Published in: *Mathematical Social Sciences*, 35, 1998, pp. 17-36.

ON THE LOGIC AND ROLE OF NEGATIVE INTROSPECTION OF COMMON BELIEF

Giacomo Bonanno

and

Klaus Nehring [*]

Department of Economics,
University of California,
Davis, CA 95616-8578.

E-mail: gfbonanno@ucdavis.edu
kdnehring@ucdavis.edu

October 1996. Revised, June 1997

Abstract

We provide two local characterizations of Negative Introspection of common belief (NI). The first shows NI be equivalent to the conjunction of Truth of individuals' belief about what is commonly believed (TCB) and common belief in TCB. According to the second, NI corresponds to a general reducibility property of higher-order beliefs about common belief to either common belief or its negation.

Negative Introspection of common belief and its characterizing conditions help provide epistemic foundations for game-theoretic solution concepts. We show this for correlated equilibrium under incomplete information as well as backward induction in a class of extensive games.

[*] The authors are grateful to an anonymous referee for helpful suggestions.

1. Introduction

The concepts of common knowledge and common belief have been discussed extensively, both syntactically and semantically.[1] In the economics literature the most common approach to modeling the beliefs (or knowledge) of an individual *i* is through either a *belief operator* $B_i : 2^{\Omega} \to 2^{\Omega}$ (where Ω is a set of states and 2^{Ω} is the set of subsets of Ω, called events), or through a *possibility correspondence* $I_i: \Omega \to 2^{\Omega}$.[2] A typical interpretation of the operator B_i is in terms of "probability 1 beliefs" (see, for example, Aumann and Brandenburger, 1995, and Dekel and Gul, 1997). At the *individual* level the difference between knowledge and belief is usually identified with the presence or absence of the *Truth Axiom*

$$\text{T.} \qquad B_i E \subseteq E,$$

which is interpreted as "if individual *i* believes E, then E is true". In such a case the individual is said to *know* E. Thus it is possible for an individual to *believe* a false proposition but she cannot *know* a false proposition.

At the interpersonal level, the literature then distinguishes between *common* knowledge

---

[1] See, for example, Aumann (1976), Aumann and Brandenburger (1995), Bacharach (1985), Bonanno (1996), Brandenburger and Dekel (1987), Dekel and Gul (1997), Fagin et al (1995), Geanakoplos (1992), Heifetz (1996), Kaneko and Nagashima (1991, 1997), Lewis (1969), Lismont and Mongin (1994, 1995), Milgrom (1981), Rubinstein and Wolinski (1990), Samet (1990).

[2] Given an event $E \subseteq \Omega$, $B_i E$ is the event that individual *i* believes E. Alternatively, given a state $\omega \in \Omega$, $I_i(\omega)$ is the set of states that individual *i* considers possible at ω. The two approaches are equivalent, in the following sense. Given a possibility correspondence I_i, one can define a corresponding belief operator as follows: for every $E \subseteq \Omega$, $B_i E = \{\omega \in \Omega : I_i(\omega) \subseteq E\}$. Conversely, given a normal belief operator B_i (as defined below) one obtains a possibility correspondence I_i as follows (the symbol $\neg$ denotes complement): for every $\omega \in \Omega$, $I_i(\omega) = \{\omega' \in \Omega : \omega \in \neg B_i \neg \{\omega'\}\}$. It is easily verified that these two maps are one the inverse of the other. By a *normal* belief operator we mean an operator B_i that satisfies Necessity ($B_i \Omega = \Omega$), Conjunction [$B_i E \cap B_i F = B_i(E \cap F)$] and Monotonicity (if $E \subseteq F$ then $B_i E \subseteq B_i F$) (see Section 2; cf. also Chellas, 1984, p.115). The belief operator obtained from a possibility correspondence is always normal.

and common belief on the basis of whether or not the Truth Axiom is postulated for the individuals. However, while at the individual level T captures merely a relationship between the individual's beliefs and the external world, at the interpersonal level it has strong implications, such as the following: $B_iB_jE \subseteq B_iE$, that is, if individual *i* believes that individual *j* believes E, then individual *i* herself believes E[3]. Thus, in contrast to other axioms, T does not merely reflect individual agents' "logic of belief". Given its logical force, it is not surprising to find that the Truth Axiom has strong implications for the logic of common belief. In particular, it is well known that, if each individual's beliefs satisfy the strongest logic of knowledge (namely S5 or KT5),[4] the associated common knowledge operator satisfies this logic too. Such is not the case for belief: bereft of the Truth Axiom, even the strongest logic for individual belief (KD45) is insufficient to ensure the satisfaction of the "Negative Introspection" axiom for common belief (B_* denotes the common belief operator, which is defined in Section 2):

$$5^*. \qquad \neg B_*E \subseteq B_*\neg B_*E,$$

which says that if E is not common belief, then it is common belief that E is not common belief.

In this paper we investigate the implications of Negative Introspection as a property of common belief. There are two sources of interest in such an analysis, epistemic as well as behavioral. From the epistemic point of view, an adequate understanding of common belief presupposes an understanding of when and how one of the major logical properties of belief operators, Negative Introspection, fails. We show that the possibility of such failure is closely related to the somewhat counterintuitive possibility that individuals may be mistaken in their beliefs about what is commonly believed (Theorem 1, Section 2). Epistemically, Negative

---

[3] The Truth Axiom applied to individual *j* gives $B_jE \subseteq E$. It follows from Monotonicity of B_i (see Footnote 2) that $B_iB_jE \subseteq B_iE$. The intersubjective implications of the Truth Axiom are explored in Bonanno and Nehring (1997).

[4] This corresponds to the following assumptions on the possibility correspondences: $\forall\, \alpha, \beta \in \Omega$ (1) $\alpha \in I_i(\alpha)$ and (2) if $\beta \in I_i(\alpha)$ then $I_i(\alpha) = I_i(\beta)$. That is, each possibility correspondence gives rise to a partition of the set Ω.

Introspection of common belief turns also out to be useful due to its being *equivalent* to the reducibility of complex *mixed* iterations of belief operators such as $B_1\neg B_2 B_* \neg B_1 B_* E$ to either $B_* E$ or $\neg B_* E$ (Theorem 2, Section 3; for instance, $B_1\neg B_2 B_* \neg B_1 B_* E$ is equal to $B_* E$).

From the point of view of behavioral/game-theoretic applications one would expect that Negative Introspection makes a significant difference in contexts in which properties involving the *negation* of common belief play are important. Paradigmatic examples are agreement scenarios and no trade in betting environments. A central issue that has emerged in the recent literature on the epistemic foundations of game theory concerns the justification and indeed meaningfulness of the Common Prior Assumption under incomplete information (see in particular Dekel and Gul, 1997, Gul, 1996, and Lipman, 1995) which is required for an epistemic grounding of correlated equilibrium as a general solution concept (Aumann, 1987). In response to this skepticism, it has recently been shown that the Common Prior Assumption has a sound foundation in a general Agreement property (absence of agreeing to disagree à la Aumann, 1976). However, the local version of the Common Prior Assumption characterized by Agreement is by itself too weak a property to entail the play of correlated equilibrium strategies. What is needed to obtain the desired implication is exactly Negative Introspection of common belief, as shown in Section 4 (Theorem 4).

While these examples show that satisfaction of Negative Introspection of common belief can make a significant difference, it is highly problematic as a *primitive* assumption, as it is neither formulated as, nor straightforwardly translatable into, a property of individual beliefs. The main result of this paper explicates Negative Introspection of common belief in terms of the property of correctness of individual beliefs about common belief, **TCB**. We provide the following *local* characterization. Let **NI** be the set of states where the Negative Introspection property for the common belief operator is satisfied: $\alpha \in \mathbf{NI}$ if and only if, for every event E, if $\alpha \in \neg B_* E$ then $\alpha \in B_* \neg B_* E$. Let **TCB** be the following event: $\alpha \in \mathbf{TCB}$ if and only if, for every individual *i* and every event E, if $\alpha \in B_i B_* E$ then $\alpha \in B_* E$. Theorem 1 (Section 2) states that **NI = TCB** $\cap$ B_***TCB.**

A further application of Negative Introspection of common belief is given in Section 4 where we provide a decomposition of common belief in no error, $B_*\mathbf{T}$ (where **T** is the event that no individual has any false beliefs). Epistemically common belief in no error is of interest as a local and intersubjective version of the Truth Axiom. Behaviorally it has recently been used to justify backward induction in an interesting class of perfect information games (Stalnaker, 1996, Stuart, 1997). One part of the decomposition is an epistemic condition which is behaviorally equivalent to the absence of *unbounded* gains from betting. From an economic point of view this seems to be a highly acceptable assumption. Thus the burden of the justification for common belief in no error falls on the second part of the decomposition which $B_*\mathbf{NI}$ respectively $B_*\mathbf{TCB}$, which, by itself, is shown to be substantially weaker than common belief in no error.

2. Intersubjective Characterization of Negative Introspection of Common Belief

Throughout, individual beliefs are assumed to satisfy the logic of KD45 (or weak S5)[5].

DEFINITION 1. A *qualitative interactive belief frame* (or frame, for short) is a tuple

$$\mathcal{Q} = \langle \mathrm{N},\ \Omega,\ \tau, \{\mathrm{I_i}\}_{\mathrm{i}\in\mathrm{N}} \rangle$$

where

- $\mathrm{N} = \{1, ..., n\}$ is a finite set of *individuals*.
- Ω is a (possibly infinite) set of *states* (or possible worlds). The subsets of Ω are called *events.*
- $\tau \in \Omega$ is the "true" or "actual" state[6].

---

[5] In particular, this will be the case in a Bayesian setting: see, for example, Aumann and Brandenburger (1995) and Dekel and Gul (1997). See also Section 4.

[6] We have included the true state in the definition of an interactive belief frame in order to stress the interpretation of the frame as a representation of a particular profile of hierarchies of beliefs.

- for every individual $i \in N$, $I_i : \Omega \to 2^{\Omega} \setminus \varnothing$ is i's *possibility correspondence*, satisfying the following properties (whose interpretation is given in Remark 1 below): $\forall \alpha, \beta \in \Omega$,

  Transitivity: if $\beta \in I_i(\alpha)$ then $I_i(\beta) \subseteq I_i(\alpha)$,

  Euclideanness: if $\beta \in I_i(\alpha)$ then $I_i(\alpha) \subseteq I_i(\beta)$.

For every $\alpha \in \Omega$, $I_i(\alpha)$ is the set of states that individual *i* considers possible at α.

Given a frame and an individual *i*, i's *belief operator* $B_i : 2^{\Omega} \to 2^{\Omega}$ is defined as follows: $\forall E \subseteq \Omega$, $B_i E = \{\omega \in \Omega : I_i(\omega) \subseteq E\}$. $B_i E$ can be interpreted as the event that (i.e. the set of states at which) individual *i believes* that event E has occurred.

REMARK 1. It is well known (see Chellas, 1984, p. 164) that non-empty-valuedness of the possibility correspondence is equivalent to *consistency* of beliefs: $\forall E \subseteq \Omega$, $B_i E \subseteq \neg B_i \neg E$ (the individual cannot simultaneously believe E and not E). Transitivity of the possibility correspondence is equivalent to *positive introspection* of beliefs: $\forall E \subseteq \Omega$, $B_i E \subseteq B_i B_i E$ (if the individual believes E then she believes that she believes E). Finally, euclideanness of the possibility correspondence is equivalent to *negative introspection* of beliefs: $\forall E \subseteq \Omega$, $\neg B_i E \subseteq B_i \neg B_i E$ (if the individual does not believe E, then she believes that she does not believe E).

Notice that we have allowed for false beliefs by not assuming reflexivity of the possibility correspondences ($\forall \alpha \in \Omega$, $\alpha \in I_i(\alpha)$), which – as is well known – is equivalent to the *Truth Axiom*: $\forall E \subseteq \Omega$, $B_i E \subseteq E$ (if the individual believes E then E is indeed true).

REMARK 2. The following fact is also well known. Let $I : \Omega \to 2^{\Omega}$ be a possibility correspondence and $B : 2^{\Omega} \to 2^{\Omega}$ the corresponding belief operator (that is, $\forall E \subseteq \Omega$, $BE = \{\omega \in \Omega : I(\omega) \subseteq E\}$). Then B satisfies the following properties: $\forall E,F \subseteq \Omega$,

Necessity: $B\Omega = \Omega$

Conjunction: $B(E \cap F) = BE \cap BF$

Monotonicity: if $E \subseteq F$ then $BE \subseteq BF$.

The common belief operator B_* is defined as follows. First, for every $E \subseteq \Omega$, let $B_eE = \bigcap_{i \in N} B_iE$, that is, B_eE is the event that everybody believes E. Then, for every $E \subseteq \Omega$, the event that E is commonly believed is defined as the infinite intersection:

$$B_*E = B_eE \cap B_eB_eE \cap B_eB_eB_eE \cap \ldots$$

The corresponding possibility correspondence I_* is then defined as follows: for every $\alpha \in \Omega$, $I_*(\alpha) = \{\omega \in \Omega : \alpha \in \neg B_* \neg \{\omega\}\}$. It is well known[7] that I_* can be characterized as the *transitive closure* of $\bigcup_{i \in N} I_i$, that is,:

> $\forall\, \alpha, \beta \in \Omega$, $\beta \in I_*(\alpha)$ if and only if there is a sequence $\langle i_1, \ldots i_m \rangle$ in N and a sequence $\langle \eta_0, \eta_1, \ldots, \eta_m \rangle$ in Ω such that: (i) $\eta_0 = \alpha$, (ii) $\eta_m = \beta$ and (iii) for every $k = 0, \ldots, m-1$, $\eta_{k+1} \in I_{i_{k+1}}(\eta_k)$.

Note that, although I_* is always non-empty-valued and transitive, it need not be euclidean, as the following example shows (cf. Colombetti, 1993 and Lismont and Mongin, 1994, 1995). Let $N = \{1, 2\}$, $\Omega = \{\tau, \beta\}$, $I_1(\tau) = \{\tau\}$, $I_1(\beta) = \{\beta\}$, $I_2(\tau) = I_2(\beta) = \{\beta\}$. Hence $I_*(\beta) = \{\beta\}$ and $I_*(\tau) = \{\tau, \beta\}$. Note that I_1 and I_2 are non-empty-valued, transitive and euclidean, while I_* is not euclidean: $\beta \in I_*(\tau)$ but $I_*(\tau) \not\subseteq I_*(\beta)$. It follows (cf. Remark 1) that B_* does not satisfy Negative Introspection. In fact, let $E = \{\beta\}$. Then $B_*E = \{\beta\}$ and $B_*\neg B_*E = B_*\{\tau\} = \varnothing$. Thus $\neg B_*E = \{\tau\} \not\subseteq B_*\neg B_*E = \varnothing$.

The game-theoretic relevance of Negative Introspection of common belief will be

---

[7] See, for example, Bonanno (1996), Fagin et al (1995), Lismont and Mongin (1994, 1995).

discussed in Section 4. *In this section we address the issue of how to understand it in terms of conditions on individual beliefs*. Theorem 1 provides a local characterization, further illuminated by Proposition 1, while Corollary 1 draws the "global" implications.

Properties such as Negative Introspection of common belief are to be defined locally, i.e. with respect to the true state τ. An equivalent, and mathematically more elegant, alternative is to define a property as an event, i.e. a set of states; the property is then satisfied at the true state τ if and only if τ belongs to that event. A characterization result will correspondingly be stated as the equality of two events.

Let (NI stands for "Negative Introspection")

$$\mathbf{NI} = \bigcap_{E \in 2^{\Omega}} (B_*E \cup B_*\neg B_*E).$$

Thus $\alpha \in \mathbf{NI}$ if and only if – for every event E – whenever at α it is not common belief that E, then, at α, it is common belief that E is not commonly believed (if $\alpha \in \neg B_*E$ then $\alpha \in B_*\neg B_*E$).

REMARK 3. It is well known that $\alpha \in \mathbf{NI}$ if and only if, $\forall \beta, \gamma \in I_*(\alpha)$, $\gamma \in I_*(\beta)$.[8]

Let **TCB** ('TCB' stands for "Truth about Common Belief") be the following event:

$$\mathbf{TCB} = \bigcap_{i \in N} \bigcap_{E \in 2^{\Omega}} (\neg B_i B_*E \cup B_*E).$$

TCB captures the notion that individuals are correct in their beliefs about what is commonly believed: $\alpha \in \mathbf{TCB}$ if and only if – for every event E and for every individual i – if, at α, individual i believes that E is commonly believed, then, at α, E is indeed commonly believed (if

---

[8] *Proof.* (i) Suppose that $\beta, \gamma \in I_*(\alpha)$ and $\gamma \notin I_*(\beta)$. Let $E = I_*(\beta)$. Since $\gamma \in I_*(\alpha) \cap \neg E$, $I_*(\alpha) \not\subseteq E$, that is, $\alpha \in \neg B_*E$. Since $I_*(\beta) = E$, $\beta \in B_*E$. Hence, since $\beta \in I_*(\alpha)$, $I_*(\alpha) \cap B_*E \neq \varnothing$, that is, $\alpha \in \neg B_*\neg B_*E$. Thus $\alpha \in \neg B_*E \cap \neg B_*\neg B_*E$. Hence $\alpha \notin \mathbf{NI}$. (ii) Conversely, suppose that $\alpha \notin \mathbf{NI}$. Then there exists an $E \subseteq \Omega$ such that $\alpha \in \neg B_*E \cap \neg B_*\neg B_*E$. Since $\alpha \in \neg B_*\neg B_*E$, there

$\alpha \in B_i B_* E$ then $\alpha \in B_* E$). Note that it follows from the definition of common belief that every individual must be correct in her belief that something is *not* common belief: $B_i \neg B_* E \subseteq \neg B_* E$[9]. Thus $\alpha \in$ **TCB** if and only if at α individual beliefs *about* common beliefs are correct.

THEOREM 1. NI = TCB $\cap$ B_*TCB.

Proof. (**NI** $\subseteq$ **TCB**). Let $\alpha \in$ **NI**. Fix an arbitrary $i \in N$ and $E \subseteq \Omega$. We want to show that $\alpha \in B_* E \cup \neg B_i B_* E$. Since $\alpha \in$ **NI**, $\alpha \in B_* E \cup B_* \neg B_* E$. Suppose that $\alpha \notin B_* E$. Then $\alpha \in B_* \neg B_* E$. By definition of B_*, $B_* \neg B_* E \subseteq B_i \neg B_* E$. By Consistency (cf. Remark 1), $B_i \neg B_* E \subseteq \neg B_i B_* E$. Thus $B_* \neg B_* E \subseteq \neg B_i B_* E$. Hence $\alpha \in \neg B_i B_* E$.

In order to prove that **NI** $\subseteq B_*$**TCB**, we need the following lemma.

LEMMA 1. **NI** $\subseteq B_*$**NI.**

Proof of Lemma 1. Let $\alpha \in$ **NI**. Fix an arbitrary $\beta \in I_*(\alpha)$. We need to show that $\beta \in$ **NI**, that is (cf. Remark 3), for all $\gamma, \delta \in I_*(\beta)$, $\delta \in I_*(\gamma)$. Since $\beta \in I_*(\alpha)$, by transitivity of I_*, $I_*(\beta) \subseteq I_*(\alpha)$ and, by Remark 3, since $\alpha \in$ **NI**, $I_*(\alpha) \subseteq I_*(\beta)$. Hence $I_*(\beta) = I_*(\alpha)$. Fix arbitrary $\gamma, \delta \in I_*(\beta)$. Since $I_*(\beta) = I_*(\alpha)$, $\gamma, \delta \in I_*(\alpha)$. Since $\gamma \in I_*(\alpha)$ and $\alpha \in$ **NI**, by Remark 3 and transitivity of I_*, $I_*(\gamma) = I_*(\alpha)$. Hence $\delta \in I_*(\gamma)$. ■

Proof of Theorem 1 continued (**NI** $\subseteq B_*$**TCB**). Since **NI** $\subseteq$ **TCB** (proved above), by Monotonicity of B_* (cf. Remark 2), B_***NI** $\subseteq B_*$**TCB**. By Lemma 1, **NI** $\subseteq B_*$**NI.** Hence **NI** $\subseteq B_*$**TCB**.

In order to prove the converse, namely that **TCB** $\cap B_*$**TCB** $\subseteq$ **NI**, we need the following lemma.

---

exists a $\beta \in I_*(\alpha)$ such that $\beta \in B_* E$, that is, $I_*(\beta) \subseteq E$. Since $\alpha \notin B_* E$, there exists a $\gamma \in I_*(\alpha)$ such that $\gamma \notin E$. Hence $\gamma \notin I_*(\beta)$.

[9] By definition of B_*, $B_* E \subseteq B_i B_* E$. By consistency of i's beliefs (cf. Remark 1), $B_i B_* E \subseteq \neg B_i \neg B_* E$. Thus $B_* E \subseteq \neg B_i \neg B_* E$ which is equivalent to $B_i \neg B_* E \subseteq \neg B_* E$.

LEMMA 2. $\alpha \in \mathbf{TCB}$ if and only if $I_*(\alpha)$ satisfies the following property:

P_{TCB} $\quad \forall i \in N,\ \forall \beta \in I_*(\alpha),\ \exists \gamma \in I_i(\alpha)$ such that $\beta \in I_*(\gamma)$.

Proof of Lemma 2. (Necessity). Let $\alpha \in \Omega$ be such that $I_*(\alpha)$ satisfies property P_{TCB}. We want to show that $\alpha \in \mathbf{TCB}$. Fix arbitrary $i \in N$ and $E \subseteq \Omega$. We need to show that $\alpha \in \neg B_i B_* E \cup B_* E$, that is, either $I_i(\alpha) \not\subseteq B_* E$ or $I_*(\alpha) \subseteq E$. Suppose that $I_*(\alpha) \not\subseteq E$. Then there exists a $\beta \in I_*(\alpha)$ such that $\beta \notin E$. By P_{TCB}, there exists a $\gamma \in I_i(\alpha)$ such that $\beta \in I_*(\gamma)$. Since $\beta \notin E$, $I_*(\gamma) \not\subseteq E$. Hence $\gamma \notin B_* E$. Thus $I_i(\alpha) \not\subseteq B_* E$. (Sufficiency). Let $\alpha \in \mathbf{TCB}$. We want to show that $I_*(\alpha)$ satisfies property P_{TCB}. Fix arbitrary $i \in N$ and $\beta \in I_*(\alpha)$. We need to show that there exists a $\gamma \in I_i(\alpha)$ such that $\beta \in I_*(\gamma)$. Since $\beta \in I_*(\alpha)$, $I_*(\alpha) \not\subseteq \neg\{\beta\}$, that is, $\alpha \notin B_* \neg\{\beta\}$. Since $\alpha \in \mathbf{TCB}$, $\alpha \in \neg B_i B_* \neg\{\beta\} \cup B_* \neg\{\beta\}$. Hence $\alpha \in \neg B_i B_* \neg\{\beta\}$, that is, $I_i(\alpha) \not\subseteq B_* \neg\{\beta\}$. But this means that there exists a $\gamma \in I_i(\alpha)$ such that $\gamma \notin B_* \neg\{\beta\}$, that is, $I_*(\gamma) \not\subseteq \neg\{\beta\}$, i.e. $\beta \in I_*(\gamma)$. ■

Proof of Theorem 1 continued ($\mathbf{TCB} \cap B_* \mathbf{TCB} \subseteq \mathbf{NI}$). Let $\alpha \in \mathbf{TCB} \cap B_* \mathbf{TCB}$. By Lemma 2,

$$I_*(\alpha) \text{ satisfies property } P_{TCB}. \tag{1}$$

Since $\alpha \in B_* \mathbf{TCB}$, $I_*(\alpha) \subseteq \mathbf{TCB}$. Thus, by Lemma 2,

$$\forall\, \eta \in I_*(\alpha),\ I_*(\eta) \text{ satisfies property } P_{TCB}. \tag{2}$$

We want to show that $\alpha \in \mathbf{NI}$. The proof is illustrated in Figure 1, where an arrow labeled "i" from ω to ω' represents the fact that $\omega' \in I_i(\omega)$. By Remark 3, it is enough to show that for every $\beta, \gamma \in I_*(\alpha)$, $\gamma \in I_*(\beta)$. Fix arbitrary $\beta, \gamma \in I_*(\alpha)$. Since $\beta \in I_*(\alpha)$, there is a sequence $\langle i_1, \ldots i_m \rangle$ in N and a sequence $\langle \beta_0, \beta_1, \ldots, \beta_m \rangle$ in Ω such that: $\beta_0 = \alpha$, $\beta_m = \beta$ and, for every $k = 0, \ldots, m-1$, $\beta_{k+1} \in I_{i_{k+1}}(\beta_k)$. By definition of I_*, for every $k = 0, \ldots, m-1$, $\beta_{k+1} \in I_*(\alpha)$. By (1), since $\gamma \in I_*(\alpha)$, there exists a $\delta_1 \in I_{i_1}(\alpha)$ such that $\gamma \in I_*(\delta_1)$. Since $\delta_1, \beta_1 \in I_{i_1}(\alpha)$, by euclideanness of I_{i_1} it follows that $\delta_1 \in I_{i_1}(\beta_1)$. Thus, since $\gamma \in I_*(\delta_1)$ and $\delta_1 \in I_{i_1}(\beta_1)$, it follows from the definition of I_*

that $\gamma \in I_*(\beta_1)$. Since $\beta_1 \in I_*(\alpha)$, by (2) $I_*(\beta_1)$ satisfies P_{TCB}. Hence there exists a δ_2 such that $\delta_2 \in I_{i_2}(\beta_1)$ and $\gamma \in I_*(\delta_2)$. Since $\delta_2, \beta_2 \in I_{i_2}(\beta_1)$, by euclideanness of I_{i_2} it follows that $\delta_2 \in I_{i_2}(\beta_2)$. Hence, by definition of I_*, $\gamma \in I_*(\beta_2)$. Repeating this argument m times we get $\gamma \in I_*(\beta_m) = I_*(\beta)$, since $\beta_m = \beta$. ■

Insert Figure 1

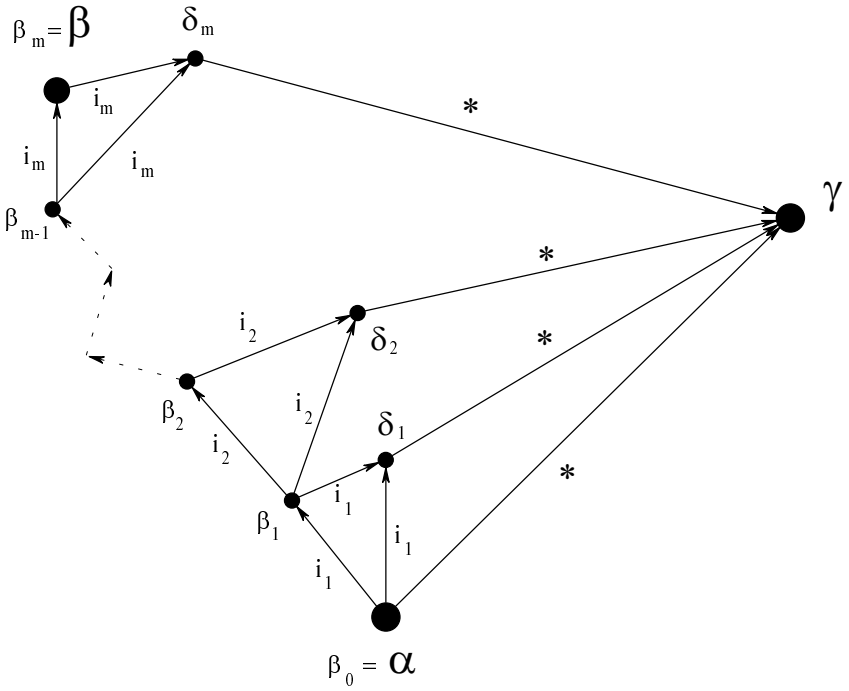

Figure 1

According to Theorem 1, Negative Introspection of common belief hinges on common knowledge of truth restricted to beliefs about common belief. One may wonder whether there is something qualitatively different about the truth of this very special type of beliefs. This question can be answered affirmatively, in that truth about common belief is necessary and sufficient for individuals' beliefs about common belief to *coincide*: we call this "Shared Worlds". (By comparison, having correct beliefs about what others believe, in general, does not imply sharing their beliefs.) Let **SW** be the following event:

$$\mathbf{SW} = \bigcap_{i \in N} \bigcap_{j \in N} \bigcap_{E \in 2^{\Omega}} (\neg B_i B_* E \cup B_j B_* E)$$

SW captures the notion that individuals agree on what is commonly believed: $\alpha \in \mathbf{SW}$ if and only if, for every event E, whenever one individual believes that it is common belief that E, then every other individual believes that too[10].

PROPOSITION 1. SW = TCB.

To prove Proposition 1 we need the following lemma.

LEMMA 3. For every $E \subseteq \Omega$, $\bigcap_{i \in N} B_i B_* E = B_* E$.

Proof of Lemma 3. That $B_* E \subseteq \bigcap_{i \in N} B_i B_* E$ follows from the definition of common belief (see, for example, Bonanno, 1996). Thus we only need to show that $\bigcap_{i \in N} B_i B_* E \subseteq B_* E$. Suppose not, that is, suppose that, for some $\alpha \in \Omega$ and for all $i \in N$, $I_i(\alpha) \subseteq B_* E$ and for some $\beta \in I_*(\alpha)$, $\beta \notin E$. By definition of I_*, there exist $i \in N$ and $\delta \in I_i(\alpha)$ such that $\beta \in I_*(\delta)$ (here we are using the fact that I_* is secondary reflexive: $\alpha \in I_*(\beta)$ implies $\alpha \in I_*(\alpha)$; secondary reflexivity of I_* is a direct consequence of secondary reflexivity of each I_i, which, in turn, is implied by euclideanness of I_i). Since $I_i(\alpha) \subseteq B_* E$ and $\delta \in I_i(\alpha)$, $\delta \in B_* E$, that is, $I_*(\delta) \subseteq E$. Hence $\beta \in E$, yielding a contradiction. ■

---

[10] Note that $\alpha \in \mathbf{SW}$ requires that at α the individuals share the same "model of the world" $\Omega_i(\alpha) \equiv \bigcup_{\omega \in I_i(\alpha)} I_*(\omega)$, that is, $\alpha \in \mathbf{SW}$ if and only if for all $i, j \in N$, $\Omega_i(\alpha) = \Omega_j(\alpha)$. Note that *common belief* in Shared Worlds rules out, by definition, even uncertainty about the others' model of the world, as the following example shows: $N = \{1,2\}$, $\Omega = \{\tau, \beta, \gamma\}$, $I_1(\tau) = \{\tau\}$, $I_1(\beta) = I_1(\gamma) = \{\beta\}$, $I_2(\tau) = I_2(\gamma) = \{\tau, \gamma\}$, $I_2(\beta) = \{\beta\}$. Thus $I_*(\tau) = I_*(\gamma) = \{\tau, \beta, \gamma\}$ and $I_*(\beta) = \{\beta\}$. Here $\mathbf{SW} = \{\tau, \beta\}$. However, while $\tau \in \mathbf{SW}$, $\tau \notin B_* \mathbf{SW} = \{\beta\}$: at τ (and γ) individual 2 is uncertain as to whether 1's personal model is $\{\beta\}$ or Ω.

Proof of Proposition 1. (**SW** $\subseteq$ **TCB)** Let $\alpha \in$ **SW**. Fix an arbitrary individual i and event E. We need to show that if $\alpha \in B_i B_* E$ then $\alpha \in B_* E$. Suppose that $\alpha \in B_i B_* E$. Then, since $\alpha \in$ **SW**, $\alpha \in B_j B_* E$ for every $j \in N$. Thus $\alpha \in \bigcap_{j \in N} B_j B_* E$. Hence, by Lemma 3, $\alpha \in B_* E$.

(**TCB** $\subseteq$ **SW**) Suppose that $\alpha \in$ **TCB**. Fix arbitrary $i, j \in N$ and $E \subseteq \Omega$. We want to show that if $\alpha \in B_i B_* E$ then $\alpha \in B_j B_* E$. Suppose that $\alpha \in B_i B_* E$. Then, since $\alpha \in$ **TCB**, $\alpha \in B_* E$. By definition of common belief, for every $j \in N$, $B_* E \subseteq B_j B_* E$. ■

Finally, since **NI** can be viewed as describing the "logic" of common belief, a global (or "axiomatic") version of Theorem 1 which incorporates Proposition 1 is of some interest. It is provided in the following corollary.

COROLLARY 1. **NI** = Ω if and only if **SW** = Ω [11].

Proof. Suppose that **NI** = Ω. Then, by Theorem 1 and Proposition 1, **SW** $\cap$ B_***SW** = Ω which implies that **SW** = Ω. Suppose that **SW** = Ω. Then B_***SW** = $B_* \Omega = \Omega$ (the latter equality follows from Necessity: see Remark 2). Thus **SW** $\cap$ B_***SW** = Ω. It follows from Theorem 1 and Proposition 1 that **NI** = Ω. ■

3. Further characterization of Negative Introspection of Common Belief: reducibility of *mixed* iterations of belief operators

In this section we prove another interesting characterization of Negative Introspection of common belief which throws light on the role of "transfinite" levels of a belief hierarchy (see

---

[11] That is, (i) and (ii) below are equivalent:

(i) $\forall E \subseteq \Omega$, $\neg B_* E \subseteq B_* \neg B_* E$,

(ii) $\forall i, j \in N$, $\forall E \subseteq \Omega$, $B_i B_* E \subseteq B_j B_* E$.

Dekel and Gul, 1997 and references therein). For common knowledge (i.e. when individual beliefs satisfy the logic of S5), the extension of higher-order events such as $B_1\neg B_2 B_* \neg B_1 B_* E$ is determined by the extension of the corresponding common knowledge events (for instance, $B_1\neg B_2 B_* \neg B_1 B_* E$ is equal to $B_* E$). Negative Introspection of common belief can be viewed as a special instance of such a reduction. Theorem 2 shows Negative Introspection to be equivalent, in general, to the reducibility of *mixed*[12] higher-order beliefs about common belief to either common belief or its negation. If m is a non-negative integer, define, for every event E, $\neg^m E$ to be E if m is even and $\neg E$ if m is odd.

THEOREM 2. The following conditions are equivalent:

(i) $\alpha \in \mathbf{NI}$,

(ii) for every event E, every sequence $\langle i_1, i_2, ..., i_k \rangle$ in N and every sequence $\langle \lambda_0, \lambda_1, ..., \lambda_k \rangle$ in $\{0,1\}$,

$$\alpha \in \neg^{\lambda_k} B_{i_k} \neg^{\lambda_{k-1}} B_{i_{k-1}} ... \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \quad \Leftrightarrow \quad \alpha \in \neg^{(\lambda_k + ... + \lambda_1 + \lambda_0)} B_* E.$$

REMARK 4. In the statement of Theorem 2 some or all of the B_{i_j} can be allowed to be the common belief operator B_* rather than the belief operator of a particular individual.

Proof. not (i) $\Rightarrow$ not (ii). Suppose that $\alpha \notin \mathbf{NI}$. Then there exists an event E such that $\alpha \in \neg B_* E$ and $\alpha \in \neg B_* \neg B_* E$. By definition of B_*, the latter implies that there exists a sequence $\langle i_1, i_2, ..., i_k \rangle$ in N such that $\alpha \in \neg B_{i_1} B_{i_2} ... B_{i_k} \neg B_* E$. This, together with $\alpha \in \neg B_* E$ yields a violation of (ii).

---

[12] It is well known that Negative Introspection of common belief implies the reducibility of finite iterations *of the common belief operator* B_* (such as $B_* \neg B_* B_* \neg B_*$) to either B_* or $\neg B_*$ (see Chellas, 1984, p. 154). The interest of Theorem 2 lies in the fact that it deals with *mixed* iterations, that is, iterations of *different* belief operators.

(i) $\Rightarrow$ (ii). We prove this in two steps. The first step (Lemma 4) deals with iterations of length k = 1. This step exploits the definition of **NI** as well as the implication **NI** $\subseteq$ **TCB** (cf. Theorem 1). The extension to iterations of arbitrary length is shown inductively and makes use of the belief-closedness of **NI**, that is, for all $i \in N$, **NI** $\subseteq B_i$**NI**.

LEMMA 4. For every $i \in N$, every $E \subseteq \Omega$ and every $\lambda, \mu \in \{0,1\}$,

$$\neg^{\lambda} B_i \neg^{\mu} B_* E \cap \mathbf{NI} = \neg^{(\lambda+\mu)} B_* E \cap \mathbf{NI}. \tag{3}$$

Proof of Lemma 4. Fix arbitrary $i \in N$, $E \subseteq \Omega$ and $\lambda, \mu \in \{0,1\}$. We need to consider four cases.

CASE 1: $\lambda = \mu = 0$. In this case (3) is

$$B_i B_* E \cap \mathbf{NI} = B_* E \cap \mathbf{NI} \tag{4}$$

By definition of common belief, $B_* E \subseteq B_i B_* E$. Hence $B_* E \cap \mathbf{NI} \subseteq B_i B_* E \cap \mathbf{NI}$. To prove the converse, let $\alpha \in B_i B_* E \cap \mathbf{NI}$. By Theorem 1, since $\alpha \in \mathbf{NI}$, $\alpha \in \mathbf{TCB}$. Hence, since $\alpha \in B_i B_* E$, it follows that $\alpha \in B_* E$. Thus $\alpha \in B_* E \cap \mathbf{NI}$.

CASE 2: $\lambda = 0$ and $\mu = 1$. In this case (3) is

$$B_i \neg B_* E \cap \mathbf{NI} = \neg B_* E \cap \mathbf{NI}. \tag{5}$$

By definition of common belief, $B_* E \subseteq B_i B_* E$. By consistency (see Remark 1), $B_i B_* E \subseteq \neg B_i \neg B_* E$. Hence $B_* E \subseteq \neg B_i \neg B_* E$, which is equivalent to $B_i \neg B_* E \subseteq \neg B_* E$. It follows that $B_i \neg B_* E \cap \mathbf{NI} \subseteq \neg B_* E \cap \mathbf{NI}$. To prove the converse, let $\alpha \in \neg B_* E \cap \mathbf{NI}$. Then $\alpha \in B_* \neg B_* E$. By definition of B_*, $B_* \neg B_* E \subseteq B_i \neg B_* E$. Thus $\alpha \in B_i \neg B_* E \cap \mathbf{NI}$.

CASE 3: $\lambda = 1$ and $\mu = 0$. In this case (3) is

$$\neg B_i B_* E \cap \mathbf{NI} = \neg B_* E \cap \mathbf{NI}. \tag{6}$$

which is equivalent to (4).

CASE 4: $\lambda = \mu = 1$. In this case (3) is

$$\neg B_i \neg B_* E \cap \mathbf{NI} = B_* E \cap \mathbf{NI}. \tag{7}$$

which is equivalent to (5). ■

Proof of Theorem 2 ctnd. The statement (i) $\Rightarrow$ (ii) is equivalent to the following: for every event E, every sequence $\langle i_1, i_2, ..., i_k \rangle$ in N and every sequence $\langle \lambda_0, \lambda_1, ..., \lambda_k \rangle$ in $\{0,1\}$,

$$\neg^{\lambda_k} B_{i_k} \neg^{\lambda_{k-1}} B_{i_{k-1}} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap \mathbf{NI} = \neg^{(\lambda_k + \ldots + \lambda_1 + \lambda_0)} B_* E \cap \mathbf{NI}. \quad (8)$$

We prove this by induction. If k = 1 then (8) is true by Lemma 4. Suppose that (8) is true for every sequence of length k. We want to show that it is true for every sequence of length k+1. Fix an arbitrary event E and arbitrary sequences $\langle i_1, i_2, \ldots, i_k, i_{k+1} \rangle$ in N and $\langle \lambda_0, \lambda_1, \ldots, \lambda_k, \lambda_{k+1} \rangle$ in {0,1}. By the induction hypothesis, letting $\lambda = \lambda_k + \lambda_{k-1} + \ldots + \lambda_0$,

$$\neg^{\lambda_k} B_{i_k} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap \mathbf{NI} = \neg^{\lambda} B_* E \cap \mathbf{NI}. \quad (9)$$

Hence, by Conjunction (cf. Remark 2),

$$B_{i_{k+1}} \neg^{\lambda_k} B_{i_k} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap B_{i_{k+1}} \mathbf{NI} = B_{i_{k+1}} \neg^{\lambda} B_* E \cap B_{i_{k+1}} \mathbf{NI}. \quad (10)$$

It follows from (10) that

$$\neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda_k} B_{i_k} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap B_{i_{k+1}} \mathbf{NI} = \neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda} B_* E \cap B_{i_{k+1}} \mathbf{NI}. \quad (11)$$

By Lemma 1, $\mathbf{NI} \subseteq B_* \mathbf{NI}$. By definition of B_*, $B_* \mathbf{NI} \subseteq B_{i_{k+1}} \mathbf{NI}$. Hence $\mathbf{NI} \subseteq B_{i_{k+1}} \mathbf{NI}$. It follows from this and (11) that

$$\neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda_k} B_{i_k} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap \mathbf{NI} = \neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda} B_* E \cap \mathbf{NI}. \quad (12)$$

By Lemma 4,

$$\neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda} B_* E \cap \mathbf{NI} = \neg^{(\lambda_{k+1} + \lambda)} B_* E \cap \mathbf{NI}. \quad (13)$$

From (12) and (13) we obtain,

$$\neg^{\lambda_{k+1}} B_{i_{k+1}} \neg^{\lambda_k} B_{i_k} \ldots \neg^{\lambda_1} B_{i_1} \neg^{\lambda_0} B_* E \cap \mathbf{NI} = \neg^{(\lambda_{k+1} + \lambda)} B_* E \cap \mathbf{NI}, \quad (14)$$

as desired. ■

4. Game-Theoretic Relevance of Negative Introspection of Common Belief

We conclude by discussing the role of Negative Introspection of common belief in the epistemic foundations of game theory.

§ 4.1. Correlated equilibrium under incomplete information

Perhaps the first and one of the most ambitious contributions to the epistemic foundations of game theory is Aumann's (1987) claim that correlated equilibrium can be viewed as an expression of Bayesian rationality. In contrast to his seminal paper that introduced the notion of correlated equilibrium (Aumann, 1974), the new characterization appeals to the notion of a common prior in a situation of incomplete information, where there is no *ex ante* stage and the primitives of the model are the individuals' belief hierarchies. As pointed out recently (Gul, 1996, Dekel and Gul, 1997, Lipman, 1995), the *meaning* of a common prior in situations of incomplete information is highly problematic. This skepticism can be developed along the following lines. As Mertens and Zamir (1985) showed in their classic paper, the description of the "actual world" in terms of belief hierarchies generates a collection of "possible worlds", one of which is the actual world. This set of possible worlds, or states, gives rise to a formal similarity between situations of incomplete information and those of asymmetric information (where there is an ex ante stage at which the individuals have identical information and subsequently update their beliefs in response to private signals). However, while a state in the latter represents a real contingency, in the former it is "a fictitious construct, used to clarify our understanding of the real world" (Lipman, 1995, p.2), "a notational device for representing the profile of infinite hierarchies of beliefs" (Gul, 1996, p. 3). As a result, notions such as that of a common prior, "seem to be based on giving the artificially constructed states more meaning than they have" (Dekel and Gul, 1997, p.115). Thus an essential step in providing a justification for correlated equilibrium under incomplete information is to provide an interpretation of the common prior based on "assumptions that do not refer to the constructed state space, but rather are assumed to hold in the true state", that is, assumptions "that

only use the artificially constructed states the way they originated – namely as elements in a hierarchy of belief" (Dekel and Gul, 1997, p.116).

An interpretation of the desired kind of the common prior assumption under incomplete information was provided recently (Bonanno and Nehring, 1996; see also Feinberg, 1995) in terms of a generalized notion of absence of agreeing to disagree à la Aumann (1976), called consistency of expectations. In order to introduce the relevant notions we need to refine qualitative frames by adding degrees of belief. In what follows the set of states Ω is assumed to be finite.

DEFINITION 2. A *Bayesian frame based on the qualitative frame* $\mathcal{Q}$ is a tuple $\mathcal{B} = \langle \mathcal{Q}, \{p_i\}_{i \in N} \rangle$ where

- for every individual $i \in N$, $p_i : \Omega \to \Delta(\Omega)$ (where $\Delta(\Omega)$ denotes the set of probability distributions over Ω) is a function that specifies i's *probabilistic beliefs,* satisfying the following properties[13] [we use the notation $p_{i,\alpha}$ rather than $p_i(\alpha)$]: $\forall\, \alpha, \beta \in \Omega$,

  (i) $\text{supp}(p_{i,\alpha}) = I_i(\alpha)$, and

  (ii) if $I_i(\alpha) = I_i(\beta)$ then $p_{i,\alpha} = p_{i,\beta}$.

Thus $p_{i,\alpha} \in \Delta(\Omega)$ is individual i's subjective probability distribution at state α and the above two conditions say that every individual knows her own beliefs. We denote by $\|p_i = p_{i,\alpha}\|$ the event $\{\omega \in \Omega : p_{i,\omega} = p_{i,\alpha}\}$. It is clear that $\|p_i = p_{i,\alpha}\| = \|I_i = I_i(\alpha)\|$, where $\|I_i = I_i(\alpha)\| = \{\omega \in \Omega : I_i(\omega) = I_i(\alpha)\}$, and that the set $\{\|p_i = p_{i,\omega}\| : \omega \in \Omega\}$ is a partition of Ω; it is often referred to as individual i's *type partition*.

DEFINITION 3. At state α there is Consistency of Expectations if there do not exist random variables $Y_i : \Omega \to \mathbb{R}$ ($i \in N$) such that: (1) $\forall\, \omega \in \Omega$, $\sum_{i \in N} Y_i(\omega) = 0$, and (2) at α it is

common belief that, for every individual i, i's subjective expectation of Y_i is positive, that is, $\alpha \in B_*(\|E_1 > 0\| \cap \ldots \cap \|E_n > 0\|)$, where $\|E_i > 0\| = \{\omega \in \Omega: \sum_{\omega' \in \Omega} Y_i(\omega') p_{i,\omega}(\omega') > 0\}$.

Consistency of Expectations turns out to be *equivalent* to a particular local version of the Common Prior Assumption defined as follows.

DEFINITION 4. For every $\mu \in \Delta(\Omega)$, let $\mathbf{HQC}_\mu$ (for Harsanyi Quasi Consistency with respect to the "prior" μ) be the following event: $\alpha \in \mathbf{HQC}_\mu$ if and only if

(1) $\forall i \in N$, $\forall \omega, \omega' \in I_*(\alpha)$, if $\mu(\|p_i = p_{i,\omega}\|) > 0$ then $p_{i,\omega}(\omega') = \dfrac{\mu(\omega')}{\mu(\|p_i = p_{i,\omega}\|)}$ if $\omega' \in \|p_i = p_{i,\omega}\|$ and $p_{i,\omega}(\omega') = 0$ otherwise (that is, $p_{i,\omega}$ is obtained from μ by conditioning on $\|p_i = p_{i,\omega}\|$)[14], and

(2) $\mu(I_*(\alpha)) > 0$.

If $\alpha \in \mathbf{HQC}_\mu$, μ is a *local common prior at* α. Furthermore, let $\mathbf{HQC} = \bigcup_{\mu \in \Delta(\Omega)} \mathbf{HQC}_\mu$.

PROPOSITION 2.[15] At α Consistency of Expectations is satisfied if and only if $\alpha \in \mathbf{HQC}$.

The above proposition shows that **HQC** is the natural way of expressing Harsanyi consistency locally.

---

[13] If μ is a probability distribution over Ω, we denote by supp(μ) the support of μ, that is, the set of states to which μ assigns positive probability.

[14] Where, for every event E, $\mu(E) = \sum_{\omega \in E} \mu(\omega)$. Note that, for every $\omega \in \Omega$ and $i \in N$, $\omega \in \|p_i = p_{i,\omega}\|$. Thus $\mu(\omega) > 0$ implies $\mu(\|p_i = p_{i,\omega}\|) > 0$.

[15] For a proof see Bonanno and Nehring (1996). This result is a local version of Morris's (1994) characterization of no trade under asymmetric information. See also Feinberg (1995).

Harsanyi Quasi Consistency may seem weaker than expected in that condition (2) of its definition only requires the derived common prior to assign positive probability to some commonly possible state but allows the true state to be assigned zero "prior" probability. As illustrated in the example of Figure 2, Agreement and No Trade-type arguments cannot deliver more. [16]

Insert Figure 2

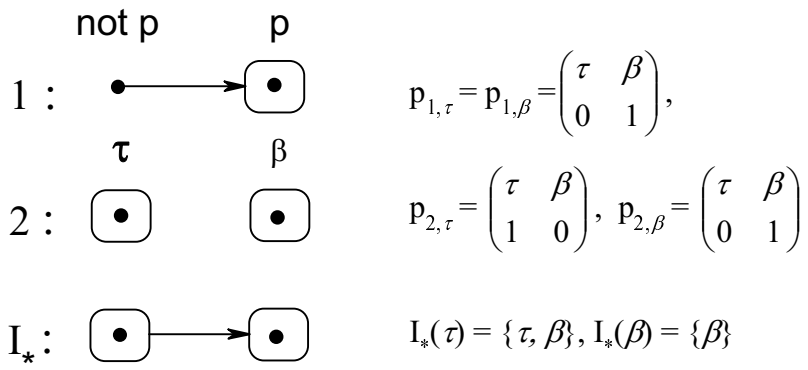

Figure 2

In this example, at the true state individual 1 wrongly believes that it is common belief that p, while individual 2 correctly believes that not p is the case and knows 1's incorrect beliefs. Expectation consistency is satisfied at the true state (as well as at β). In fact, let Y_1 and Y_2 be random variables on $\{\tau, \beta\}$ such that $Y_2 = -Y_1$ and suppose that $\tau \in B_*\|E_1 > 0\|$, that is, at τ it is common belief that individual 1's expectation of Y_1 is positive. Then $Y_1(\beta) > 0$, hence $Y_2(\beta) < 0$. Thus $\beta \notin \|E_2 > 0\|$, that is, at β 2's expectation of Y_2 cannot be positive. Since $\beta \in I_*(\tau)$, it follows that $\tau \notin B_*\|E_2 > 0\|$. Thus Agreement is necessarily satisfied at τ. By Proposition 2 there must be a μ such that $\tau \in \mathbf{HQC}_\mu$. Indeed such a local common prior is given by $\mu(\beta) = 1$.

---

[16] In Figures 2 - 5 for every $k \in \{1, 2, *\}$, $\omega' \in I_k(\omega)$ if and only if either ω and ω' belong to the same cell (represented as a rounded rectangle) or there is an arrow from ω, or the cell containing ω, to ω', or the cell containing ω'.

Is Harsanyi Quasi Consistency an adequate epistemic basis for correlated equilibrium? Perhaps not too surprisingly in view of the previous example, Harsanyi Quasi Consistency is insufficient *by itself*, as demonstrated by the following example. Figures 3a and 3b show a two-person zero-sum game with a unique correlated equilibrium (B,R), and an epistemic model of it.

Insert Figure 3

		Player 2		
		L	C	R
Player 1	T	10, –10	–10, 10	–7 , 7
	M	–10, 10	10, –10	–7 , 7
	B	7 , –7	7 , –7	0 , 0

Figure 3a

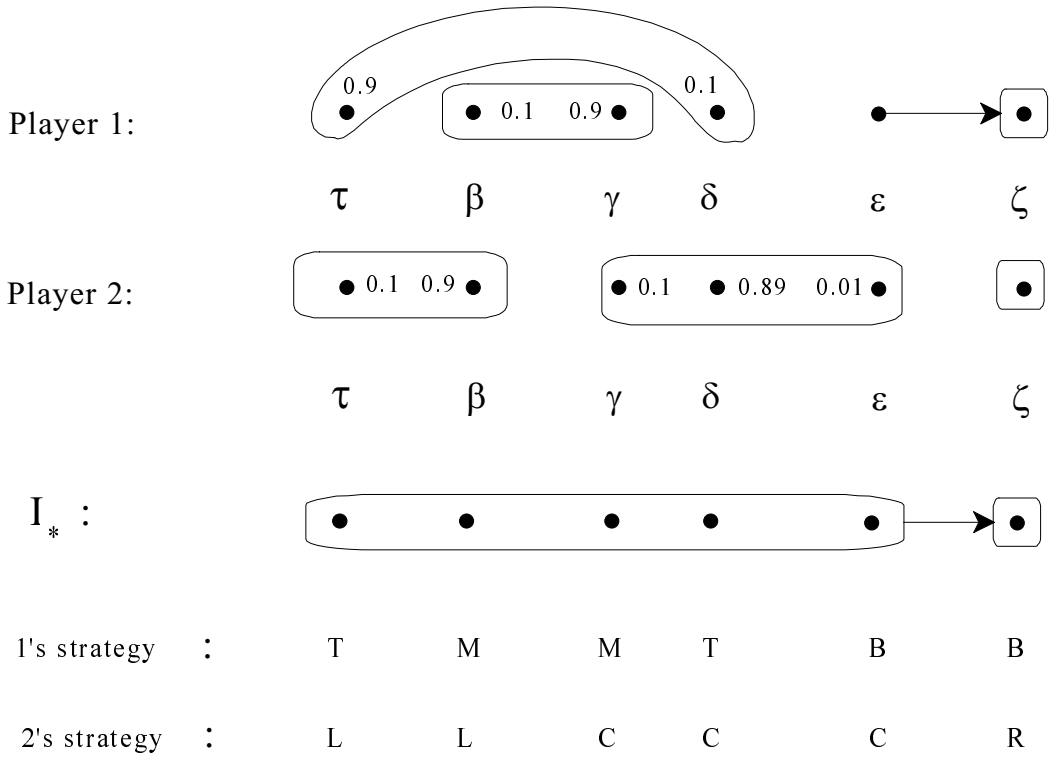

Figure 3b

In this example, at τ (i) the players' beliefs satisfy Harsanyi Quasi Consistency ($\tau \in \mathbf{HQC}_{\mu} = \Omega$ where $\mu(\zeta) = 1$), (ii) there is common belief in rationality ($I_*(\tau) = \Omega$ and at every state each player's strategy is optimal given her beliefs) and (iii) no individual has any false beliefs. Yet at τ the players play (T,L) which is not a correlated equilibrium.

Note that in the above example, although the derived common prior assigns zero probability to τ, there is no sense in which the belief hierarchies described by the true state are "improbable" and constitute a null event. Indeed the actual beliefs of all players assign positive probability to τ.

The above example is in fact quite general. By a straightforward generalization of its construction any profile of correlated rationalizable strategies – where one strategy is a unique best response to some distribution over correlated rationalizable strategies of the other players – can be

realized at the true state τ of a Bayesian frame where $\tau \in$ **HQC** (and no individual has false beliefs).

What seems to go wrong in the example is that, while Player 2 believes Player 1 to be wrong at ε, this does not show up as disagreement – and hence as a violation of Harsanyi Quasi Consistency – since Player 1 falsely believes at ε that there is agreement that the true state is ζ. Hence **TCB** is violated at ε, and therefore B_***TCB** at τ.

Indeed – in the absence of false beliefs at the true state – B_***TCB** is exactly what needs to be added to **HQC** to ensure the play of a correlated equilibrium strategy-profile, as the following theorem shows.

Given a finite normal-form game G and a Bayesian frame $\mathcal{B}$ one obtains a model for G by associating with each state $\omega \in \Omega$ a strategy profile in such a way that for every player i, if $p_{i,\alpha} = p_{i,\beta}$ then i's strategy at α is the same as i's strategy at β (that is, the type of a player specifies not only her beliefs but also her strategy). Given a model of G we denote by **RAT** the event (i.e. set of states) where every player is rational in the sense that her strategy maximizes her expected utility given her beliefs. Let **T** (for Truth) be the following event:

$$\mathbf{T} = \bigcap_{i \in N} \bigcap_{E \in 2^{\Omega}} \neg(B_i E \cap \neg E).$$

Thus, for every $\alpha \in \Omega$, $\alpha \in$ **T** if and only if no individual has any false beliefs at α (for every $i \in N$ and for every $E \subseteq \Omega$, if $\alpha \in B_i E$ then $\alpha \in E$)[17]. To take account of the incomplete information context, we call a strategy profile a *correlated equilibrium* if it is played with positive probability in some correlated equilibrium (in the ordinary sense).

---

[17] It is well known that $\alpha \in$ **T** if and only if $\alpha \in \bigcap_{i \in N} I_i(\alpha)$. It follows that $\alpha \in B_*$**T** if and only if, for all $\beta \in I_*(\alpha)$, $\beta \in \bigcap_{i \in N} I_i(\beta)$.

THEOREM 3. Fix an arbitrary finite normal-form game G and an arbitrary model of G satisfying the following properties at the true state:

(1) $\tau \in \mathbf{T} \cap B_*\mathbf{TCB}$ (the actual beliefs of the players are correct and there is common belief in Truth about common belief),

(2) $\tau \in B_*\mathbf{RAT}$, (there is common belief in rationality)

(3) $\tau \in \mathbf{HQC}$ (Harsanyi Quasi Consistency of beliefs, that is, Agreement, is satisfied).

Then the strategy profile associated with τ (i.e. the strategy profile actually played) is a correlated equilibrium.

On the other hand, as the example of Figure 3 shows, if (2) and (3) are satisfied and (1) is weakened to $\tau \in \mathbf{T}$ then the strategy profile associated with τ need not be a correlated equilibrium.

Proof. In Bonanno and Nehring (1996) it is shown that, for every $\mu \in \Delta(\Omega)$, if $\tau \in \mathbf{HQC}_\mu \cap \mathbf{T} \cap \mathbf{NI}$ then $\mu(\omega) > 0$ for every $\omega \in I_*(\tau) \cup \{\tau\}$. In view of Theorem 1 of Section 2, since $\mathbf{T} \subseteq \mathbf{TCB}$, the same holds for any μ such that $\tau \in \mathbf{HQC}_\mu \cap \mathbf{T} \cap B_*\mathbf{TCB}$. Furthermore, the frame restricted to $I_*(\tau) \cup \{\tau\}$ is a partitional frame. Thus a straightforward adaptation of Aumann's (1987) proof yields that the strategy profile associated with τ is a correlated equilibrium. ■

REMARK 5. Mathematically, it is clear from the proof that **NI** is the critical property to bridge the gap. If condition (1) is weakened to $\tau \in \mathbf{NI}$ (or, equivalently, $\tau \in \mathbf{TCB} \cap B_*\mathbf{TCB}$) then the conclusion is that $\tau \in B_*\mathbf{CE}$, where **CE** is the event that a correlated equilibrium is played; that is, at the true state it is common belief that a correlated equilibrium is played.

Thus one sees that once the rather mild-looking property of Negative Introspection of common belief is satisfied, **HQC** is re-instated with the proper strength.

§ 4.2. Backward induction.

Another important issue within the epistemic foundations of game theory concerns the justification for backward induction. For a class of extensive games, which includes the finitely repeated prisoners' dilemma and the centipede game, it has recently been shown (Stalnaker, 1994,

1996, Stuart, 1997) that if at the true state τ of an epistemic model of a game there is common belief in rationality and common belief that no individual has any false beliefs (that is, $\tau \in B_*\mathbf{T}$), then the strategy profile associated with that state gives rise to the backward induction outcome. In this class of games the backward induction outcome is the unique Nash equilibrium outcome. A Stuart/Stalnaker justification is much more compelling than the standard one (cf. Aumann and Brandenburger, 1995) in that *it does not make any assumptions on beliefs (or conjectures) being mutually known*.

In game-theoretic contexts where zero-probability events play a crucial role, the assumption of common belief in no error ($B_*\mathbf{T}$) may be viewed as rather strong. Additional arguments are needed to asses its strength. We will do so by accounting for common belief in no error as the conjunction of individually weaker properties.

One such property is *quasi-coherence* of beliefs, defined as the common *possibility* of common belief in no error: $\mathbf{Q} = \neg B_*\neg B_*\mathbf{T}$. In Bonanno and Nehring (1997) this is shown to be equivalent to the requirement that it cannot be the case that every individual can make *unbounded* gains from betting (assuming individuals to be "moderately" risk-averse). From an economic point of view this seems as compelling a property as one might imagine[18].

Is quasi-coherence strong enough to obtain the backward-induction result? The answer is negative, as the following example (taken from Stuart, 1997, p. 138) shows. Figures 4a and 4b show the one-period version of the prisoners' dilemma and a model of a two-period repetition (T-T denotes the second-period strategy of imitating the opponent's first-period choice).

Insert Figure 4

[18] Quasi-coherence can be thought of as the qualitative counterpart to Harsanyi Quasi Consistency. Bonanno and Nehring (1997) show that it is equivalent to the impossibility of agreeing to disagree about "union-consistent" qualitative belief indices.

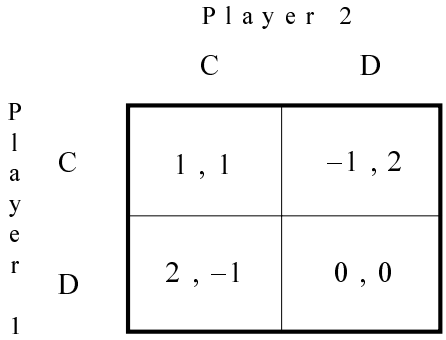

Figure 4a

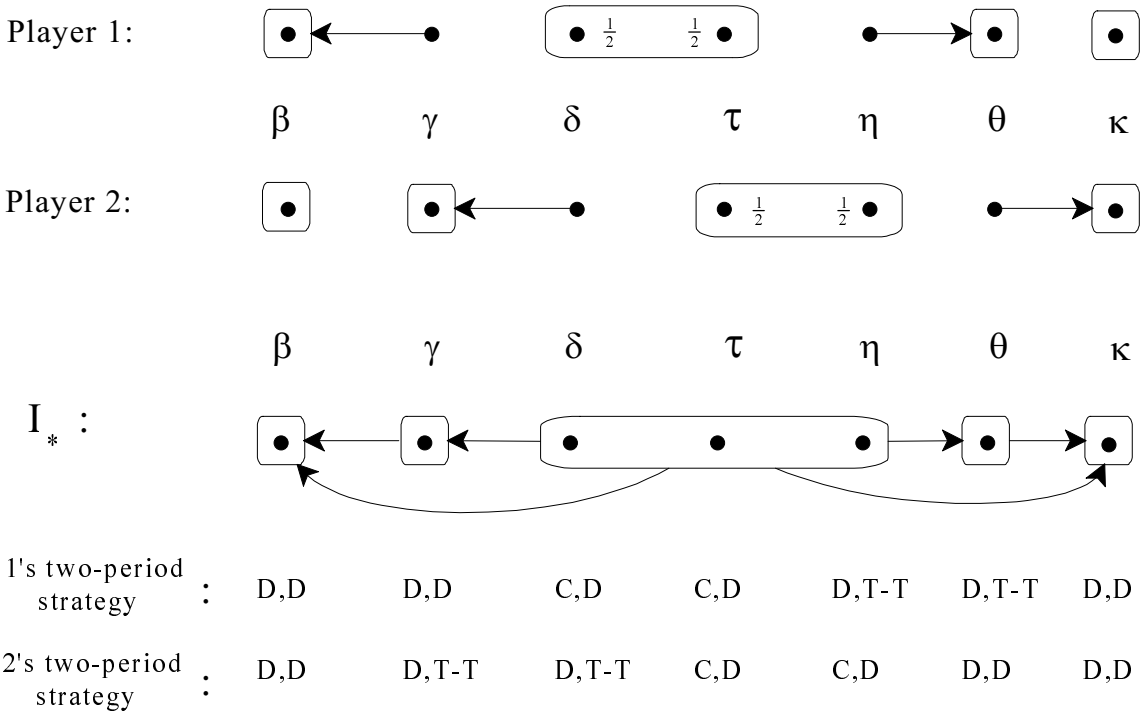

Figure 4b

In this example (i) $\tau \in \mathbf{Q}$ ($\neg B_* \neg B_* \mathbf{T} = \Omega$ since $B_* \mathbf{T} = \{\beta, \kappa\}$), (ii) there is common belief in rationality (at every state both players are maximizing the expected sum of their two-period payoffs) and (iii) no player has any false beliefs ($\mathbf{T} = \{\beta, \tau, \kappa\}$). Yet the players cooperate in the first period. (Note that in this example $\mathbf{Q} = \mathbf{HQC}$, while in general $\mathbf{HQC} \subseteq \mathbf{Q}$.)

In analogy to the case of correlated equilibrium, the gap between quasi-coherence and common belief in no error is bridged by **NI** (or, equivalently, **TCB** $\cap$ B_***TCB**): **Q** $\cap$ **NI** = B_***T**.[19] This decomposition of B_***T** helps since **Q** seems uncontroversial and B_***TCB** by itself is substantially weaker than B_***T**, as shown in the following example where there is maximal disagreement.

Insert Figure 5

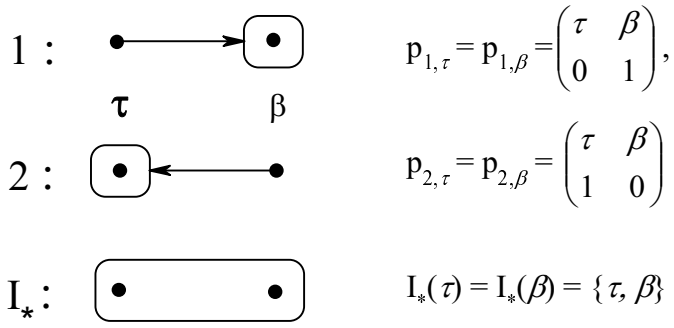

Figure 5

In this example, **NI** = **TCB** = B_***TCB** = Ω; notice also that here common belief and common knowledge[20] coincide; in any such Bayesian frame common belief must satisfy Negative Introspection, since common knowledge does by definition.

---

[19] In the example of Figure 4, Negative Introspection of common belief fails at τ: **NI** = $\{\beta, \kappa\}$.

[20] Where the individual knowledge operators are obtained from the type partitions (see Definition 2).

References

Aumann, R. (1974), Subjectivity and correlation in randomized strategies, *Journal of Mathematical Economics*, 1, 67-96.

Aumann, R. (1976), Agreeing to disagree, *Annals of Statistics,* 4, 1236-1239.

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

Aumann, R. and A. Brandenburger (1995), Epistemic conditions for Nash equilibrium, *Econometrica*, 63, 1161-80.

Bacharach, M. (1985), Some extensions of a claim of Aumann in an axiomatic model of knowledge, *Journal of Economic Theory*, 37, 167-190.

Bonanno, G. (1996), On the logic of common belief, *Mathematical Logic Quarterly*, 42, 305-311.

Bonanno, G. and K. Nehring (1996), How to make sense of the Common Prior Assumption under incomplete information, Working paper, University of California Davis.

Bonanno, G. and K. Nehring (1997), Assessing the Truth Axiom under incomplete information, Working paper, University of California Davis.

Brandenburger, A. and E. Dekel (1987), Common knowledge with probability 1, *Journal of Mathematical Economics*, 16, 237-246.

Chellas, B. (1984), *Modal logic*, Cambridge University Press, Cambridge.

Colombetti, M. (1993), Formal semantics for mutual beliefs, *Artificial intelligence*, 62, 341-353.

Dekel, E. and F. Gul (1997), Rationality and knowledge in game theory, in Kreps D. M. and K. F. Wallis (eds.), *Advances in Economics and Econometrics*, vol. 1, Cambridge University Press.

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

Feinberg, Y. (1995), A converse to the Agreement Theorem, Discussion Paper # 83, Center for Rationality and Interactive Decision Theory, Jerusalem.

Geanakoplos, J. (1992), Common knowledge, *Journal of Economic Perspectives*, 6, 53-82.

Gul, F. (1996), A comment on Aumann's Bayesian view, mimeo [forthcoming in *Econometrica*].

Harsanyi, J. (1967-68), Games with incomplete information played by "Bayesian players", Parts I-III, *Management Science*, 8, 159-182, 320-334, 486-502.

Heifetz, A. (1996), Common belief in monotonic epistemic logic, *Mathematical Social Sciences*, 32, 109-123.

Kaneko, M. and T. Nagashima (1991), Final decisions, the Nash equilibrium and solvability in games with common knowledge of logical abilities, *Mathematical Social Sciences*, 22, 229-255.

Kaneko, M. and T. Nagashima (1997), Indefinability of the common knowledge concept in finitary logics, in M. Bacharach, L.A. Gérard-Varet, P. Mongin and H. Shin (Eds.), *Epistemic logic and the theory of games and decisions*, Kluwer Academic.

Lewis, D. (1969), *Convention: a philosophical study,* Harvard University Press, Cambridge (MA).

Lipman, B. (1995), Approximately common priors, mimeo, University of Western Ontario.

Lismont, L. and P. Mongin (1994), On the logic of common belief and common knowledge, *Theory and Decision*, 37, 75-106.

Lismont, L. and P. Mongin (1995), Belief closure: a semantics for common knowledge for modal propositional logic, *Mathematical Social Sciences*, 30, 127-153.

Mertens, J-F. and S. Zamir (1985), Formulation of Bayesian analysis for games with incomplete information, *International Journal of Game Theory,* 14, 1-29.

Milgrom, P. (1981), An axiomatic characterization of common knowledge, *Econometrica*, 49, 219-222.

Morris, S. (1994), Trade with heterogeneous prior beliefs and asymmetric information, *Econometrica*, 62, 1327-47.

Rubinstein, A. and A. Wolinsky (1990), On the logic of "Agreeing to Disagree" type of results, *Journal of Economic Theory*, 51, 184-193.

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

Stalnaker, R. (1994), On the evaluation of solution concepts, *Theory and Decision*, 37, 49-74.

Stalnaker, R. (1996), Knowledge, belief and counterfactual reasoning in games, *Economics and Philosophy*, 12, 133-163.

Stuart, H. (1997), Common belief of rationality in the finitely repeated Prisoners' Dilemma, *Games and Economic Behavior*, 19, 133-143.