (For each $i \in N$ the relations $R'_i \setminus R_i$ are represented by directed edges.)[8]

[8]Consider another example in which there is a partitional structure $\mathcal{S}$ with $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $N = \{a, b\}$, $\mathcal{I}_a = \{\{\omega_1\}, \{\omega_2, \omega_3\}\}$, and $\mathcal{I}_b = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$. In the counterfactual structure, there is a counterfactual state $\lambda^{\{\omega_2,\omega_3\}}_{b,\omega_2}$ in which b is made "more ignorant" relative to ω_2 that reaches both ω_2 and ω_3 via R'_b. However, there is no counterfactual state $\lambda^{\{\omega_2,\omega_3\}}_{b,\omega_1}$. That is, if the actual state were ω_1 it is irrelevant for the purposes of this paper to determine b's information when he no longer distinguishes between the states in the partition elements $I_b(\omega_2)$ and $I_b(\omega_3)$, neither of which contain the actual



(a) A partitional structure $\mathcal{S}$ (b) The counterfactual structure $\mathcal{S}'$ of $\mathcal{S}$

Figure 1: A partitional structure and its corresponding counterfactual structure

As discussed in Section 4.3, $B_i(.)$ is interpreted as a "belief" operator in counterfactual structures. In particular, $b_i(\sigma) \subseteq e$ is interpreted as "i believes that e at σ".[9] Several facts are worth noting: (1) a cannot distinguish between ω_1 and ω_2 at the counterfactual state $\lambda^{\Omega}_{a,\omega_1}$ (but a can distinguish between them at the original state ω_1). Therefore a is said to be "more ignorant" at $\lambda^{\Omega}_{a,\omega_1}$ relative to ω_1, in a sense to be made precise. (2) At $\lambda^{\Omega}_{a,\omega_1}$, b believes that a *can* distinguish between ω_1 and ω_2. In fact, b's information is unchanged relative to his information at the original state ω_1, so a was made "more ignorant" in a manner that is "secret". Furthermore note that the Truth property is violated. (3) At $\lambda^{\Omega}_{a,\omega_1}$, a does not believe that $\{\omega_1\}$, but she also does not believe that she does not believe that $\{\omega_1\}$, so Negative Introspection is also violated. ■

The following important remark shows the precise manner in which $b_i(.)$ and $I_i(.)$ are related in counterfactual structures.

Remark 1. Suppose that $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i\in N})$ is the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$. (a) For any state $\sigma \in \Sigma$,

state. It would also be unclear how to interpret such a state.

[9]The framework proposed here is purely semantic, so precisely which "basic facts" are true at the counterfactual states has not been specified. But for sake of argument, suppose that the fact p (e.g. "It is very sunny") is true at ω_1 but false at ω_2, while the fact q is true at both states. At the counterfactual state $\lambda^{\Omega}_{a,\omega_1}$, a is uncertain as to whether p is true but still believes that q is true. Whether a's belief that q is true at the counterfactual state is correct depends on whether q is in fact true at that state. For simplicity one could impose that a counterfactual state $\lambda^{e}_{i,\omega}$ is a "metaphysical replica" of the state ω in the sense that exactly the same basic facts are true at both states.

$b_i(\sigma) \in \Gamma_i$, and (b) for any $e \in \Gamma_i$, there is state $\sigma \in \Sigma$ such that $b_i(\sigma) = e$.

This essentially states that, in a counterfactual structure, Γ_i is *exactly* the set of all possible beliefs for agent i. As mentioned in the motivating example (Section 4.1) and as discussed at the end of Section 5, this fact is important in responding to the flaws raised by Moses and Nachum (1990).

4.3. Properties of the simple epistemic operator in counterfactual structures

In the epistemic logic literature, partitional structures (in which the reachability relations are equivalence relations) are known as "S5" structures, while structures in which the reachability relations are serial, transitive, and Euclidean, are known as "KD45" structures, and finally those in which the reachability relations are serial and transitive are "KD4" structures. The simple epistemic operator $B_i(.)$ satisfies Krikpe, Consistency, Positive and Negative Introspection, but not Truth in KD45 structures, and satisfies Krikpe, Consistency, and Positive Introspection, but not Negative Introspection or Truth in KD4 structures (see Fagin et al., 1995).

Remark 2. Suppose that $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i \in N})$ is the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$. The reachability relations $\{R'_i\}_{i \in N}$ are serial and transitive.

This remark shows that counterfactual structures belong to the class of KD4 information structures; and the properties of the epistemic operator are thus easily identified. Because Truth is dropped in counterfactual structures, the operators $B_i(.)$ and $C_G(.)$ must properly be interpreted as *belief* and *common belief* operators respectively.

The fact that $B_i(.)$ satisfies neither Truth nor Negative Introspection is consistent with the existing literature. For example, in his model of counterfactual reasoning in games, Stalnaker (1996) adopts a KD45 structure (in which Truth is violated) since by the very nature of counterfactual reasoning players must be capable of entertaining false beliefs. The fact that Negative Introspection is dropped is consistent with the insight of Samet (2010) that it precludes "more ignorance" from being meaningfully expressible in partitional structures since it implies that at any partition element, an agent must know something that the agent does not know at another.[10]

[10]To quote Samet (2010): "suppose the agent knows a fact f in state ω and does not know it in state ω'. Then, she *knows* in ω' that she does not know f, while in ω, she *does not know* that she does not know f". Note that Samet (1990) proves a probabilistic agreement theorem in information structures in which Negative Introspection is dropped (though not as a solution to the conceptual flaws that are dealt with here).

4.4. The interpretation of "ignorance" in counterfactual structures

Proposition 1. *Suppose that* $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i\in N})$ *is the counterfactual structure of a partitional structure* $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$. *For any* $e \subseteq \Omega$, *and* $\omega, \omega' \in \Omega$, $I_i(\omega) \subseteq e$ *and* $I_i(\omega') \subseteq e$ *if and only if* $b_i(\lambda_{i,\omega}^{I_i(\omega)\cup I_i(\omega')}) \subseteq e$.[11]

Proposition 1 shows that for any event e, i believes that e at the counterfactual state $\lambda_{i,\omega}^{I_i(\omega)\cup I_i(\omega')}$ if and only if i also believes that e at the states within *each* of the partition elements $I_i(\omega) \in \mathcal{I}_i$ and $I_i(\omega') \in \mathcal{I}_i$ (where $\omega, \omega' \in \Omega$). Informally, if one can call a belief in an event "information", then the information that i has at the counterfactual state preserves *exactly* the information that is the same across both the partition elements. In this sense, the information that i has at the counterfactual state is the information that i would have if i were "just more ignorant" than in the original situation: The agent is *unambiguously* more ignorant (with all states in $I_i(\omega) \cup I_i(\omega')$ being indistinguishable), but also not more ignorant than necessary relative to the original situation (in the sense that at the counterfactual state it is not the case that the agent loses information that the agent had at every state in $I_i(\omega) \cup I_i(\omega')$).[12]

Importantly, an agent does not *imagine* or *believe* him/herself to be "more ignorant" at a counterfactual state. An agent simply *is* more ignorant at a counterfactual state *tout-court*. In fact,

[11] The statement of this proposition is given for the union of two partition elements for ease of exposition. It is trivially generalizable to the union of multiple partition elements.

[12] Samet (2010) compares the knowledge of agents i and j at a given state ω by saying that j is at least as knowledgeable as i at ω if j knows every event that i knows there. In this paper, the beliefs of a single agent i are compared across states, and i is "just more ignorant" at the counterfactual state $\lambda_{i,\omega}^{e}$ relative to ω if at every state $\omega' \in e$, i believes every event that i believes at the counterfactual state.

Let us pause on the meaning of "unambiguously" here: Suppose another agent c is added to Example 1 with $\mathcal{I}_c = \{\{\omega_1\}, \{\omega_2\}\}$. In the original situation, at ω_1, a believes that $\{\omega_1\}$ and believes that c knows that $\{\omega_1\}$, and at ω_2, a believes that $\{\omega_2\}$ and does not believe that c believes that $\{\omega_1\}$. At the counterfactual state $\lambda_{a,\omega_1}^{\Omega}$, a can no longer distinguish between ω_1 and ω_2, but also no longer believes that c believes that $\{\omega_1\}$. It would appear as though too much information has been lost since one could conceive of a counterfactual in which a no longer distinguishes between ω_1 and ω_2 but still believes that c believes that $\{\omega_1\}$. However such a case would be problematic since a would believe something at the counterfactual state that she did not previously believe in the original situation (namely, that c believes that $\{\omega_1\}$), and so a would not *unambiguously* be more ignorant than in the original situation.

Remark 3. Agents in a counterfactual structure commonly exclude the counterfactual states.[13]

4.5. The interpretation of "secrecy" in counterfactual structures

In addition to showing that an agent i is "just" more ignorant at a counterfactual state $\lambda^e_{i,\omega}$, it is also possible to show that i is "secretly" more ignorant.

Remark 4. Suppose that $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i\in N})$ is the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$. For any event $h \subseteq \Omega$, any $\omega \in \Omega$, any $j \in N$ and any $\lambda^e_{i,\omega} \in \Lambda$ such that $i \neq j$, $b_j(\lambda^e_{i,\omega}) \subseteq h$ if and only if $I_j(\omega) \subseteq h$.[14]

This means that at a counterfactual state $\lambda^e_{i,\omega}$, i may have become "more ignorant" but the information of any other agent j is unchanged. The information at a counterfactual state therefore captures the fact that i is "secretly" more ignorant. (For example at the state $\lambda^\Omega_{a,\omega_1}$ in Figure 1b, agent a cannot distinguish between ω_1 and ω_2, but agent b believes that a can distinguish between them.)[15]

4.6. Further discussion about counterfactual structures

This section briefly relates counterfactual structures to existing models in epistemic logic and epistemic game theory.

The "impossible-worlds" approach (e.g. Wansing, 1990) augments information structures with a new set of states and with modified reachability relations. The states in the original structure are then referred to as "possible", or "normal", worlds, while the new states are referred to as "impossible", or "non-normal". Counterfactual structures can therefore be seen as

[13] For any state $\sigma \in \Sigma$ and any group of agents $G \subseteq N$, suppose $\sigma' \in T_G(\sigma)$. Since no counterfactual state reaches itself, and every counterfactual state must reach a state within Ω, it must be the case that $\sigma' \in \Omega$.

[14] Indeed, by the construction of the counterfactual structure, for any agent $j \neq i$ the state $\lambda^e_{i,\omega}$ only reaches the states in $I_j(\omega)$ via R'_j, therefore $b_j(\lambda^e_{i,\omega}) = I_j(\omega)$.

[15] One might wonder whether one could re-define counterfactual structures to include a counterfactual state for each agent i, each $e \in \Gamma_i \setminus \mathcal{I}_i$, *and each partition element* $I_i \in \mathcal{I}_i$, rather than for each state $\omega \in \Omega$. But doing so would no longer preserve secrecy: In the partitional structure of the example given in footnote 8, consider some counterfactual states in which a is made more ignorant; for example, $\lambda^\Omega_{a,\{\omega_1\}}$, and $\lambda^\Omega_{a,\{\omega_2,\omega_3\}}$, each of which reaches every state in Ω. Since there are only two such counterfactual states but three partition elements for agent b, there is no obvious way to re-wire these states to the original ones without altering b's information.

specific "impossible-worlds" structures. More recently, "action models" are being widely used in epistemic logic to formalize how the underlying structure (both the state space and the reachability relations) must be modified (by adding or deleting states and re-wiring) to model various protocols by which agents may gain some new information (Baltag and Moss, 2005). For example, Van Eijck (2008, Theorem 17) showed that in the case of secretly gaining new information, a partitional structure would have to be transformed into a KD45 structure. In this paper, the notion of becoming more ignorant transforms a partitional structure into a KD4 structure.

One approach to modeling counterfactuals in epistemic game theory proceeds in roughly the following manner (e.g. see Halpern, 1999): Define a "closeness" relation on states. A state ω is said to belong to the event "If f were the case, then e would be true" if e is true in all the closest states to ω where f is true. The framework of Halpern (1999) is quite general, and describes a framework for modeling counterfactuals for essentially any type of hypothetical situation.[16] This paper does not develop a fully fledged theory of counterfactuals but rather has the modest aim of finding a solution to the conceptual flaws raised by Moses and Nachum (1990) in a well specified framework – which requires defining counterfactuals for only very specific types of events (namely, being "more ignorant"). The approach developed here is well-suited for this purpose; nevertheless a formal relationship between this paper and the framework of Halpern (1999) is established in Appendix A.

5. The agreement theorem in counterfactual structures

Throughout this section, consider a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$, and its corresponding counterfactual structure $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i \in N})$. As before, define $I_i(\omega) = \{\omega' \in \Omega | \omega R_i \omega'\}$, the partition $\mathcal{I}_i = \{I_i(\omega) | \omega \in \Omega\}$, and the set Γ_i for every $i \in N$.

A decision function D_i for an agent $i \in N$ maps from Γ_i to a set of actions $\mathcal{A}$. That is,

$$D_i : \Gamma_i \to \mathcal{A} \tag{7}$$

An action function $\delta_i : \Sigma \to \mathcal{A}$ is now said to *follow* decision function D_i if

[16] Halpern (2001) similarly employs a "closeness" relation to resolve a debate between Aumann and Stalnaker on the implications of the common knowledge of substantive rationality in games of perfect information. Also see Artemov (2010) for a recent application of the framework.

for all states $\sigma \in \Sigma$, $\delta_i(\sigma) = D_i(b_i(\sigma))$. This is well-defined by the fact that for any $\sigma \in \Sigma$, $b_i(\sigma) \in \Gamma_i$ (see Remark 1(a)).

The new definitions for the Sure-Thing Principle and for like-mindedness are given below.

Definition 5. The decision function D_i of agent i satisfies the *Sure-Thing Principle* if for any non-empty subset $\mathcal{E}$ of $\mathcal{I}_i$, whenever for all $e \in \mathcal{E}$, $D_i(e) = x$ then $D_i(\cup_{e \in \mathcal{E}} e) = x$.

The domain Γ_i includes all possible unions of elements of the partition $\mathcal{I}_i$, so this is well-defined. Furthermore, note that $\mathcal{E}$ must be a set of disjoint events.[17] Suppose that over two partition elements $I_i(\omega)$ and $I_i(\omega')$, agent i's decision function is such that $D_i(I_i(\omega)) = D_i(I_i(\omega')) = x$. The Sure-Thing Principle would require that $D_i(I_i(\omega) \cup I_i(\omega')) = x$. But the informational content of the union $I_i(\omega) \cup I_i(\omega')$ is now well-defined within the information structure by the possibility set $b_i(\lambda_{i,\omega}^{I_i(\omega) \cup I_i(\omega')})$.

From the discussions in Sections 4.4 and 4.5, one can interpret this version of the Sure-Thing Principle as capturing the intuition that *If an agent i takes the same action in every case when i is better informed, i takes the same action in the case in which i is secretly "just" more ignorant.* Importantly, because the agents in a counterfactual structure commonly exclude the counterfactual states (Remark 3), the decision $D_i(I_i(\omega) \cup I_i(\omega')) = x$ must not be interpreted as i takes the action x when i *imagines* him/herself to be more ignorant than at an original state in $I_i(\omega) \cup I_i(\omega')$ (by no longer distinguishing between the states in this union). Rather, the decision $D_i(I_i(\omega) \cup I_i(\omega')) = x$ should be interpreted as i takes the action x when i *is* more ignorant than at an original state in $I_i(\omega) \cup I_i(\omega')$ (by no longer distinguishing between the states in this union). That is, if $D_i(I_i(\omega)) = D_i(I_i(\omega')) = x$ then the decision function D_i satisfies the Sure-Thing Principle if it happens to also

[17] Note that the Sure-Thing Principle is imposed only on events in $\mathcal{I}_i$, which *happen* to be disjoint because of the partitionality of the information structure. The condition is not imposed on *all* events and there is no *imposed* requirement that the events be disjoint. This contrasts with Moses and Nachum (1990) who, in their solution, propose adopting a version of the Sure-Thing Principle that is imposed on possibly non-disjoint events. The disjointness of events arises naturally if one thinks of decision functions as being conditional probabilities. Indeed, if one indexes a decision function by an event e and let $D_i^e(f) = \Pr_i(e|f)$, then such a decision function will satisfy the Sure-Thing Principle, since conditional probabilities satisfy $\Pr(e|f \cup f') = x$ if $\Pr(e|f) = \Pr(e|f') = x$ *when* $f \cap f' = \emptyset$ (see Bacharach, 1985, p. 180). In fact, Cave (1983) notes that conditional probabilities, expectations, and actions that maximize conditional expectations all naturally satisfy the Sure-Thing Principle.

satisfy $D_i(I_i(\omega) \cup I_i(\omega')) = x$ regardless of agent i's unawareness at an original state of their counterfactual ignorance. This is compelling: Suppose i decides to jump off the bridge when i knows that their coin landed face up and to jump off the bridge when i knows that their coin landed face down. It seems convincing to think that i must jump off the bridge when i does not know whether the coin landed face up or face down (i.e. when i *is* more ignorant), but it is relatively less convincing to think that i must jump off the bridge when i *imagines* him/herself to not know whether the coin landed face up or face down. Savage's (1972) principle is better illustrated by the first case than the second.

This version of the Sure-Thing Principle has some desirable features. Firstly, the manner in which ignorance is defined, as preserving only the information that is the same across partition elements, is intuitive. Secondly, the principle finds its origins in single-agent decision theory (see Savage, 1972), and the fact that the counterfactual situation is one in which the information of other agents is unchanged means that it is the *same* principle that is carried over from the single-agent setting to a multiple-agent setting.

Definition 6. Agents i and j are said to be *like-minded* if for any $e \in \Gamma_i$ and any $e' \in \Gamma_j$, if $e = e'$ then $D_i(e) = D_j(e')$.[18]

This notion of like-mindedness is straightforward: Over the same information, like-minded agents take the same action. However, this definition has an advantage over Bacharach's: Since the decision function D_i of agent i is now defined only over events in Γ_i, agent i is not required to consider which action to take over the possible belief of another agent j.

Theorem 1. *Let $\mathcal{S}' = (\Sigma, N, \{R'_i\}_{i \in N})$ be the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$.*
If the agents $i \in N$ are like-minded (Definition 6) and follow the decision functions $\{D_i\}_{i \in N}$ (as defined in (7)) that satisfy the Sure-Thing Principle (Definition 5), then for any $G \subseteq N$, if $C_G(\cap_{i \in G}\{\sigma \in \Sigma | \delta_i(\sigma) = x_i\}) \neq \emptyset$ then $x_i = x_j$ for all $i, j \in G$.

[18] In contrast with the previous definition, agents are not simply said to be like-minded if they have the "same" decision functions since the domains of the decision functions will now typically be different for different agents.

5.1. Discussion

Although Theorem 1 might appear to have similarities with that of Bacharach (1985), it is conceptually entirely distinct.[19] In particular, the theorem is obtained while avoiding the conceptual flaws that were discussed in Section 3: Remark 1 shows that in counterfactual structures, the domain of the decision function of every agent is *exactly* the set of all possible beliefs for that agent. Indeed, the decision functions are defined over unions of partition elements, but these are possible beliefs for the agents because, for every such union, there exists a counterfactual state at which the possibility set is precisely that union. The first point in the conceptual flaws raised by Moses and Nachum (1990, Lemma 3.2) is therefore avoided. Regarding the second point (Moses and Nachum, 1990, Lemma 3.3), the decision function D_i of agent i is now defined only over events in Γ_i. As mentioned above, there is therefore no requirement for the function to determine the agent's action in the case where the event corresponds to another agent's possible belief.[20]

6. Relation to alternative solutions

Moses and Nachum (1990) propose a solution to the conceptual flaws that they found in the result of Bacharach (1985). Essentially, they define a "relevance projection", which maps from sets of states to the "relevant information" at that set of states (Moses and Nachum, 1990, p. 158). They then impose conditions on this projection and on the decision functions to derive a new agreement theorem. However, it is not always obvious how a projection satisfying their conditions ought to be found. In contrast, the approach presented here offers a *constructive method* of obtaining a structure in which the analysis can be carried out.[21]

[19] Note that the proof of the theorem itself does not have to rely on the counterfactual structure. Indeed, with the appropriate restrictions, one could have stated the result as holding in standard partitional structures. However, it is the fact that the decision functions are embedded in the larger structure that allows for a proper interpretation of the information over which the decisions are defined.

[20] Bacharach (1985) required decisions to be defined over arbitrary events rather than over possible beliefs (see Definition 5) because he required decisions to be defined over unions of partition elements (and these cannot be possible beliefs in partitional structures). In contrast, Definition 7 requires decisions to be defined only over possible beliefs for the agents but these *can* be unions of partition elements in counterfactual structures.

[21] Note that the counterfactual structures along with the decision functions (as defined in (7)) do satisfy properties that resemble, *in spirit*, the conditions imposed on the relevance projection.

Aumann and Hart (2006) also propose a solution using a purely syntactic approach (which expresses information by means of a syntactic language comprising purely syntactic statements such as propositions). To derive their result, Aumann and Hart (2006) impose a condition, which is not imposed here, that higher-order information must be irrelevant to the agents' decisions. If first-order information refers to the information that agents have about the "basic facts", such as "It is raining", then second-order information refers to the information that an agent i has about an agent j's information about the "basic facts", and third-order information refers to the information that i has about j's information about k's information about the "basic facts", and so on. The restriction of Aumann and Hart (2006) requires agents' decision functions to not depend on anything above first-order information. But one can easily imagine scenarios in which higher-order information is relevant. Indeed, any situation in which an agent's decision depends on the information of another agent will suffice.[22]

Finally, Samet (2010) presented an ingenious solution to the conceptual flaws by redefining the Sure-Thing Principle entirely. Roughly, Samet's "Interpersonal Sure-Thing Principle" states that if agent i knows that agent j is more informed than i is, and knows that j's action is x, then i takes action x. Combining this with the assumption of the existence of an "epistemic dummy" – an agent who is less informed than every other agent – Samet (2010) proves a new agreement theorem in partitional structures. Informally, the proof proceeds as follows: Suppose i takes action x and j takes action y. There is an epistemic dummy k, so by the Interpersonal Sure-Thing Principle, k must take the same action as i and as j, but k cannot take two distinct actions, therefore $x = y$. Obviously, Theorem 1 does not require the existence of an epistemic dummy, and unlike the version of the Sure-Thing Principle presented here (Definition 5), the Interpersonal Sure-Thing Principle does not have a straightforward single-agent version.

[22] Consider a situation in which agent a is an analyst, and agent b requires some advice. Agent b is willing to pay a to obtain some advice if and only if b knows that a is more informed than b is. Here, b's decision does not depend on b's information regarding "basic facts", but on high-order knowledge; namely, on b knowing that a is more informed than b.

7. Appendix A: Relation to Halpern (1999)

For any state $\omega \in \Omega$, event $e \subseteq \Omega$, and agent i, Halpern (1999) defines a *selection function* $f_i(\omega, e)$, which maps into events, that identifies the closest states to ω in which the event e is true for agent i. The framework then admits conditional events with the introduction of a binary operator $>_i$ over events, which is defined as

$$h >_i e = \{\omega \in \Omega : f_i(\omega, h) \subseteq e\} \tag{8}$$

In this case, $\omega \in h >_i e$ if the closest states to ω in which h is true all satisfy e for agent i. For events $e, h \subseteq \Omega$ and each agent i, Halpern (1999) then defines a conditional epistemic operator $\bar{K}_i^h(e) = \neg K_i(\neg h) >_i K_i(e)$, where $K_i(e) = \{\omega \in \Omega : I_i(\omega) \subseteq e\}$ corresponds to the simple epistemic operator $B_i(.)$ when it is defined over the set of states Ω in a partitional or counterfactual information structure. The event $\bar{K}_i^h(e)$ is read as "if agent i considered h to be possible, then i would have known e".

Now, let $\mathcal{S}' = (\Sigma, N, \{R_i'\}_{i \in N})$ be the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i \in N})$. For any agent i, $h \in \Gamma_i \setminus \mathcal{I}_i$, and any event $e \subseteq \Omega$, define the operator $Z_i^h(e) = \{\omega \in \Omega : b_i(\lambda_{i,\omega}^h) \subseteq e\}$. So $Z_i^h(e)$ is the set of states in Ω for which there is a counterfactual state in which i believes e when i no longer distinguishes among the states in h.

Claim 1. Within $\mathcal{S}'$, suppose that for agent i, every $h \in \Gamma_i \setminus \mathcal{I}_i$ and every state $\omega \in h$, the selection function of Halpern (1999) satisfies

$$f_i(\omega, \neg K_i(\neg h)) = h \tag{9}$$

Then for any $i \in N$, $h \in \Gamma_i \setminus \mathcal{I}_i$, and for any $e \subseteq \Omega$, $\bar{K}_i^h(e) = Z_i^h(e)$.[23]

Proof. By defintion $Z_i^h(e) = \{\omega \in \Omega : b_i(\lambda_{i,\omega}^h) \subseteq e\}$ and $\bar{K}_i^h(e) = \{\omega \in \Omega : f_i(\omega, \neg K_i(\neg h)) \subseteq K_i(e)\}$. But for any $h \in \Gamma_i \setminus \mathcal{I}_i$ and any $\omega \in h$, $b_i(\lambda_{i,\omega}^h) = f_i(\omega, \neg K_i(\neg h)) = h$. Since h is a union of partition elements for i, $h \subseteq e$ if and only if $h \subseteq K_i(e)$. Therefore $\bar{K}_i^h(e) = Z_i^h(e)$. □

This claim shows that when the selection function satisfies Equation (9), the operator $Z_i^h(e)$ corresponds precisely to the operator $\bar{K}_i^h(e)$ of Halpern (1999) – at least, when agents condition over events that are unions of partition elements (that is, over events $h \in \Gamma_i \setminus \mathcal{I}_i$). Equation (9) does not

[23] The fact that the selection function equals h only for $\omega \in h$ reflects the reasoning outlined in the example given in footnote 8.

fully define the selection function since it does not assign value a $f_i(\omega, h)$ for *every possible* event h. But for the purposes of this paper it only matters to identify the information that agents would have for a very particular set of hypothetical events – namely those in which they are made "more ignorant" –, so other types of hypothetical event are ignored. Furthermore, the approach adopted in this paper allows for a detailed focus on the properties of hypothetical events in which agents are "more ignorant" such as "secrecy".

8. Appendix B: Proofs

Proof of Remark 2. Consider an arbitrary $i \in N$, and suppose $\sigma \in \Sigma$. If $\sigma \in \Omega$, then σ belongs to some equivalence class within $\mathcal{I}_i$. If $\sigma \in \Lambda$, then by construction of R'_i, there exists some $\omega \in \Omega$ such that $\sigma R'_i \omega$. In either case, there exists some $\omega \in \Sigma$ such that $\sigma R'_i \omega$. To establish transitivity, suppose $\sigma, \omega', \omega'' \in \Sigma$ such that $\sigma R'_i \omega'$ and $\omega' R'_i \omega''$. If $\sigma \in \Omega$, then σ, ω', and ω'' all belong to the same equivalence class, and therefore $\sigma R'_i \omega''$. If $\sigma \in \Lambda$, then since $\sigma R'_i \omega'$, it follows that $\omega' \in \Omega$, and since $\omega' R'_i \omega''$ it follows that ω'' and ω' belong to the same partition element $I_i(\omega')$. But by construction of the counterfactual states, since σ reaches ω', it must also reach every state that belongs to the equivalence class of ω'. It follows that $\sigma R'_i \omega''$. □

Proof of Remark 1. For any original state $\omega \in \Omega$, $b_i(\omega) = I_i(\omega)$ since the reachability relations R'_i and R_i are identical over $\Omega \times \Omega$. For any counterfactual state $\lambda^e_{i,\omega} \in \Lambda$, $b_i(\lambda^e_{i,\omega})$ is a union of partition elements in $\mathcal{I}_i$ since any event $e \in \Gamma_i \setminus \mathcal{I}_i$ is a union of partition elements and $\lambda^e_{i,\omega}$ reaches exactly all states in e via R'_i. And, for any $\lambda^e_{j,\omega} \in \Lambda$ (with $j \neq i$), $b_i(\lambda^e_{j,\omega}) = I_i(\omega)$ since $\lambda^e_{j,\omega}$ reaches exactly every state in $I_i(\omega)$ via R'_i. All the above implies that (a) for any state $\sigma \in \Sigma$, $b_i(\sigma) \in \Gamma_i$. Conversely, consider any $e \in \Gamma_i$. If e is a partition element in $\mathcal{I}_i$ then trivially, there is a state $\omega \in \Omega$ such that $e = I_i(\omega) = b_i(\omega)$. On the other hand, suppose e is a union of partition elements (so $e \in \Gamma_i \setminus \mathcal{I}_i$). Then there is a counterfactual state $\lambda^e_{i,\omega}$ that reaches exactly every state in e via R'_i, so $b_i(\lambda^e_{i,\omega}) = e$. Therefore, (b) for any $e \in \Gamma_i$, there exists a state $\sigma \in \Sigma$ such that $b_i(\sigma) = e$. □

Proof of Proposition 1. By definition of Γ_i, $I_i(\omega), I_i(\omega') \in \Gamma_i$ and $I_i(\omega) \cup I_i(\omega') \in \Gamma_i$. By construction of the counterfactual structure, the state $\lambda^{I_i(\omega) \cup I_i(\omega')}_{i,\omega}$ reaches exactly every state in $I_i(\omega) \cup I_i(\omega')$ via R'_i. Therefore, $b_i(\lambda^{I_i(\omega) \cup I_i(\omega')}_{i,\omega}) = I_i(\omega) \cup I_i(\omega')$. Now suppose $I_i(\omega) \subseteq e$ and $I_i(\omega') \subseteq e$. So,

$I_i(\omega) \cup I_i(\omega') \subseteq e$. Then clearly, $b_i(\lambda_{i,\omega}^{I_i(\omega)\cup I_i(\omega')}) \subseteq e$. For the converse, suppose $b_i(\lambda_{i,\omega}^{I_i(\omega)\cup I_i(\omega')}) \subseteq e$. Then $I_i(\omega) \cup I_i(\omega') \subseteq e$. It follows that $I_i(\omega) \subseteq e$ and $I_i(\omega') \subseteq e$. □

Lemma 1. *Suppose that $\mathcal{S}' = (\Sigma, N, \{R_i'\}_{i\in N})$ is the counterfactual structure of a partitional structure $\mathcal{S} = (\Omega, N, \{R_i\}_{i\in N})$. For any $\sigma \in \Sigma$, $G \subseteq N$, and $i \in G$, $\cup_{\omega\in T_G(\sigma)} b_i(\omega) = T_G(\sigma)$.*

Proof of Lemma 1. Let $\sigma \in \Sigma$. Suppose that $\omega' \in \cup_{\omega\in T_G(\sigma)} b_i(\omega)$. Then, $\omega' \in b_i(\omega)$ for some $\omega \in T_G(\sigma)$. So $\omega R_i \omega'$, and by definition of $T_G(.)$, ω is reachable from σ. It follows that ω' is reachable from σ, so $\omega' \in T_G(\sigma)$. For the converse, suppose $\omega' \in T_G(\sigma)$. Since no counterfactual state reaches itself, and every counterfactual state must reach a state within Ω, it must be the case that $\omega' \in \Omega$. Since $b_i(.)$ and $I_i(.)$ coincide over Ω, it follows that $\omega' \in I_i(\omega') = b_i(\omega')$. So, for some $\omega'' \in T_G(\sigma)$, $\omega' \in b_i(\omega'')$. That is, $\omega' \in \cup_{\omega\in T_G(\sigma)} b_i(\omega)$. □

Proof of Theorem 1. Suppose that $\omega \in C_G(\cap_{i\in G}\{\sigma \in \Sigma | \delta_i(\sigma) = x_i\})$. Then, for every $i \in G$, $T_G(\omega) \subseteq \{\sigma \in \Sigma | \delta_i(\sigma) = x_i\}$. Let us focus on agent i. This means that $\delta_i(\sigma) = x_i$ for every $\sigma \in T_G(\omega)$. By Lemma 1, $\cup_{\omega'\in T_G(\omega)} b_i(\omega') = T_G(\omega)$. This implies that $T_G(\omega)$ is a (non-empty) set of disjoint possibility sets $b_i(\omega')$ such that $\omega' \in T_G(\omega)$. This implies that $D_i(b_i(\omega')) = x_i$ for every possibility set $b_i(\omega')$ that is a subset of $T_G(\omega)$. Note that for any $\omega' \in T_G(\omega)$, since no counterfactual state reaches itself, and every counterfactual state must reach a state within Ω, it must be the case that $\omega' \in \Omega$. Also since $b_i(.)$ and $I_i(.)$ coincide over Ω, for any $\omega' \in \Omega$, $b_i(\omega') = I_i(\omega')$. From this, it follows that $\{b_i(\omega') | \omega' \in T_G(\omega)\} \subseteq \mathcal{I}_i$, and by the Sure-Thing Principle, it follows that $D_i(T_G(\omega)) = x_i$. A similar argument for any other agent j leads to the conclusion that $D_j(T_G(\omega)) = x_j$. But since any agents $i, j \in G$ are like-minded, it follows that $x_i = D_i(T_G(\omega)) = D_j(T_G(\omega)) = x_j$ for all $i, j \in G$. □

9. Acknowledgements

I thank Francis Dennig, Peter Eső, Alan Kirman, Harvey Lederman, Yoram Moses, John Quah, Dov Samet, and Alex Teytelboym for valuable discussions. The paper also benefited from the comments of Ehud Lehrer, Burkhard Schipper, anonymous referees at *Mathematical Social Sciences*, and participants at *TARK*.

10. References

Artemov, S., 2010. Robust knowledge and rationality. Tech. rep., Technical Report, CUNY, Computer Science.

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

Aumann, R., 1999. Interactive epistemology (i): Knowledge. International Journal of Game Theory 28 (3), 263–300.

Aumann, R., Hart, S., 2006. Agreeing on decisions. Unpublished manuscript, The Einstein Institute of Mathematics, Jerusalem, Israel, http://math.huji.ac.il/~hart/papers/agree.pdf (Last accessed: 17/04/2015).

Bach, C. W., Cabessa, J., 2011. Agreeing to disagree with limit knowledge. In: Logic, Rationality, and Interaction. Springer, pp. 51–60.

Bach, C. W., Cabessa, J., 2016. Limit-agreeing to disagree. Journal of Logic and Computation.

Bach, C. W., Perea, A., 2013. Agreeing to disagree with lexicographic prior beliefs. Mathematical Social Sciences 66 (2), 129–133.

Bacharach, M., 1985. Some extensions of a claim of aumann in an axiomatic model of knowledge. Journal of Economic Theory 37 (1), 167–190.

Baltag, A., Moss, L., 2005. Logics for epistemic programs. Information, Interaction and Agency 139 (2), 1–60.

Bonanno, G., Nehring, K., 1997. Agreeing to disagree: A survey. Working paper series no. 97-18, Department of Economics, University of California, Davis.

Cave, J., 1983. Learning to agree. Economics Letters 12 (2), 147–152.

Chellas, B., 1980. Modal logic: an introduction. Cambridge University Press, Cambridge, UK.

Dégremont, C., Roy, O., 2012. Agreement theorems in dynamic-epistemic logic. Journal of philosophical logic 41 (4), 735–764.

Demey, L., 2014. Agreeing to disagree in probabilistic dynamic epistemic logic. Synthese 191 (3), 409–438.

Dominiak, A., Lefort, J.-P., 2013. Agreement theorem for neo-additive beliefs. Economic Theory 52 (1), 1–13.

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

Halpern, J., 1999. Hypothetical knowledge and counterfactual reasoning. International Journal of Game Theory 28 (3), 315–330.

Halpern, J. Y., 2001. Substantive rationality and backward induction. Games and Economic Behavior 37 (2), 425–435.

Heifetz, A., Meier, M., Schipper, B. C., 2013. Unawareness, beliefs, and speculative trade. Games and Economic Behavior 77 (1), 100–121.

Hellman, Z., 2013. Deludedly agreeing to agree. In: Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge. pp. 105–110.

Ménager, L., 2007. Agreeing to disagree: a review. Working paper, LEM, Université Paris II, see https://sites.google.com/site/luciemenager/research (Last accessed: 17/04/2015).

Moses, Y., Nachum, G., 1990. Agreeing to disagree after all. In: Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge. pp. 151–168.

Samet, D., 1990. Ignoring ignorance and agreeing to disagree. Journal of Economic Theory 52 (1), 190–207.

Samet, D., 2010. Agreeing to disagree: The non-probabilistic case. Games and Economic Behavior 69 (1), 169–174.

Savage, L., 1972. The Foundations of Statistics. Dover Publications, Mineola, NY.

Stalnaker, R., 1996. Knowledge, belief and counterfactual reasoning in games. Economics and Philosophy 12, 133–164.

van Benthem, J., 2010. Modal Logic for Open Minds. University of Chicago Press, Chicago, IL.

Van Eijck, J., 2008. Advances in dynamic epistemic logic. Unpublished manuscript, CWI and ILLC, Amsterdam, Netherlands, http://

homepages.cwi.nl/~jve/papers/08/ae/38-anininlijc.pdf (Last accessed: 17/04/2015).

Wansing, H., 1990. A general possible worlds framework for reasoning about knowledge and belief. Studia Logica 49 (4), 523–539.