Believing the Unbelievable: the Dilemma of Self-Belief*

Hsueh-Ling Huynh and Balázs Szentes†

November, 1999

Abstract

In this paper, 'procedural rationality' is interpreted to be the ability to state one's own beliefs, and make decisions through logical deductions. We show that if a belief-system is consistent, deductively closed, and contains arithmetic, then two forms of self-belief– the ability to define one's own belief-system and the Principle of Positive Introspection, become irreconcilable.

This shows that the notion of a 'universal type' is untenable when this form of procedural rationality is taken into account.

1 Introduction

To paraphrase Herbert Simon (1976, 1979), rationality is the pursuit of a well-conceived objective by a well-conceived procedure. Such a general idea is obviously capable of many interpretations.

In this paper, we give a particular and precise meaning to this dictum. For an objective to be 'well-conceived', the decision maker must be able to express it in a sufficiently rich language. And a 'well-conceived procedure' requires the decision maker to deduce her action logically from her stated objectives, principles, knowledge and beliefs.

*The title of this paper is a pun on a famous paper: R. Aumann (1976). We are grateful to Kenneth Binmore, Debraj Ray, Robert Rosenthal, and Aldo Rustichini for their interest and comments in this paper. Debraj Ray also drew our attention to a recent work by Brandenburger and Keisler.

†Department of Economics, Boston University, 270 Bay State Road, Boston MA 02215. Email: hlhuynh@bu.edu, szentes@bu.edu.

We feel that this represents at least one important aspect of what is commonly perceived to be 'rationality'. Similar ideas were advocated in Binmore (1987), who anticipated the new and interesting difficulties that will arise in the interpretation of game theory when this constraint on rationality is taken into account. These ideas are also explored in Bacharach (1987); indeed, he explicitly modeled the decision process of a rational agent by first-order logic, as we do here.

In the terminology of Simon (1976), our interpretation focuses upon 'procedural rationality' rather than 'substantive rationality', as nothing is said about the exact nature of the objectives that a rational decision maker has to embrace, as long as they are definable. Indeed, we do not discuss in any further detail what sort of objectives are admissible in our theory. However, we do emphasize one important consequence of the definability of objectives, namely, the decision maker must possess a certain amount of self-knowledge. For example, in order for a Bayesian decision maker to be procedurally rational in our sense, the principle "My utility function is U and my actions maximize the expected value of this utility function", or some logical equivalent, must be explicitly represented as part of her beliefs.

For the rest of the paper, our discussion is concerned solely with the decision maker's beliefs. As we shall see, it already reveals some interesting implications of our interpretation.

Some authors have considered another interpretation of procedural rationality, where a decision method is considered to be a well-conceived procedure if it is *computable,* that is, if it can be implemented by a specific type of algorithm, say a finite automaton or a Turing machine, in finite time. This approach is taken, for example, by Anderlini (1990), Binmore and Shin (1992), and Canning (1992). Their analysis and conclusions depend on the fact that decision maker effective in this sense cannot at the same time be *decisive* at all times.[1] By contrast, our investigation is based on the more general notion of *definability.* We shall derive an 'impossibility theorem', which is valid even when the decision maker is endowed with powers of deduction that go beyond what Binmore and Shin would call 'algorithmic knowledge'.

For us, a belief can be a statement of known fact, a hypothesis about nature or social institutions, a statement about one's own objectives and beliefs, a statement about another decision maker's objectives and beliefs, or an axiom of logic or mathematics.

[1] Discussion and illustration of some of these ideas can also be found in Rubinstein (1998).

A minimal requirement of rationality is that beliefs should be *consistent*, so that the decision maker does not believe in a logical contradiction. In addition, in order for a system of beliefs to be *effective* as a procedure for making decisions, the decision maker must have the ability to derive logical consequences from her beliefs (and then believe in the consequences as well). We do not, however, assume that there is an algorithm to carry this out.

As noted above, the decision maker's ability to define her objectives already entails certain self-beliefs. Beyond that, how rich should be the set of beliefs held by a decision maker about herself, if she is to be deemed 'rational'?

To make an assessment, let us imagine an interaction between two decision makers: SELF and OTHER. Whether an action of SELF is appropriate or not often depends on what she thinks OTHER's action may be. Since she has no other basis to figure out OTHER's decision than sheer logic and what she believes to be OTHER's beliefs, she might have to make a explicit conjecture specifying the entire system of OTHER's beliefs. But of course OTHER is just as capable as SELF, and therefore also able to make explicit and complete representation of SELF's beliefs. This in turn implies that the representation she makes of OTHER's beliefs contain an explicit and complete representation of her own beliefs. Note that we have not argued that these representations of beliefs should be true, only that a procedurally rational decision maker should possess the linguistic and logical ability sophisticated enough to define her own (and others') beliefs completely. This is in keeping with the spirit of procedural, as opposed to substantive, rationality.

In Bayesian decision theory, the belief-system of a rational decision-maker (together with her preferences) is called her *type.* The decision-maker's type determines her action in certain, possibly interactive, situations. It is shown in Merten and Zamir (1985) that it is possible to construct a 'universal type space', whose elements will determine actions in all situations, and the type of another agent is fully represented in the type of one's own type. Therefore it seems unproblematic to regard the whole belief-system of as a single, knowable, object. However, our results shows that the notion of a universal type is untenable when we take into account the requirements of procedural rationality, at least as interpreted here. The formal implication of these ideas in game theory is pursued in a recent paper of Brandenburger and Keisler (1999), which makes use of the same notion of definability as here.

Another seemingly natural requirement on self-beliefs stems from more psychological

considerations. This is the Principle of Positive Introspection, stating that the decision maker is fully aware of her own beliefs. Formally: if the decision maker believes in X, then she also believes the statement "I believe in X". Furthermore, the decision maker explicitly believes in this principle itself. This postulate is considered to be sound in many contexts, even when the decision maker is not assumed to be fully aware of accessible facts, or to be aware of her own ignorance of them; see, for example, Modica and Rustichini (1999) and Geanakoplos (1989, 1992).

We shall give precise formulation of the two forms of self-belief, and demonstrate that they are in fact mutually inconsistent. The argument is derived from the Gödel's ideas used in his famous Incompleteness Theorem.

Some basic knowledge of mathematical logic will be used, namely, predicate calculus and the first-order theory of arithmetic. See, for example, the book by H. B. Enderton.

2 Expressible Beliefs

We consider a formal language, which is sufficiently rich to allow its user to state propositions in arithmetic. For example: "for all n, there exists an m such that $m \gneq n$", or (more symbolically) "$\forall n, x, y, z[(\ n \geq 3$ and $x \neq 0$ and $y \neq 0$ and $z \neq 0) \rightarrow (\ x^n + y^n \neq z^n)]$". It also allows the user to make statements like "I believe φ", abbreviated $B\varphi$, where φ is another statement in the language. The language is also closed under finite applications of the Boolean operations: $\neg$, $\vee$, and $\wedge$. Denote by $\mathfrak{L}$ the collection of all the statements in this language. The totality of a decision maker's beliefs, or her *belief-system*, is a subset $\mathfrak{B} \subseteq \mathfrak{L}$.

We impose the following postulates on $\mathfrak{B}$.

(1) $\mathfrak{B}$ *is consistent.*

This says that we cannot deduce a contradiction from the statements in $\mathfrak{B}$. In particular, $\mathfrak{B}$ cannot contain both a statement φ and its negation $\neg\varphi$.

(2–) $\mathfrak{B}$ *is deductively closed.*

This means that if $\varphi_1, ..., \varphi_n$ is a (finite) collection of statements in $\mathfrak{B}$ and φ can be inferred from $\varphi_1, ..., \varphi_n$ by the rules of Predicate Calculus, then $\varphi \in \mathfrak{B}$.

(2+) *The set* $\{\varphi \in \mathfrak{L} \mid B\varphi \in \mathfrak{B}\}$ *is deductively closed.*

Postulate (2) says that the decision maker is fully aware of all the logical consequences of her beliefs. This is sometimes called "logical omniscience". From the practical standpoint, this is a very strong assumption. Indeed, no real-life decision maker has this ability

at any given moment in time. On the other hand, it is not very clear what would be a suitable weakening of this postulate, for clearly every rational being has some ability to make logical inferences.[2] We shall see that, even with this idealization, the decision-maker's self-beliefs cannot be complete. In the proof of our main result we only require the decision maker to make a handful of deductions; the postulate of logical omniscience is stated here only because it appears more natural than the special finite collection of logical rules and axioms used in the proof.

Postulate (2+) says that the decision maker is not only able to carry out logical inferences, but she is aware that she is doing so.

We shall refer to (2–) and (2+) as Postulate (2).

(3a) $\mathfrak{B}$ *contains the axioms of Arithmetic (e.g. the Peano axioms with the schema for mathematical induction).*

(3b) $\mathfrak{B}$ *contains the following axioms concerning the Self-Belief Operator* B: for every $\varphi, \phi \in \mathfrak{L}$,

$B(\varphi \wedge \phi) \longleftrightarrow B\varphi \wedge B\phi$;

$B\varphi \vee B\phi \rightarrow B(\varphi \vee \phi)$;

$B(\neg\phi) \rightarrow \neg B\phi$.

(In our formal language, $\rightarrow$ and $\longleftrightarrow$ are *defined* in terms of $\vee, \wedge$, and $\neg$ in the usual fashion.)

We shall refer to (3a) and (3b) as Postulate (3). Again, the proof of the main result only requires a special, finite subset of these axioms.

The user of this system is already able to do a great deal of mathematics, as well as socio-economic modelling (once she is allowed the use of a few more constants denoting the beliefs of other agents). We would like to point out that when the decision maker subscribes to an even more powerful language and theory, our main result continues to be valid.

[2] Savage (1954), in laying out the foundations of Baysian decision theory, has already emphasized that logical ominiscience is a very serious idealization about the rational decision maker. In recent times, attempts have been made to construct a positive theory where logical ominiscience is not assumed; see, for example, Lipman (1999).

3 Two Forms of Self-Belief

3.1 Definable Belief-Systems

Let $\mathfrak{L}^\sharp$ be the set of all formulas of the formal language. It contains not only the set of statements $\mathfrak{L}$, but also predicates involving one, two, or any finite number of free variables. For example, $P(x) \doteq$ "x is a prime number" is an element of $\mathfrak{L}^\sharp$. Each of its element is a finite string of symbols, and it is clear that we can construct a one-to-one function $\mathfrak{L}^\sharp \to \mathbb{N}$ (the set of natural numbers). Let $[\varphi]$ be the value of this function at φ, and call it the *code* of the formula φ. Furthermore, we can do so in such a way that all the syntactical and logical relations between formulas are arithmetically defined relations between their codes. (E.g. $[\varphi \wedge \phi] = f([\varphi], [\phi])$, where f is some primitive recursive function.) We omit the technical details of the construction. For Arithmetic without the B operator, a well-known coding function is the "Gödel Code"; and it can be extended to our formal language. Let us fix, once and for all, such a coding function. For simplicity, we continue to call it the Gödel Code.

Having fixed a coding, $\mathfrak{L}^\sharp, \mathfrak{L}$, and $\mathfrak{B}$ can all be regarded as subsets of $\mathbb{N}$. A subset $\mathfrak{A} \subseteq \mathbb{N}$ is said to be *definable* if there is an arithmetic formula in one free variable $A(x)$, such that $\mathfrak{A} = \{x \in \mathbb{N} \mid A(x)\}$.

If the decision maker's belief-system is definable we would have:

(4–) *There is an arithmetic formula in one free variable $A(x)$, such that for every statement $\varphi \in \mathfrak{L}$, $\varphi \in \mathfrak{B}$ if and only if $A([\varphi])$.*

One form of self-knowledge is the ability to define one's own beliefs. At least, the decision maker may believe that she possesses such self-knowledge. Accordingly, we impose the following postulate.

(4+) *There is an arithmetic formula in one free variable $A(x)$, such that $\mathfrak{B}$ contains the following proposition concerning the Self-Belief Operator B: "For every statement $\varphi \in \mathfrak{L}$, $B\varphi$ if and only if $A([\varphi])$".*

This postulate formalizes the first form of self-belief explained in the Introduction.

3.2 Positive Introspection

We now turn to the second form of self-belief, namely, the Principle of Positive Introspection. As explained in the Introduction, it states that the decision maker is fully aware of her own beliefs, and indeed explicitly believe in the Principle of Positive Introspection

itself. Formally, this amounts to the following two postulates:

(5–) *For every statement* $\varphi \in \mathfrak{L}$, *if* $\varphi \in \mathfrak{B}$ *then* $B\varphi \in \mathfrak{B}$.

(5+) *For every statement* $\varphi \in \mathfrak{L}$, $\mathfrak{B}$ *contains the following proposition: "*$B\varphi \rightarrow BB\varphi$*".*

We shall refer to (5–) and (5+) as Postulate (5).

Note that we have not assumed that the decision-maker's beliefs should be true (i.e. $B\varphi \rightarrow \varphi$), nor that she should be aware of her ignorance (i.e. the Principle of Negative Introspection $BB\varphi \rightarrow B\varphi$).

4 The Dilemma of Self-Belief

We can finally state the main observation of this paper.

Theorem. *The postulates (1),(2),(3),(4+), and (5) are inconsistent. In other words, no subset* $\mathfrak{B} \subseteq \mathfrak{L}$ *can satisfy all these postulates.*

The proof of the Theorem is based on the Gödel Fixed-Point Lemma.

Lemma. *Let* $A(x)$ *be an arbitrary arithmetic predicate in one free variable. Then there is a statement in arithmetic,* θ*, such that from the axioms of arithmetic one can deduce* $\theta \longleftrightarrow \ulcorner A([\theta])$.

Proof. For every number x, let $\langle x \rangle$ be the predicate whose Gödel code is x. Then $\langle x \rangle(x)$ is a statement, and its Gödel code is $[\langle x \rangle(x)]$.

Let $P(x)$ be the predicate $\ulcorner A([\langle x \rangle(x)])$, and let p be the Gödel code of P. Finally, let θ be the statement $P(p)$.

We have $\theta \longleftrightarrow P(p) \longleftrightarrow \ulcorner A([\langle p \rangle(p)]) \longleftrightarrow \ulcorner A([P(p)]) \longleftrightarrow \ulcorner A([\theta])$, as desired. Clearly, the constructions and deductions can all be formalized in arithmetic. (The exact arithmetic formalization of course requires a great deal more work. We refer the reader to Gödel's original work, or any standard text in mathematical logic. However, we would emphasize that to construct θ and prove the lemma for any one predicate $A(x)$,only a finite set of axioms is involved.) ■

Apply this lemma to the predicate A which appears in postulate (4+). Intuitively, the statement θ says that "I do not believe in this statement". Here, the demonstrative adjective 'this' refers to the very statement θ itself. The presence of such a self-referential statement is the cause of the dilemma of self-belief. Under our postulates, the decision maker is logically powerful enough to realize that she cannot believe such a statement,

but then positive introspection forces her to believe that she doesn't believe it; and that leads to a contradiction. We can now give the formal argument.

Proof of the Theorem. Suppose, contrary to the conclusion, the subset $\mathfrak{B} \subseteq \mathfrak{L}$ satisfies all these postulates. By the Gödel Fixed-Point Lemma, there is a statement θ such that $\theta \longleftrightarrow \neg A([\theta])$.

By postulates (2) and (3), the decision maker is able to prove this lemma for herself, and therefore "$\theta \longleftrightarrow \neg A([\theta])$" $\in \mathfrak{B}$. Hence, by (2) and (4+), we also have "$\theta \longleftrightarrow \neg B\theta$" $\in \mathfrak{B}$.

For any statement φ, $\varphi \vee \neg\varphi$ is a logical tautology; and so by (2),"$\varphi \vee \neg\varphi$" $\in \mathfrak{B}$. Setting $\varphi = B\theta$, we have "$B\theta \vee \neg B\theta$" $\in \mathfrak{B}$.

We now show that the decision maker can logically eliminate the possibility $B\theta$. Since "$\theta \longleftrightarrow \neg B\theta$" $\in \mathfrak{B}$, we have "$B\theta \longleftrightarrow B(\neg B\theta)$" $\in \mathfrak{B}$ by (2+) and (3b). Again by (3b) and (2), "$B\theta \rightarrow \neg BB\theta$" $\in \mathfrak{B}$. By (5+), we have"$\neg BB\theta \rightarrow \neg B\theta$" $\in \mathfrak{B}$. Hence, by (2), we have "$B\theta \rightarrow \neg BB\theta \rightarrow \neg B\theta$" $\in \mathfrak{B}$, and so "$B\theta \rightarrow \neg B\theta$" $\in \mathfrak{B}$. Combined with "$B\theta \vee \neg B\theta$" $\in \mathfrak{B}$ and using (2) again, we have $\neg B\theta \in \mathfrak{B}$.

However, we also have "$\theta \longleftrightarrow \neg B\theta$" $\in \mathfrak{B}$. By (2), $\theta \in \mathfrak{B}$. By (5), $B\theta \in \mathfrak{B}$.

We have shown that $B\theta \in \mathfrak{B}$ and $\neg B\theta \in \mathfrak{B}$. Thus the decision maker's beliefs are inconsistent, contrary to postulate (1). The theorem is proved. ■

References

[1] Anderlini, L. (1990): 'Some Notes on Church's Thesis and the Theory of Games,' *Theory and Decision, 29, 19-52.*

[2] Aumann, R. (1976): 'Agreeing to Disagree,' *Annals of Statistics*, 4, 1236-1239.

[3] Bacharach, M. (1987): 'A Theory of Rational Decision in Games,' *Erkenntnis*, 27, 17-55.

[4] Binmore, K. (1987): 'Modeling Rational Players, I & II,' *Economics and Philosophy*, 3 and 4: 179-214 and 9-55.

[5] Binmore, K. and H. S. Shin (1992): 'Algorithmic Knowledge,' in 'Knowledge, Belief, and Strategic Interaction' (C. Bicchieri and M.-L. Dalla Chiara, eds.), 141-154, *Cambridge University Press,* New York.

[6] Brandenburger, A. and H. J. Keisler (1999): 'An Impossibility Theorem on Beliefs in Games,' *Mimeo,* December 19, 1999.

[7] Canning, D. (1992): 'Rationality, Computability, and Nash Equilibrium,' *Econometrica*, 60(4), 877-888.

[8] Enderton, H. B. (1972): 'A Mathematical Introduction to Logic,' *Academic Press*, New York.

[9] Geanakoplos, J. (1989): 'Game Theory without Partitions, and Applications to Speculation and Consensus,' *Yale Cowles Foundation Discussion Paper, 914*, 1-45.

[10] Geanakoplos, J. (1992): 'Common Knowledge,' *Journal of Economic Perspectives, 6(4)*, 53-82.

[11] Lipman, B. (1999): 'Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality,' Review of Economic Studies, 66(2), 339-361.

[12] Mertens, J.-F. and S. Zamir (1985): 'Formulation of Bayesian Analysis for Games with Incomplete Information,' *International Journal of Game Theory*, 14, 1-29.

[13] Modica, S. and Rustichini, A. (1999): 'Unawareness and Partitional Information Structures,' *Games and Economic Behavior*, 27(2), 265-298.

[14] Rubinstein, A. (1998): 'Modeling Bounded Rationality,' *MIT Press,* Cambridge, Massachusetts and London.

[15] Savage, L. J. (1954): "Foundations of Statistics," *Wiley*, New York.

[16] Simon, H. A. (1976): 'From Substantive to Procedural Rationality,' in 'Method and Appraisal in Economics' (S. J. Latis, ed.), 129-148, *Cambridge University Press,* New York.

[17] Simon, H. A. (1979): 'Models of Thought,' *Yale University Press*, New Haven.