An Impossibility Theorem on Self-Belief*

Hsueh-Ling Huynh† and Balázs Szentes‡

May, 2002

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 three forms of self-belief – the ability to define one's own beliefs, belief in one's own correctness, and awareness of one's own beliefs, 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.

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

*This is a revision of an earlier paper entitled "Believing the Unbelievable; the Dilemma of Self-Belief" (November 1999). We are grateful to Ken Binmore, Eddie Dekel, Yossi Feinberg, Jeff Ely, Debraj Ray, Bob Rosenthal, and Aldo Rustichini for their interest and comments in this paper.

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

‡Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60636.

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. However, neither works reached the sharp conclusions reported here.

Our discussion is concerned solely with the belief's of a single decision maker. As we shall see, it already reveals some interesting implications of our interpretation.

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.

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.

The decision maker's ability to define her objectives already entails certain self-beliefs (for example, in knowing her own preference). 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 recall from Bayesian decision theory and game theory the notion of a *type*. The type of a rational decision maker consists of her beliefs (together with her preferences, which may be regarded as a kind of self-belief), relevant to certain, often interactive, situations. Her type will then determine her actions in these situations. But then the type of each decision maker is itself open to uncertainty, and forms the subject of further beliefs. Thus the question arose as to whether it is possible to construct a 'universal type space', whose elements are complete beliefs about the types of all other decision makers. For instance, Merten and Zamir (1985) showed that this can be done if beliefs are interpreted to be certain kinds of probability measures. Therefore it would seem unproblematic to regard a decision maker's whole belief-system of as a single, knowable, object; and a fully rational being is one whose belief-system can be identified with an element of the universal type space.

Under our interpretation of procedural rationality, a complete knowledge of another decision maker's belief-system would mean the ability to *define* the whole belief-system

explicitly in a some language. And since the other is endowed with the same ability, this implies the ability to define one's own belief-system. In this paper, we explore the consequences of this line of thought.

We prove an impossibility theorem, which shows that this ability for self-definition is incompatible with other natural requirements on self-beliefs. These other requirements stem from more psychological considerations. One such requirement is that one believes that one's own beliefs are correct. Indeed some may argue that this is an integral part of the meaning of the word 'believe'. Another is that one should be aware of one's own beliefs, which says: if the decision maker believes X, then she also believes the statement "I believe X". This two postulates are 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, Geanakoplos (1989, 1992) and Modica and Rustichini (1999).

We shall give precise formulation of these 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. But we include a sketch of Gödel's fixed-point argument, not only to render the paper more self-contained, but because it is illuminating.

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 $\wedge$. (These operations are to be interpreted as 'not', 'or' and 'and' respectively, and the Boolean implications $\rightarrow$ and $\leftrightarrow$ are defined in terms of them in the usual way.) 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 $\mathfrak{B}$.

(A) $\mathfrak{B}$ *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 $\neg\varphi$.
(B) $\mathfrak{B}$ *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 $\varphi \in \mathfrak{B}$. In particular, $\mathfrak{B}$ contains all the logical tautologies.

Postulate (B) says that the decision maker is able to make logical inferences, and believes in these inferences.

Indeed, it assumes that she 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, and no real-life decision maker has this ability at any given moment in time. Yet 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.[1] On the one hand, we shall see that even with this idealization the decision-maker's self-beliefs cannot be complete. On the other hand, in the proof of our main result we only require the decision maker to make a handful of deductions; logical omniscience is postulated here only because it appears more natural than the specific finite collection of logical rules and axioms needed for the proof.
(C) $\mathfrak{B}$ *contains the axioms of Arithmetic (e.g. the Peano axioms with the schema for mathematical induction).*

Postulate (C) furnishes the decision maker with the basic use of mathematical reasoning, as well as some 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 if the decision maker subscribes to a more powerful language or theory, our result continues to be valid.

3 Three Forms of Self-Belief

The first form of self-belief is that the decision maker should believe in the correctness of her own beliefs. Thus, logically at least, she is free from self-doubt. Formally, this can be stated as the following postulate:

[1] Savage (1954), in laying out the foundations of Bayesian 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).

(D) *For every sentence* φ, *the sentence* $B\varphi \rightarrow \varphi$ *is in* $\mathfrak{B}$.

Compare this with the better-known discussions of 'knowledge operator' or 'information correspondence'. There, if an event E is known to occur, then it must occur (i.e. $KE \rightarrow E$). Here we merely assume that the decision-maker believes this to be the case.

The second form of self-belief is that the decision maker should be aware of her own beliefs, sometimes called the 'Principle of Positive Introspection'. Formally, this amounts to the following postulate:

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

Note that by postulating (D) and (E) we have not assumed that the decision-maker should be aware of her own ignorance, or the 'Principle of Negative Introspection', which in our context would say that if $\varphi \notin \mathfrak{B}$ then $\neg B\varphi \in \mathfrak{B}$. [2]

We now turn to the third form of self-belief, namely, the ability to define one's belief. Before stating this postulate we must formalize the meaning of definability. 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 $\mathfrak{L}^\sharp \rightarrow \mathbb{N}$ (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 shall 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 $\mathfrak{A} = \{x \in \mathbb{N} \mid A(x)\}$.

One form of self-knowledge is the abibilty to define one's own beliefs. This can be formalized by the following postulate.

(F) *There is an arithmetic predicate in one free variable* $A(x)$ *such that, for every state-*

[2] The Principle of Negative Introspection is well-known to be quite restrictive, and the removal of this principle leads to the information systems studied in Geanakoplos (1989, 1992) and Modica and Rustichini (1999).

ment φ, $\mathfrak{B}$ *contains the following proposition "*$B\varphi \longleftrightarrow A([\varphi])$*".*

Note that we do not assume that $\mathfrak{B}$ is actually definable by the predicate A (or by any predicate at all), merely that the decision maker believes that her own belief-system can be thus defined.

4 The Dilemma of Self-Belief

Theorem. *No belief-system* $\mathfrak{B} \subseteq \mathfrak{L}$ *can satisfy all the following postulates:*

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

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

(C) $\mathfrak{B}$ *contains the axioms of Arithmetic.*

(D) *[No Self-Doubt] For every sentence* φ*, the sentence* $B\varphi \rightarrow \varphi$ *is in* $\mathfrak{B}$*.*

(E) *[Self-Awareness] If* $\varphi \in \mathfrak{B}$*, then* $B\varphi \in \mathfrak{B}$*.*

(F) *[Self-Definition] There is an arithmetic predicate in one free variable* $A(x)$ *such that* $B\varphi \longleftrightarrow A([\varphi])$ *is in* $\mathfrak{B}$ *for every* φ*.*

While the Theorem is proved by a rather straightforward application of ideas from Gödel's Incompleteness Theorem, it is not proved by direct appeal to the latter theorem. In the absence of any one of the last five postulates, there will be a $\mathfrak{B}$ satisfying them, and $\mathfrak{B}$ would be incomplete. (Namely, the decision maker is allowed to be agnostic about certain propositions φ, where neither φ nor $\neg\varphi$ are in $\mathfrak{B}$.)

Instead, we need the following lemma.

Lemma [Gödel's Fixed-Point 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* $\theta \longleftrightarrow \neg A([\theta])$*.*

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

We have $\theta \longleftrightarrow P(p) \longleftrightarrow \neg A([\langle p \rangle(p)]) \longleftrightarrow \neg A([P(p)]) \longleftrightarrow \neg A([\theta])$, 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 (F). 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.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 $\theta \longleftrightarrow \neg A([\theta])$. Then the following statements are in $\mathfrak{B}$.

1) $\theta \leftrightarrow \neg A([\theta])$, by (B), (C) and the lemma (since the lemma is deducible from arithmetic).

2) $B\theta \leftrightarrow A([\theta])$, by (F).

3) $\theta \leftrightarrow \neg B\theta$, by (1), (2) and (B).

4) $B\theta \rightarrow \theta$, by (D).

5) $B\theta \rightarrow \neg B\theta$, by (4), (3) and (B).

6) $B\theta \vee \neg B\theta$, by (B) since $B\theta \vee \neg B\theta$ is a tautology.

7) $\neg B\theta$, by (5), (6) and (B).

8) θ, by (3), (7) and (B).

9) $B\theta$, by (8) and (E).

Now (7) and (9) shows that $\mathfrak{B}$ is inconsistent, contradicting (A). ■

5 Further Discussion

To summarize, 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 seems to be the subject of a recent paper of Brandenburger and Keisler (1999), which makes use of the same notion of definability as here.

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* and takes an *optimal* action at all times.[3] Again, their results show that the notion of a

[3] A clear discussion and illustration of these ideas can also be found in Rubinstein (1998).

universal type is untenable, if it is interpreted to mean an effective algorithm which makes an optimal decision under all circumstances. Notably, their results appeal to *substantive,* and not only procedural, rationality.

By contrast, the present investigation is based on the more general notion of *definability*, and focuses almost entirely on the implications of *procedural rationality.* Our impossibility theorem is valid even when the decision maker is endowed with powers of deduction that go beyond what Binmore and Shin would call 'algorithmic knowledge'. As for the latter aspect, we only assume that a procedurally rational decision maker believes that she has an explicit and complete definition of her own beliefs, and possesses a sufficiently sophisticated linguistic and logical capability to do so. Besides the postulate of self-awareness, we do not assume the decision maker's beliefs to be true, nor her decisions to be optimal.

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] Gödel, K. (1930?): 'On the Formal Incompleteness of ???', ???

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

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

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

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

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

[17] 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.

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