Discussion Paper No. 943

RATIONALITY AND ASCRIPTIVE SCIENCE*

by

Itzhak Gilboa**

June 1991

---

*I am very grateful to many people for the discussions that motivated this work, as well as for comments and references. At the risk of leaving some names out, I would like to mention Robert Aumann, Elchanan Ben-Porath, Cristina Bicchieri, Andy Fano, Tim Fuerst, Joe Halpern, Ehud Kalai, Andrew Ortony, Jim Peck, Phil Reny, Aldo Rustichini, David Schmeidler, and, especially, Eva Gilboa.

**Department of Managerial Economics and Decision Sciences, J.L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208.

Abstract

This paper suggests definitions for two closely related terms which are (or could be) used in the social sciences. First, "rationality" is defined as a behavior which will not be altered as a result of awareness to its analysis. Next, an "ascriptive theory" is defined to be a descriptive theory which may become common knowledge among its subjects, yet remain valid.

The relation between these concepts--as well as between them and others--is studied, and an "impossibility theorem," due to Dostoyevsky, is discussed.

1. Apologetics

A definition, which may be thought of as an association of a word to a concept, may be of interest in and of itself in two cases: either that the word exists out there, in our minds and on our lips, but it is not associated with one definite concept, or in the case that the concept is clearly around but it lacks a specific word to help us completely separate it--in our thought and speech--from its neighbors, siblings and second-order cousins.

This paper suggests two definitions, one of each category, hopefully. (The alternatives are that both the word and their concept are known, together with their association, or that neither is very important. In either case, the definitions are of little interest.)

The first definition is an attempt to find a single (simple) concept which would capture (almost) all that is meant by the word "rationality." The second one is a definition of the term "ascriptive science," which should fit an old concept in search of a new name.

In attaching a concept to a word, one is typically making a claim about the "real world," i.e., that in most of the cases this word is used, that concept is actually meant. Attaching a word to a concept, on the other hand, is somewhat more modest an endeavor. Although there is an implicit suggestion that this concept is somehow salient, hence, merits a new term, this claim is weaker and fuzzier than in the previous case.

However, both types of definitions may simply be viewed as a proposed addendum to a vocabulary which is judged based on its usefulness in facilitating discourse. It is thus that the writer of these lines wishes these definitions to be viewed.

Section 2 proposes a definition of "rationality" which hopefully encompasses most of the ways this word is used, and highlights its inherently subjective nature. In Sections 3 and 4 some general terms are discussed, mostly to verify that no misunderstandings will follow. Section 3 discusses the terms "science" and "philosophy," claiming that the latter is a special case of the former, and makes a side comment on the notion of "observability" in philosophy and in economics. Section 4 discusses the distinction between descriptive and prescriptive science, and points out the ambiguity of the latter.

In Section 5 we finally get to the definition of ascriptive science and study its relation to the notions of "descriptive" and "prescriptive." Section 6 is devoted to the interrelation between ascriptive science and rationality. The cases in which the two are most distinctly separate turn out to be related to theory-dependent preferences. Section 7 quotes and discusses an "impossibility theorem" (or paradox) due to Dostoyevsky, which arises naturally in these cases. Finally, Section 8 concludes with a short comment on the models discussed and the notion of free will.

This paper is written in an informal, verbal way. In a few places some formal objects are mentioned, but this is only for notational convenience. The reason is that the few conclusions mentioned throughout the text do not warrant the complicated formal definitions one would need in order to formalize them. However, it is the author's hope that clarity is gained, rather than lost, by following this path. In particular, the verbal formulations should be clear enough to make the mathematical formalization obvious.

2. Rationality

The term "rationality" is used in such a variety of contexts, explicit meanings and implicit interpretations, that one can hardly hope to obtain a widely accepted definition of what types of behavior (say) are "rational." However, it is somewhat troubling that we do not even seem to have an understanding of what it is that one means by dubbing a certain behavior pattern as rational. This section is an attempt to clarify this point.

In the everyday usage of the word, "rationality" is often applied to general, objective actions. It is typically considered irrational, say, to use your grandfather's abacus rather than a pocket calculator, or to avoid borrowing money (when it makes sense financially) just because one does not like being a debtor.

Modern decision (and, to an extent, economic) theory, on the other hand, did not make any such presuppositions on preferences and, at least in principle, is supposed to accommodate various psychological and sociological motives. In other words, it allows for subjectivity in preferences as well as in beliefs. However, subjective utilities and priors (as in von-Neumann-Morgenstern (1944) and Savage (1954)) are derived from appropriate axioms, which were suggested as "canons of rationality." In this approach, no particular action (choice) can be dismissed as irrational; it is only a pattern of choices which may be tested for rationality, where the latter means some notion of consistency.

It would seem, however, that the appropriate "notion of consistency" is not universally agreed upon, especially in view of various alternative models suggested in the last decade. While some people insist that rational agents must play Nash equilibrium (Nash (1951)) in a one-shot game, others would

suggest normative models allowing for intransitivity of preferences. Classical axioms of rationality that have been studied in weakened versions range from Savage's sure-thing principle to the set of axioms known as "S5" on knowledge operators, from common knowledge of the model to "Bayesian rationality." Some of these models are merely trying to better describe reality, but some are claiming to be "theories of rationality" just as their classical ancestors.

Thus, one is led to the conclusion that which axioms are accepted as defining "rationality" is also a matter of taste. To a large extent, a game- or decision-theorist may be viewed as making a choice, as expressing preferences, among various axioms, and possibly also suggesting the selected ones to the theory's subjects.

This view suggests a definition of rationality which is inherently subjective: a certain mode of behavior of an intelligent decision maker (DM) is <u>rational</u> if, when the DM is confronted with its analysis, he/she does not feel uneasy, i.e., does not want to change it.

For instance, when people exhibit cyclical strict preferences, and are made <u>aware</u> of this, most of them feel there is something wrong with their decisions. For those people, transitivity (of strict preference) may be an axiom of rationality. It seems safe to conjecture that a smaller subset of decision makers will be upset by (the awareness of) their violation of the sure thing principle, for example.

Rationality is therefore closely related to introspection. Correspondingly, the term "intelligent" used in this definition should be interpreted as "capable of introspection." Thus, one would not ascribe rationality to bees (or genes) even if their behavior happens to conform with

some axiom or other (say, playing a Nash equilibrium). Since we do not have any evidence (not even of the type called "intuition") about bees' introspective abilities and their feelings once confronted with the erudite analysis of their choices, they do not qualify as "rational."

Two questions arise here, which are (and should be) almost automatically asked about a definition of this nature. First, is this definition cyclical? Second, is this distinction "metaphysical nonsense"?

As for the first question: Does this definition amount to saying that "rational" is whatever is called rational? Obviously not. The decision maker is not asked, according to this definition, whether the analyzed behavior is "rational"--rather, whether he/she would like to change it. Indeed, the DM need not know what the word "rationality" means (which is very fortunate since no one does).

The second question bears on a more fundamental issue: How can one tell whether bees are introspective or not? Moreover, how can one tell whether people would like to change their behavior? Can it be applied to decisions that cannot be reversed (and are there any others at all)? And, finally (one should hope): Is it not the case that bees are rational according to this definition after all? One can expose them to the most recent advanced research in biology and evolutionary game theory alike and, lo and behold, the bees do not change their behavior, which is supposedly the test for rationality.

All these questions stem from the identification of "observable" with "actual choice situation," the latter understood in a narrow sense. We will come back to this point below; at this stage let us just agree that what people say <u>is</u> observable, and can be meaningfully interpreted just as "actual"

choices. Indeed, choosing what speech act to perform is actually a choice (this time interpreted in a wider sense).

Thus, one can provide a reasonable test for a DM's introspectability and tell bees from humans. (Obviously, this test goes only as far as our language reaches, and what "really" occurs in a bee's mind is bee-yond us. But for this reason precisely it is also besides the point and, indeed, ill-defined. Understanding "introspective" as "capable of communicating seemingly-introspective thoughts" should resolve the issue.)

Similarly, since people are capable of expressing regret and may bang their head against the wall in a very observable manner, there should not be a problem with applying this definition of rationality in retrospect.

Assuming that the definition above is accepted as non-vacuous and meaningful, we still need to address a few issues relating to "being confronted with the analysis" of one's decisions. First, it should be made clear that this "analysis" must not contain any new information which was not available to the DM while making the decision. Betting on the wrong horse should not be considered "irrational" simply because in hindsight one would have preferred betting on the winner. The "analysis" of the decision should not involve anything that is not deducible from the DM's knowledge.

The second issue has to do with computation complexity (and cost). Consider a chess player making a move, and then being confronted with a complete computer-generated analysis of all plays of the game in the next 15 moves, realizing the original choice was not optimal. Was it also irrational? Strictly speaking, it was, as the analysis is deducible from available information. On the other hand, one may want to consider only reasonable (say, polynomial) deductions, and classify a complete analysis as described

above as "new information," since its computational complexity makes it practically logically independent from its logical progenitor.

The situation may be further complicated if the extent to which analysis is performed prior to a decision becomes a (quantitative) decision variable with, say, an increasing cost. It would then seem perfectly "rational" for one to take decisions without completing all involved computations, at least if one has bounds or priors on the resulting computations.

Deciding what kind of information should be considered in the deductive closure of one's knowledge is, to a large extent, a matter of taste. But even those who would like to assume restricted computational abilities (and costly acquisition of deductions as if they were new information) will probably agree that on some level (see Lipman (1989)) a "rational" decision should be made in awareness of all possibilities (for example, the possibility to "buy" additional deductions). And this is the crux of the matter. Regardless of what one chooses to define as "knowledge," "deduction," etc., rationality is inherently tied to awareness and introspection.

It is interesting to note that, although the main goal of this section is to provide a definition of "rationality" for scientific discourse, it appears that the definition given above may also encompass the everyday, "naive," usage of the word. When someone says, for instance, that it is irrational to use your grandfather's abacus, one typically means that if one thinks about it, i.e., should you be confronted with the analysis of this behavior, you will probably feel uneasy about it. Thus, the use of the term "irrational" in the sense of "losing one's head" is also compatible with the general definition.

Finally, a few words on representation are in place. Consider, for

instance, a decision maker who chooses different alternatives when the decision problem is represented differently (as in Tversky-Kahneman (1986)). Such a DM may be completely aware of his/her choices, but fail to observe they are, in a sense, inconsistent, because he/she is not aware of different representations of these alternatives. Similarly, people may exhibit cyclical preferences, each of which is known to them, but the cyclicity will not be obvious unless abstractly represented in a few mathematical symbols. Worse still, people may be aware of the cyclicity but not be bothered by it in one representation while they are in another.

It seems relatively safe to interpret a rational decision as one which is robust (in the sense discussed above) with respect to all various representations of the analysis. (Or, if you will, with respect to all theories compatible with the evidence.) This does not necessarily mean that rational decisions should be independent of the representation in the actual choice situation. For instance, a surgeon may readily admit, without a shred of inconvenience, that his/her operation decisions may change if information is provided in terms of probability of survival or of death. For such a doctor, representation-dependent decision may be rational. The only requirement (according to this definition) on rationality is that the representation of the analysis would not matter.

This definition seems attractive since it does not allow for different scientists, analyzing the same DM, to reach different conclusions regarding his/her rationality, but it puts no restrictions on the type of actual choices which may be "rational."

3. Science, Philosophy and Economics

In order to avoid misunderstandings, let us clarify what is meant by the term "science." For the sake of this discussion (and some others, though presumably not all) it will be useful to define "science" as the activity of constructing mathematical models for some extra-mathematical (preferably "real-world") phenomenon. The term "mathematical models" should be understood broadly, to include all forms of logical reasoning that may be mathematically formalized, even if the resulting theorems are trivial.

Thus interpreted, "science" encompasses not only physics and economics, but also psychology and history, and even philosophy. While scientific disciplines are defined by their subject matter, i.e., the reality they attempt to model, it is not always a simple task to come up with a clear-cut definition of this "reality." Indeed, it is often quite challenging to find the common denominator of all the topics dealt with in a given "discipline," and to draw the lines between it and others.

Philosophy is probably one of the least well-defined disciplines (and also one of the least disciplined). In bold strokes one may attempt to define its subject matter as human thought. Ethics and aesthetics, science and religion, language and logic can all be considered as various aspects of human thinking, where the latter involves reasoning, judgment, and so forth. Indeed, it is not surprising, according to this definition, that philosophy often overlaps psychology, economics and the social sciences in general, to the extent that they bear on people's values and opinions. Similarly, philosophy intersects mathematics, which is not a science, when mathematically modeling mathematical activities, i.e., in logic. The fact that these intersections exist should not be any more intriguing than the fuzzy borderlines between physics and chemistry, psychology and sociology, and so

forth.

The question of whether philosophy is a science or not is, as many of the issues in this paper, ultimately a matter of definition. Apart from purely aesthetic preferences on definitions, they may also be judged on grounds of their usefulness, and from this perspective it would seem that philosophy should better be counted as science. In particular, this raises the issue of applicability of philosophy-of-science theories to philosophical activities and, indeed, to themselves.

However, there is another benefit in unifying our way of thinking about sciences and philosophy. As mentioned above, one of the cornerstones of modern economic theory, as well as of decision and game theory, is the issue of actual observable choices. The notion of a cardinal utility function, for instance, is rejected unless supported by observable choices as in von-Neumann and Morgenstern (1944). In particular, a unique function which represents not only preferences but also a binary relation over pairs of alternatives (interpreted as "x is preferred to y more than z is to w") is usually considered theoretically flawed, since the second binary relation is not observable from actual choices. Similarly, there is a tendency to accept as "economic evidence" only results of experiments in which choices involving "real money" were observed, as opposed to answers to hypothetical questions.

This approach, which probably has its roots in logical positivism, is undoubtedly very helpful inasmuch as it relates theoretical concepts to observable ones. Indeed, a lot of time, energy and ink may be wasted on arguments between proponents of observationally equivalent models. But it sometimes appears to be the case that empirical evidence which is not "real-life actual choices" is completely ignored in principle, thereby needlessly

restricting the scope of economic theories.

Consider, for instance, the assumption that players' utility functions in a game are common knowledge. These functions, typically von-Neumann-Morgenstern utilities, are supposed to reflect the players' preference relations over lotteries, and can only be derived from observation of past choices between pairs of lotteries over the same outcomes, combined with the assumption that these preferences have not changed. However, in many situations, the outcomes are so novel that they cannot be identified with any old ones, choices among which have actually been observed. In principle, then, utilities cannot be assumed known, not even approximately. Yet in many situations of this nature, the players may have a very clear intuition regarding their preferences, i.e., their hypothetical choices, and they may be willing to share this information. Decision and game theory models can still be applicable and insightful if we accept this evidence as legitimate "observation." Of course, the players may have an incentive to lie and, as always, some additional assumptions are needed to make the models useful. But the point is that situations in which agents say different things are <u>not</u> observationally equivalent.

This does not mean that economics should relinquish the idea of "actual choices." As in every discipline, an experiment's validity increases with its similarity to the reality one is interested in. Furthermore, there are examples (though these are surprisingly few) in which a hypothetical questionnaire provided vastly different results from an experiment which involved "real money." But this distinction, even if it may be considered qualitative, is not the distinction between meaningful and meaningless data. And, more specifically, in the absence of "real money" evidence, verbal

reports may still be helpful.

Reconsidering the definition of rationality in Section 2 above, which may be viewed as a philosophy-of-science endeavor (trying to explain what theorists mean by the term "rational"), it should not be surprising, nor considered a flaw, that this definition resorted to verbal reports. The science of philosophy typically accepts this type of evidence more readily than does economics.

Let us conclude this section with a note on formal models of scientific activities. The process of model construction may be formulated in several ways. The most widely accepted one in the philosophy of science is by logic--a theory is formally represented as a set of axioms, evidence--as propositions, and the correspondence between them--as a map from theoretical to observable symbols. (For a classical survey, see Suppe (1974).) Alternatively, one can represent theories as Turing machines and evidence as bit strings (see Gilboa (1990b). A similar idea appears in the artificial intelligence literature (see, for example, Jain and Sharma (1990)).

For our purposes, though a complete formal model is not needed, it will sometimes be helpful to bear in mind the "possible worlds" formalization (which can be thought of as a somewhat more abstract description of the logic model): suppose that a model (of reality) is simply a set of states of the world Ω_M, which is to be interpreted as the formal (mathematical) representation of all conceivable states of the world. A theory is a subset $\Omega_T \subseteq \Omega_M$, which may be interpreted as what is predicted and/or recommended to occur (we will discuss this distinction below). Reality, on the other hand, is a separate (and disjoint) set, Ω_R, which does not appear in the scientist model. Finally, together with Ω_M and Ω_T, a scientist typically describes

(informally) the scope of the model and the theory. This extra-mathematical information may be formalized as a set of correspondences $\{\phi_\alpha\}_\alpha$ from Ω_M to Ω_R, each of which specifies one particular application of the theory. (For instance, a game theorist may study "the battle of the sexes," in which Ω_M has four elements, and claim that it captures the interaction between any two married people. This would give rise to one correspondence, ϕ_α for each such pair, where each state-of-the-world ω in Ω_M is mapped to an event $\phi_\alpha(\omega) \subseteq \Omega_R$.)

The discussion in the sequel will be kept mostly informal. However, in the few cases where formalization will be more likely to resolve, rather than cause, confusion, we will refer to this simple model.

4. Descriptive and Prescriptive Science

In the informal definition of science, as well as in the interpretation of the formal model, there was no mention of what is actually meant by the scientific model of reality. This can hardly be considered (in general) implicit because, at least for the social sciences, very different interpretations are suggested for various theories.

Experience has shown that it is very useful to have a vocabulary for such interpretations, as such a vocabulary greatly simplifies discussions, and makes it easier to understand why a proposed theory does not make sense.

The most prominent distinction is, probably, between descriptive and prescriptive theories. A descriptive (or "positive") theory purports to say what the world <u>is</u> like, whereas a prescriptive (or "normative") theory should be understood as saying what it <u>ought to</u> be like.

It is important to emphasize that, in principle, there is nothing in a mathematical model which will help us make this distinction. It is only a

matter of interpretation. Put differently, "descriptive" and "prescriptive" are not attributes of the theory itself, but rather of the act of modeling, of the scientific endeavor, not of its product. Thus, when we say that "x proposed a descriptive (prescriptive) theory," this should be understood as "x proposed a model for reality, and a theory in this model and, furthermore, x thinks that this is what reality is (should be) like," or, "in this model, x engages in descriptive (prescriptive) science."

As an introduction to the next section, it may be helpful to make these terms slightly more precise. With descriptive science there is little room for ambiguity: there is one state of the world, $\omega_0 \in \Omega_R$, which represents "truth" or "the case," and by proposing a descriptive theory $(\Omega_M, \Omega_T, \{\phi_\alpha\}_\alpha)$, a scientist makes the claim that $\omega_0 \in \phi_\alpha(\Omega_T)$ for all α, i.e., that for all suggested applications, the prediction Ω_T will be correct. In other words, claiming that the theory reflects reality may be identified with the event $\cap_\alpha \phi_\alpha(\Omega_T)$, which may or may not obtain.

Prescriptive theories, on the other hand, are not that simple to define: What is actually meant by "ought"? One can think of at least three interpretations.

1. The weakest one would actually allow for every meaning. Any subset of Ω_M (or of Ω_R) may be appealing to the theorist, and will thus qualify for the imperative embodied in a prescriptive theory.

More formally, one could define a formal symbol, say, A, to designate the subset of Ω_R which is considered "acceptable," similar to $\omega_0 \in \Omega_R$ which is considered "the case." Then a prescriptive theory would be understood to claim that $A \subseteq \cap_\alpha \phi_\alpha(\Omega_T)$, just like a descriptive one claims $\omega_0 \in \cap_\alpha \phi_\alpha(\Omega_T)$. The difference is, however, that ω_0 appears elsewhere in our model--as the

definition of reality we actually assume that all the observations obtained by our scientists will be determined by ω_0. The set A has no corresponding assumption (in this interpretation) to imbue it with meaning, and thus can always be defined to equal $\cap_\alpha \phi_\alpha(\Omega_T)$, which actually means that anything may qualify as a prescriptive theory, and that this term means, as Humpty-Dumpty put it, "just what I choose it to mean--neither more, nor less." (See Carrol's Through the Looking Glass, Ch. VI.)

2. The second, more restrictive interpretation would hold that a theory's imperative (especially if it is interpreted as a "moral" one) should be echoed by some ("moral") intuition in the minds of the theory's subjects. For example, interpreting the Shapley value (Shapley (1953)) from this viewpoint, one may argue--and actually test the claim--that (most) people in a certain society find the axioms morally appealing or, at least, that the axioms of the dummy player and interchangeable players seem "fair" or "just" to those people to whom the theory is suggested to apply.

This "intuition" that people are supposed to have may be formalized as a preference order, on axioms or on various courses society may take. However, this preference order should be distinguished from the one represented by utilities, since it is supposed to capture some "moral" intuition which is divorced from "selfish" motives. It can be thought of as the preference of some "super-ego." Alternatively, one may prefer a simpler definition, which directly states that the subset Ω_T (to be precise, $\phi_\alpha(\Omega_T)$ for all α) will be called "fair," "good," or "desirable" by the individuals involved.

3. The third interpretation of "prescriptive" is even stricter, yet conceptually simpler than the second; it would state that, if given the choice, each of the decision makers involved would rather restrict the

possible states of the world to Ω_T. Although this definition is more directly related to "actual choices," it does not always capture what is meant by "prescriptive." Considering the symmetry (or interchangeable players) axiom in the Shapley value again, people may not necessarily want to impose it since they hope to do better otherwise. (Suppose, for instance, that the Shapley value is not in the core.)

Note, however, that both of the last two definitions of "prescriptive" are translatable to a certain claim about reality, which may be falsified by evidence. In a way, then, "prescriptive" is defined by "descriptive," while relating to a more complicated, and typically hypothetical, choice situation.

Undoubtedly, other interpretations are also possible; in the following we will bear in mind this variety of definitions, and try to specify which one is used in case of ambiguity.

Finally, let us also mention that there is, of course, no theoretical reason to classify all theories to "descriptive" or "prescriptive" (or both). Scientists may propose theories without making neither claim about them. However, such theories are of little interest and, luckily, not very common.

5. Ascriptive Science

In classifying theories according to their intended interpretation to "descriptive" and "prescriptive," it seems useful to make yet another distinction.

Some theories of intelligent agents may be a faithful description of reality as long as they are not commonly known among their subjects (in the sense of Lewis (1969) and Aumann (1976)). Consider, for example, a brilliant scientist who developed a remarkable model which perfectly matches past

observations. With such a model, this scientist may be able to predict with astounding accuracy the next stock market crash, car accident, or terrorist act. Naturally, once our scientist makes the theory known to (even if not commonly known among) the agents involved, it is unlikely to retain its precision.

If, however, a descriptive theory is robust to its own publication, we can also a priori ascribe it to its subjects, and their (common) awareness of it should not reduce its accuracy or correctness. Let us call such a theory ascriptive.

The notion of ascriptive theories, though typically not explicitly distinguished from descriptive ones, pervades game theory as well as modern economic theory. Indeed, the concepts of Nash equilibrium, competitive equilibrium (Arrow and Debreu (1954)) and rational expectations (Lucas (1972)) all rely on the intuition behind it (at least as one possible interpretation). However, not all descriptive theories of intelligent agents are necessarily ascriptive. Many of them, including some of the recent developments in learning models, tend to forego ascriptivity for the sake of more precise predictions (see, for instance, "post-ascriptive" models such as Bray (1982) and the variety of game theoretic models with myopic players or evolution-based concepts).

At the risk of redundancy, let us emphasize that, as in the case of "descriptive" and "prescriptive," saying that a theory of "ascriptive" is merely a classification of the intended interpretation. It does not imply that the theory is correct or known by its subjects. It only means that "the proponent of this theory suggests that this is what reality is like and, furthermore, that his/her paper need not be kept a secret in order to

truthfully reflect reality." Once this claim is clear, both of its assertions may be challenged and tested.

While the distinction between descriptive and ascriptive theories is obvious, it may be worthwhile to highlight the differences between ascriptive and prescriptive theories. First, ascriptive theories are supposed to reflect reality. Prescriptive ones, on the other hand, are free of this constraint, and cannot, in principle, be classified to "true" or "false." (We have seen above that some of the possible interpretations can be translated to descriptive theories; however, these are on a different level of testing. That is, a prescriptive theory $(\Omega_M, \Omega_T, \{\phi_\alpha\}_\alpha)$ does not claim that $\omega_0 \in \cap_\alpha \phi_\alpha(\Omega_T)$. At most it may claim that $\omega_0 \in \cap_\alpha B(\phi_\alpha(\Omega_T))$ where B is some operator describing the subjects' "moral intuition," with the interpretation that B(A) is the event in Ω_R where people believe that A is "just" or "desirable.") Furthermore, a prescriptive theory <u>should not</u>, in a sense, reflect reality, as this renders it pointless. There is little point in having a prescriptive theory stating "Thou shalt not kill" if no one has ever/will ever commit homicide. Or, considering a complementary example, even if it were agreed that walking 50 miles on one's head is morally wrong (say, it accelerates the ozone layer disintegration), it would still be pointless to pass such a law, as no one does that anyway.

One could have expected, though, that at least the second part of the ascriptive claim would be shared by prescriptive ones; that is, that at least when the subjects of the theory become aware of it, it <u>begins</u> to be a valid description of reality. But this need not be the case either. Consider, for instance, a one-shot prisoner's dilemma game (see figure 1) in which one may claim that it is morally desirable to cooperate (C). Such a prescriptive

theory would qualify even according to the stringest definition: if each player could choose to a priori restrict the possible plays to the cooperative outcome, each would have done so. Yet this prescriptive theory need not (and should not) make any claim to truthfulness even if it is known to the players. In the same example, noncooperative (NC) behavior on part of both players may be a viable ascriptive theory, without making any claim whatsoever to the way things ought to be.

Figure 1

	C	NC
C	(3,3)	(0,4)
NC	(4,0)	(1,1)

If one understands the "reality" which should be described in a broad enough sense that would include players' knowledge, a theory can be truly ascriptive only if, according to itself, it is common knowledge among its subjects. The question of whether this is possible or not is, to an extent, a matter of taste, and depends on the type of models and axioms one would wish to use. However, for rich enough finitely axiomatizable models using first-order modal logic where the modality is interpreted as a predicate, it was shown that such models do not exist (see Tarski (1956), Montague (1974), Thomason (1982,1986)). Yet once this point is understood we will understand "ascriptive" in a weaker sense, according to which the model is formulated in a higher-level language (in Tarski's hierarchy) than the subjects' knowledge.

One more clarification is called for. In the definition of "ascriptive"

theories we require that the theory will not be any worse description of reality once it is common knowledge among its subjects. But it may also be a better description of reality. For instance, consider a completely symmetric pure-coordination game as in figure 2 and the descriptive theory predicting the play (T,L). Once it is common knowledge between the two players, it may serve as a "focal point" and be a more accurate description of reality than before.

Figure 2

	L	B
T	(1,1)	(0,0)
B	(0,0)	(1,1)

To an extent, letting the theory serve as a communication device among its subjects is similar to the discussion in Section 2 above, where the "analysis" provides more information than could be deduced by the players. It appears that in such cases the borderline between "reality" and "theory" was not drawn in the "right" place, and that the communication device should have been included in the modeled "reality." In the example above, a more aesthetic model would have included the possible signals, and an ascriptive theory may have predicted that, given the signal "(T,L)", (T,L) would indeed be the outcome of the game. Such an ascriptive theory will probably not enhance its own accuracy simply by being known, but then again it may. And, thus, we may be led into an infinite regress, without ever being guaranteed that the theory's being common knowledge is independent of its correctness.

For the sake of this discussion, we may agree that this independence is the appropriate definition of "ascriptive" theory.

6. Rationality and Ascription

Given the definitions suggested above, it is natural to ask how the concepts of rationality and ascriptive science interrelate. More specifically, can "ascriptive theory" simply be defined as "a descriptive theory of rational agents"?

For the case of a single decision maker, the answer is, indeed, in the affirmative. An ascriptive theory is a descriptive one, whose correctness is not affected once it becomes common knowledge among its subjects. In particular, the single subject will not change his/her choice if he/she knows (is aware of) the theory (the analysis of his/her behavior). Conversely, a descriptive theory of a rational decision maker can be known by its subject and retain its accuracy, and since (under the usual assumptions) knowledge of a fact by an agent implies common knowledge of this fact (by the same agent), the theory is also ascriptive.

When more than one decision maker is involved, however, this equivalence no longer holds in general. The subjects of ascriptive theories are still (claimed to be) rational, but not every descriptive theory of rational agents necessarily claims to be ascriptive. This is due to the fact that an agent's choice may depend on other agents' knowledge directly--that is, not only via their choices.

Indeed, if we assume that players' preferences in a game depend on other players' actual choices alone, and if no player would want to change his/her choices when confronted with their analysis--which would have to include the

environment, i.e., the specification of the other players' behavior--then common knowledge of the theory would not undermine its predictive power either. (And, if the theory selects a unique choice for each player, it will have to predict a Nash equilibrium.)

However, if the very fact that others know the theory makes one want to change one's choice (say, because one would like to be surprising), this no longer holds. The next section is devoted to this type of preferences. At this point we will only note that with a single DM the problem does not arise: even if an agent's preferences depend on his/her knowledge (say, if one likes to surprise oneself), the requirement of rationality and of ascription boil down to the same thing.

Finally, it should be noted that for a single DM one may think of ascriptive theories as the intersection between descriptive and prescriptive ones: using a strict definition of "prescriptive theory," as "desirable" in the sense of the DM not wanting to change the behavior it suggests, every descriptive theory is prescriptive if and only if it is ascriptive. However, we have seen before that even with such a narrow definition of "prescriptive" this equivalence does not hold with more than one decision maker.

7. Dostoyevsky's Impossibility Theorem

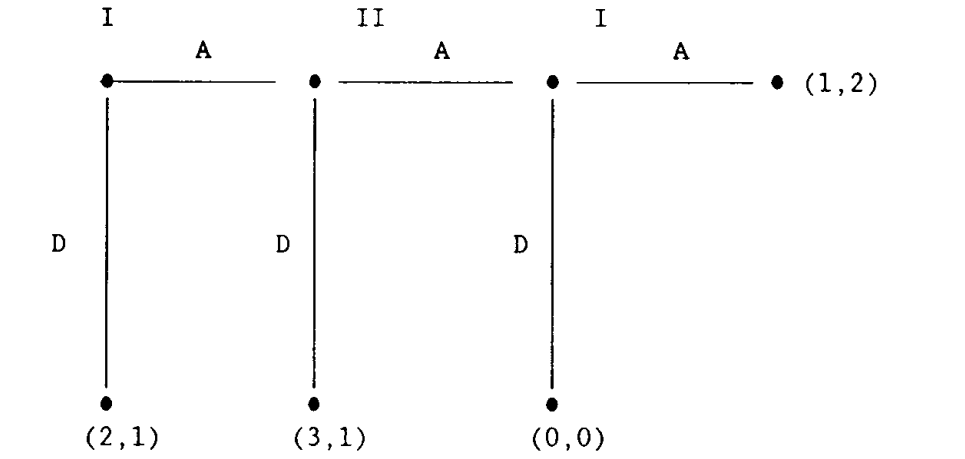
The fact that payoffs may, in general, depend on what other players (or even the DM) know poses some problem for non-vacuous ascriptive theories. In Notes from Underground, we find: "Even in the case he really might turn out to be nothing but a piano key, and even if this were proved to him by the natural sciences and mathematics, man still won't come to his senses, and will do something deliberately contrary, solely out of ingratitude, and to insist

on his own way" (Dostoyevsky, pp. 34-35).

In other words, a non-vacuous ascriptive theory can never be true (irrefutable) if its subjects prefer not to be predicted. Similar issues arise in "information-dependent games" (Gilboa and Schmeidler (1988)) or "psychological games" (Geanakoplos, Pearce and Stacchetti (1989)). Indeed, these models may be used to formalize Dostoyevsky's preferences and the negative conclusion which follows.

The fact that an agent may derive utility from the mere refutation of a theory should be distinguished from a separate issue, which is discussed in game theory recently--namely, the fact that a player may want to cause others not to believe in a theory, in hope of changing their actual choices. Consider the game of figure 3.

Figure 3



It is sometimes argued that, should the players believe in the backward induction solution at the beginning of the game, player I, instead of playing D (as dictated by this solution) will consider playing A. This move would obviously contradict the backward induction solution, and when II is called to play, she may not want to risk playing A, since I has already exhibited violation of the backward induction solution. Thus II may play D, which yields the best possible outcome from I's viewpoint. Hence, the argument continues, it may be "rational" for I to refute the theory saying he is rational.

The discussion of this topic, and of common knowledge of rationality, is beyond the scope of this paper. (See Reny (1988), Bichierri (1988a, 1988b, 1989), Gilboa (1990), and Ben-Porath (1991).) In our context only the following points should be made:

a. every non-vacuous descriptive theory may be refuted;

b. a satisfactory theory of rational choice should (by definition of "satisfactory") compare all alternatives of the DM;

c. it follows that a (satisfactory) ascriptive theory should specify what will happen if the players decide to deviate from it, i.e., it should describe what is the outcome not only for the possible events (complying with it), but also for <u>conceivable</u> ones, which are considered by an agent while making a choice;

d. in particular, the backward induction solution (for finite extensive form games, say, with perfect information and no ties) is such a theory. It predicts that the backward induction solution will apply to all conceivable subgames. Furthermore, as such, it may well be ascriptive, since knowledge of the theory

does not induce any player to violate it. This does not, of course, mean that it is a reasonable or correct theory.

Let us emphasize the difference between this discussion and Dostoyevsky's "impossibility theorem": in the game of figure 3, refutation of a theory is designed to bring about a change in others' behaviors. Thus, its desirability for an agent depends on his/her beliefs regarding the effect of this manipulation. In Dostoyevsky's case, by contrast, the very fact the theory is refuted makes the DM better off. Therefore, with preferences as in regular (knowledge-independent) games there are ascriptive theories that could, at least in principle, be true, whereas with knowledge-dependent games one can find preferences for which no non-vacuous ascriptive theory could possibly hold.

Finally, let us compare this phenomenon with simple nonexistence of a self-enforcing solution. Consider the "matching pennies" game (shown in Figure 4) which is played precisely once. Even if we accept the fact that utilities (and the game) are common knowledge, a strict interpretation of a "one-shot game" devoids mixed strategies from observational content. Mixed strategies may be interpreted as subjective probabilities of other players regarding one's choice, which may be derived from hypothetical choices they would make on a "Savage questionnaire." But, in a truly one-shot situation, they do not pass the falsifiability test (see, for instance, Popper (1934)), nor will a totally mixed strategy ever maximize the likelihood function after a single observation.

Figure 4

	L	R
T	(1,-1)	(-1,1)
B	(-1,1)	(1,-1)

Thus, a testable theory would have to restrict itself to predict a certain subset of the four states of the world and, together with the assumption that the players' choices are independently taken, to a set of pure strategies for each player. Obviously, for the "matching pennies" game no non-vacuous ascriptive theory may hold true (and be robust to its publication).

However, for many games there will be dominated strategies, or Nash equilibria in pure strategies, or other features which will allow for ascriptive theories to say something which is neither trivial nor false. The point with Dostoyevsky's "paradox" is that for these preferences, in no choice situation will there be such theories.

8. A Comment on Free Will

The fact that an ascriptive theory may be known by its agents, combined with the fact that in order for it to be true the agents should not deviate from it, sometimes raises the question of determinism versus free will. How come, it is sometimes asked, that the agents know their own behavior and still

are claimed to make a free choice? If their choice is truly free, it cannot be a priori known by anyone, including themselves.

A similar point was raised and dealt with in Aumann (1987). The answer in our case is basically the same: for each agent we should consider his/her decision problem separately, assuming the agent knows all that the theory specifies about others (including their knowledge--to be considered as belief--about the agent) and nothing about oneself. An outside observer may now test the theory's prediction about the individual agent with the agent's actual choice. Should the two be compatible, we take the next step and reveal the theory to the agent.

If the agent wants to change his/her choice, it was not a rational choice for him/her by definition. Rationality, as well as ascriptive science, deals with situations in which this does not occur. But the very fact that some theories are not ascriptive shows that the notion does not contradict free choice: it is precisely free choice which is used to test rationality/ascriptivity.

Obviously, one can cast Dostoyevsky's "impossibility theorem" into a single DM problem. Then, by definition, no behavior will be rational for a DM who wants to surprise oneself. If this decision maker is introspective enough, we may arrive at a paradox. However, this paradox does not show that such a DM has no free will. To the contrary: it is free will which is needed in order to arrive at the paradox.

Whereas, for this kind of preferences, no consistent theory of choice seems to be possible, for less problematic (and surely for knowledge-independent) preferences no problem arises. Hence, free will may be perfectly compatible with the theory, though it requires a specification of (each

agent's beliefs on) the outcome of every choice, including those excluded by the theory.

References

Arrow, K. J. and G. Debreu (1954), "Existence of Equilibrium for a Competitive Economy," Econometrica, 22.

Aumann, R. J. (1976), "Agreeing to Disagree," Annals of Statistics, 4, 1236-1239.

Aumann, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," Econometrica, 55, 1-18.

Ben-Porath, E. (1991), "Common Belief in Rationality in Extensive Form Games," mimeo.

Bicchieri, E. (1988a), "Strategic Behavior and Counterfactuals," Synthese, 76, 135-169.

Bicchieri, C. (1988b), "Common Knowledge and Backward Induction: A Solution to the Paradox," in the Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge, M. Vardi (ed.), Morgan-Kaufmann, 381-393.

Bicchieri, C. (1989), "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," Erkenntnis, 30, 69-85.

Bray, M. (1982), "Learning, Estimation and the Stability of Rational Expectations," Journal of Economic Theory, 26, 318-339.

Carrol, L., Through the Looking Glass, in The Works of Lewis Carrol, E. Guiliano, ed., Crown Publishers, 1982.

Dostoyevsky, F. M., Notes From Underground (originally published in 1864). Quoted from the translation of Mirra Ginsburg, Bantam Books, 1974.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989), "Psychological Games and Sequential Rationality," Games and Economic Behavior, 1, 60-79.

Gilboa, I. (1990a), "A Note on the Consistency of Game Theory," in the Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge, R. Parikh (ed.), Morgan-Kaufmann, 201-208.

Gilboa, I. (1990b), "Philosophical Applications of Kolmogorov's Complexity Measure," mimeo.

Gilboa, I. and D. Schmeidler (1988), "Information Dependent Games: Can Common Sense Be Common Knowledge?", Economic Letters, 27, 215-221.

Jain, S. and A. Sharma (1990), "Hypothesis Formation and Language Acquisition with an Infinitely-Often Correct Teacher," in the Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge, R. Parikh (ed.), Morgan-Kaufmann, 225-239.

Lewis, D. (1969), Conventions: A Philosophical Study, Cambridge: Cambridge University Press.

Lipman, B. (1989), "How to Decide How to Decide How to. . . : Limited Rationality in Decisions and Games," Carnegie-Mellon University working paper.

Lucas, R. (1972), "Expectations and the Neutrality of Money," Journal of Economic Theory, 4, 103-124.

Montague, R. (1974), "Syntactical Treatment of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability," in Formal Philosophy, by R. Montague, Yale University Press, 286-302.

Nash, J. F. (1951), "Non-Cooperative Games," Annals of Mathematics, 54, 286-295.

Popper, K. R. (1934), Logik der Forschung; English edition (1958), The Logic of Scientific Discovery, London: Hutchinson and Co. Reprinted 1961. New York: Science Editions.

Reny, P. (1988), "Rationality, Common Knowledge and the Theory of Games," mimeo.

Savage, L. J. (1954), The Foundations of Statistics, New York: Wiley.

Shapley, L. S. (1953), "A Value of n-Person GAmes," in Contributions to the Theory of Games, Vol. II, H. W. Kuhn and A. W. Tucker (eds.), Princeton: Princeton University Press, 307-317.

Suppe, F. (1974), The Structure of Scientific Theories (edited with a critical introduction by F. Suppe), Urbana, Chicago, London: University of Illinois Press.

Tarski, A. (1956), "The Concept of Truth in Formalized Languages," in Logic, Semantics, Meta-mathematics, by A. Tarski, Oxford University Press, pp. 152-278.

Thomason, R. (1982), "Paradoxes of Intentionality?", mimeo.

Thomason, R. (1986), "Paradoxes and Semantic Representations," in the Proceedings of the First Conference on Theoretical Aspects of Reasoning About Knowledge, J. Halpern (ed.), Morgan Kaufmann, 225-239.

Tversky, A. and D. Kahneman (1986), "Rational Choice and the Framing of Decisions," The Journal of Business, 59, S251-S278.

von Neumann, I. and O. Morgenstern (1944), Theory of Games and Economic Behavior, Princeton: Princeton University Press.