האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

RULE-RATIONALITY VERSUS ACT-RATIONALITY

by

ROBERT J. AUMANN

Discussion Paper # 497 December 2008

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Rule-Rationality versus Act-Rationality

Robert J. Aumann*

Abstract

People's actions often deviate from *rationality,* i.e., self-interested behavior. We propose a paradigm called *rule-rationality,* according to which people do not maximize utility in each of their acts, but rather follow rules or modes of behavior that usually—but not always—maximize utility. Specifically, rather than choosing an act that maximizes utility among all possible acts in a given situation, people adopt rules that maximize average utility among all applicable rules, when the same rule is applied to many apparently similar situations. The distinction is analogous to that between Bentham's "act-utilitarianism" and the "rule-utilitarianism" of Mill, Harsanyi, and others. The genesis of such behavior is examined, and examples are given. The paradigm may provide a synthesis between rationalistic neo-classical economic theory and behavioral economics.

1. Introduction

The assumption of rationality—that people act in their own best interests, given their information—underlies most of economic theory and indeed of economics as a whole. Economic policy revolves largely around the creation of incentives for people to act as the policy maker would like; and to act in accordance with one's incentives is, of course, to act rationally. Courses in Eco 101, in price theory, and in micro revolve around maximizations, first and second-order conditions, and so on. Applications of economic theory to various and sundry areas such as law, criminology, marriage, patents, health, finance, pensions, sports, what have you, work with maximizations—i.e., rationality—on the most basic level.

Even before the advent of Behavioral Economics, the rationality assumption was called into question, or modified, in one way or another. Herb Simon [1947]

*Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem

suggested the notion of satisficing: that people do not necessarily maximize, but only seek some acceptable level of utility; also, that people use heuristics rather than calculating optima [Newell, Shaw, and Simon 1962]. Milton Friedman promulgated the "as if" dogma: that people do not consciously optimize, but only act as if they do.[1] In experiments such as probability matching [Siegel and Goldstein 1959] and the ultimatum game [Güth et. al. 1982], subjects deviated systematically from utility maximization. Tversky and Kahneman [1974] pointed to various systematic violations of rationality. Admittedly, most of their work, and the subsequent development of Behavioral Economics, is based on polls and laboratory experiments; yet there are also actual empirical data that point to systematic deviations from rationality. Nevertheless, rationality remains the central paradigm of mainstream economics.

Viewing the rationality paradigm as a "thesis" (in the sense of Hegel [1807]), and the apparent irrationality discussed above as its "antithesis," we here suggest a "synthesis," namely *rule-rationality.* Ordinary rationality means that when making a decision, economic agents choose an act that yields maximum utility among all acts available in that situation; to avoid confusion, we henceforth call this *act-rationality.* In contrast, under rule rationality people do not maximize over acts. Rather, they adopt *rules,* or modes of behavior, that maximize some measure of total or average or expected utility, taken over all decision situations to which that rule applies; then, when making a decision, they choose an act that accords with the rule they have adopted. Often this is the act that maximizes their utility in that situation, but not necessarily always; the maximization is over rules rather than acts.

Four remarks are in order:
(i) The rule need not—in general, will not—be consciously adopted. Its adoption could be the result of evolutionary forces, genetic or memetic.[2] Or it could be the result of a learning process, which again, may or may not be conscious.
(ii) Often, the rule will be executed by means of a *mechanism,* which expresses the rule only indirectly. One example of such a mechanism is the notion of "honor" (as in O'Neill [1999]); we will encounter others in the sequel.

[1] See, e.g., his remarks (indented) on p.8 of Hetzel (2007), where he explicitly discusses the "as if" doctrine in connection with the rationality assumption. The doctrine is also discussed in Friedman (1953), in connection with assumptions such as perfect competition, or with theory building in general, not specifically with rationality.

[2] Richard Dawkins [1976] coined the word "meme" for the social analogue of a gene—a trait that propagates itself in Society because it is generally successful. We will encounter many memes in the sequel.

(iii) The distinction between act and rule rationality is analogous to the distinction between the *act-utilitarianism* of Jeremy Bentham [1789] and the *rule-utilitarianism* of John C. Harsanyi [1980] and others. With both act-rationality and act-utilitarianism, one optimizes an *act*—does the best one can in a specific situation; in contrast, with both rule-rationality and rule-utilitarianism, one optimizes a *rule* so as to do well "in general," but not necessarily always. But utilitarianism (of both kinds) is very different from rationality (of both kinds). Utilitarianism is a moral imperative; it is about how moral people "should" act, for the benefit of Society as a whole. Rationality is egoistic; it is about advancing the interests of the decision maker himself[3] only.
(iv) Much of the material presented here is based on ideas that have been "in the air" for years. What we do believe to be new is the perspective—putting together ideas such as evolution, the "as if" doctrine, rule utilitarianism, perceived deviations from rationality, and so on.

2. Evolution and Rationality

2.1. The Formal Analogy

The connection between evolution and rationality has been recognized for decades [Maynard Smith and Price 1973, Dawkins 1976]. To start with, there is a purely formal analogy. With rationality, a decision maker chooses an act that maximizes utility; analogously, with evolution, a population selects a trait that maximizes *fitness,* defined as the expected number of offspring. The decision maker corresponds to the population; his choice corresponds to evolutionary selection; and utility corresponds to fitness. In each case, an element (act or trait) is chosen from a set (feasible acts or traits) to maximize some function (utility or fitness); but whereas acts are consciously chosen by decision makers, traits are selected, totally unconsciously, by an evolutionary process that operates automatically.

As there has been confusion over the workings of the process, it is worthwhile to enter into some detail. Every individual in a population has a genetic endowment, which is passed on from generation to generation. In one way or another, e.g. by mutation, occasional alterations in the gene pool of the population occur. Such alterations usually affect a single individual only. If the trait that an alteration prescribes increases the fitness of that individual, then by the definition of

[3]Masculine pronouns indicate indeterminate as well as male gender, as was customary in the past.

"fitness", that individual may expect to have more offspring than the rest of the population. These offspring also possess the altered gene, so they, too, may expect to have more offspring than the rest of the population. The proportion of the altered gene in the population thus increases exponentially, and eventually takes over the whole population. If the trait still does not maximize fitness, then in the course of time there will be a mutation or other alteration that further improves fitness; like before, it, too, will eventually take over the population. Thus a trait that maximizes fitness emerges.

On the other hand, if the alteration in question decreases fitness, it will not propagate in the population, and so has no significant effect.

To be sure, in practice the process is perhaps not so cut-and-dried. The "set of feasible traits" is not clearly defined, expectations need not necessarily translate into realizations, the environment may change in mid-stream, sexual reproduction screws up the story, and so on. Nevertheless, the essence of the description applies.

The vital element to be noted is that the process is *entirely mechanical.* No one consciously—or even unconsciously—chooses, or maximizes, anything; no volition is involved. It is "as if" somebody was trying to maximize fitness; but that's not really what's going on. In contrast, rational decision-making is all conscious, all volitional, without any "as if."

2.2. The Substantive Relation

In the above discussion, rational, utility-maximizing decision making is the primary element; evolution plays a secondary, "as if" role. We now reverse the roles, and assert that evolution is the fundamental driving force, that ordinary utility-maximizing rationality is a *product* of evolution. Rationality has evolved, alongside of physical features like eyes, stomachs, limbs, and breasts, *because it maximizes fitness.* A person shopping around for lower prices is maximizing fitness, because the money saved can be used to purchase food, theater tickets for a date, shelter, attractive clothing, education for the children, and so on—all of which increase fitness.

To be sure, there is a missing link: the relation between utility and fitness. Rationality maximizes utility; evolution maximizes fitness. Utility expresses preferences—what an individual likes, what he wants to do. Does he always single-mindedly want to increase fitness—the number of his offspring? An obese person craving another piece of chocolate will maximize his utility by eating it, but doing so is unlikely to enhance his fitness.

Nevertheless, as a rule, utility and fitness do in large measure go together. For now we leave it at that; the exact relationship is explored more carefully below (Section 6.2).

3. Some Examples

3.1. Bees and Flowers

In an experiment conducted some twenty years ago, a biologist by the name of Andreas Bertch, from Marburg, Germany, studied the behavior of bees in a "field" of artificial flowers. The field consisted of a rectangular array of several dozen disks, each with a diameter of several centimeters, and each colored either blue or yellow. In the center of each disk was a tube that could supply "nectar"—i.e., sugar water. Initially, only blue "flowers" were programmed to supply nectar. When a batch of bees emerged from their cocoons, it was let loose on the field, and soon learned to visit blue flowers only. After some time, Bertch changed the programming: now only yellow flowers supplied nectar. The bees, however, continued to visit blue flowers only, and eventually died of starvation.[4]

This seems highly irrational. Upon finding that there is no nectar in the blue flowers, rational bees "should" have at least tried the yellow flowers. Why did they "prefer" death to trying something different?

To understand this, one must first understand why one would expect rationality from a bee. The answer is set forth in the previous section—rationality is an expression of evolutionary forces. In these terms, one may rephrase the question as follows: Why didn't evolution program the bee so that when blue flowers cease giving nectar, it turns to alternative sources?

The answer is that there was no evolutionary pressure for this kind of development: the situation in Bertch's laboratory *never occurs in nature.* In nature, the colors of nectar-supplying flowers do not change during the lifetime of a bee. Therefore, it is sufficient in nature for the bee to learn in its youth which flowers supply nectar, and stick to this throughout its life. So evolutionary pressures have produced a *learning window*—a period of time during the bee's youth when it learns which flowers give nectar. After that, it *cannot* learn anything new. This

[4] Private communication from Prof. Avi Shmida of the Department of Ecology and the Center for the Study of Rationality at the Hebrew University of Jerusalem. Shmida adds that Bertch never published his results, but that Prof. Reinhard Selten of the University of Bonn witnessed the experiments, and they were recorded on videotape by Shmida.

is sufficient for the requirements of bees in all situations that occur naturally.

This is a good example of rule-rationality. The rule is, "stick to what you learned in your youth." The *mechanism* for executing the rule is the learning window. In Bertch's laboratory, the rule does not lead to *act*-rationality, which would call for the bee to try yellow flowers when the blue ones stop giving nectar.

3.2. The Ultimatum Game

In 1982, Güth et. al. conducted an experiment that has come to be known as the *Ultimatum Game;* here we discuss just one version. The rules are simple. Two players, the *proposer* and the *responder,* must divide DM 100.[5] If they agree on the division, each receives his agreed share. If they do not agree, neither receives anything.

The game was played with many pairs of players, each player participating just once. The players did not sit face to face, and could not communicate directly. Rather, they sat at computer consoles in separate rooms. The proposer started by making an offer to the responder; the offer was numerical only, with no accompanying words. The responder could respond only by typing "yes" or "no" into the computer; no other response was allowed. Once he had responded, the game was over. After that, the players received their payoff (if any) and left by separate doors. At no stage did they see each other or learn each other's identity. The subjects were students—presumably not particularly long on money.

In this situation, one might expect the proposer to offer the responder a non-negligible amount—say DM 10, taking DM 90 for himself—and for the responder to accept. That is because there is no rational reason for the responder to walk away from a non-negligible amount of money; and taking this as given, a rational proposer should maximize his payoff.

But that is not what happened. Most offers were in the neighborhood of 65-35. And when they were considerably less—say 80-20—they were *rejected:* the responder actually walked away from as much as DM 20.

On the face of it, this seems to be a clear violation of act-rationality. Not on the part of the proposer, who—perhaps foreseeing the response—is rational in not risking rejection; but on the part of the responder.

Possible explanations include wounded pride, a feeling of being insulted, self-respect, and a desire for revenge. Another explanation that might be suggested is

[5] The Deutsch Mark (DM) was the currency of Germany at the time the experiment was conducted. Roughly, DM 1 in 1982 is equivalent to 1 Euro in 2008.

that the responder wishes to establish a reputation for rejecting lop-sided offers, so that in future negotiations, he will not get such offers. But that explanation does not hold water, because the game was played entirely anonymously; no one was told the players' identities, so reputations could be neither established nor destroyed.

There are two ways of viewing pride, insults, self-respect and revenge. One is that they themselves are legitimate sources of utility and disutility, so the responder is behaving entirely rationally when he rejects a 90-10 offer; he actually gets positive utility from taking revenge, and he would get negative utility from accepting an insulting offer. That is a perfectly consistent, logical position.

But conceptually and methodologically, it is not quite satisfactory; one might wish to delve deeper. Rather than taking emotions like the above as given, one might wish to *account* for them in terms of more fundamental human needs. What purpose—evolutionary or otherwise—does it serve to feel insulted, or to take revenge? What is the function of self-respect?

That's where rule rationality comes in. We suggest that even though it isn't act-rational for the responder to reject an 80-20 offer, it *is* rule-rational to do so. As a rule, one should reject lop-sided offers, precisely for the reputational reason discussed above: so as to be treated more even-handedly in the future. People use this rule because it is *usually* act-rational: specifically, in almost all—or all—natural, "real-world" negotiations, which are not anonymous. The *mechanism* for executing the rule is a combination of the emotions discussed above—self-respect, wounded pride, a desire for revenge, and so on, which evolved, genetically or memetically, because they usually maximize fitness. In Güth's laboratory, the rule does not lead to act-rationality, which would call for the responder to accept any positive sum.

Note that this entire discussion is about the *responder;* it is his behavior that is rule-, but not act-rational. The behavior of the proposer, who in these experiments usually proposes at least DM 30 to the responder, *is* act-rational, since he fears a rejection by the responder—rightfully.

3.3. Food

Eating is an excellent example of rule-rationality. From the evolutionary point of view, it is rational as a *rule*; one needs food for energy and growth. But as we all know, one can overeat, and then eating becomes act-irrational. Nevertheless, people continue to eat even then.

The *mechanism* for executing the rule is hunger; also the other side of the same coin, namely the enjoyment of food. Usually, the direct motivation for eating is not to get energy, but hunger and food enjoyment. Both are genetic; they evolved in order to motivate organisms to eat. Evolution did not "design" the mechanism to cope with the sedentary nature of much of modern life, so it sometimes misfires, so to speak. Thus, in spite of the rule-rationality of eating, it is sometimes act-irrational.

Overeating may be act-irrational not only for overfed people, but at the opposite end of the spectrum, also for the severely undernourished; there are documented cases of people who survived the concentration camps during the Holocaust, only tragically to die of overeating upon being liberated. Again, evolution did not design the system to deal with this situation, because it is unusual.

3.4. Sex

Sex is another good example of rule rationality. Engaging in sex is rational as a rule, because it increases the expected number of offspring. To execute the rule, nature evolved the *mechanism* of the sex drive: The reason that people (and animals) have sex is, in general, not that they consciously want children, but that they enjoy sex, it fulfills a physiological need.

But often, the mechanism misfires; or at least, does not serve the purpose for which it was "designed." The sex drive leads to many activities that have no chance of producing offspring: sex with birth control, sex after the reproductive age, homosexuality, masturbation, oral sex, bestiality, pornography, and so on. At best, such activities are evolutionarily neutral—neither increase nor decrease fitness. But they can also be harmful, as when a sexually transmitted disease is contracted. So sex is always rule-rational, but may be act-irrational.

Human beings are not alone in engaging in act-irrational—but rule-rational!—sex. Orchids of the genus *Ophrys* resemble female bees, give no nectar, and are constructed so that visiting bees cannot eat the pollen; it sticks to their brows. Male bees visit these orchids, ejaculate on them, and then visit other orchids and pollinate them with the pollen on their brows. Biologists like to say that the orchid "fools" the visiting bee into thinking that it is a female bee. It is doubtful that that assertion is meaningful, as it implies that bees have conscious desires and make conscious decisions. But even if meaningful, it need not be true. People who masturbate over provocative erotic pictures or stories are not "fooled" into thinking that they are having sex; it is simply that their sex drives make them

masturbate. Bees are no different. The rule-rationality of the bee thus plays an important role in the natural history of the orchid.

3.5. Arrow's Pacific Island Story

The following story was related by Kenneth Arrow, professor emeritus at Stanford University and 1972 Economics Nobel Laureate (private communication). During World War II, a squadron of American bombers based on a Pacific island was assigned the mission of flying twenty-five bombing sorties to a Japanese-held island 800 miles away. Because of the great distance, most of the weight that the bombers could carry was needed for fuel; very little could be used for the payload: the bombs. This mission was very dangerous; in similar previous missions, only a quarter of the airmen had survived. Just as the mission was about to begin, an Operations Research officer arrived from Washington with a brilliant proposal: instead of the planned mission, *half* the airmen—to be chosen by lot—would fly just one single sortie, but this sortie would be *one-way.* As a result, much more weight could be devoted to bombs, and in that single one-way sortie, as many bombs could be delivered as in the twenty-five round-trip sorties. And, the survival probability of each airman would increase from 1/4 to 1/2.

The airmen unanimously refused the generous offer of the OR officer from Washington. When asked the reason for their refusal in individual interviews, each one replied that *he* is a much better pilot than average, that *he* will not be shot down.

Clearly, this behavior of the airmen was act-irrational. But it *was* rule-rational! In the army, especially in a war, things change so rapidly and unexpectedly that it makes no sense for individual soldiers to make long-term plans. Even in a trivial matter like getting leave, if you're given a choice between this weekend and next weekend, you always take this weekend; by next weekend, your service may be cancelled, or in the opposite direction, all leaves may be cancelled. So the rule that soldiers *subconsciously* adopt is, "look ahead just one day—stay alive today—tomorrow will take care of itself." That is what the airmen were doing, without being aware of it. Though act-irrational, they were being rule-rational.

The story has a beautiful, surprising denouement. After three sorties, an order came from Washington cancelling the whole mission. So the airmen had after all been right—the unconsciously adopted rule worked!

3.6. Selten's Umbrella

Until recently,[6] Reinhard Selten, professor emeritus at the University of Bonn and 1994 Economics Nobel Laureate, *always* carried an umbrella; even in Israel's Negev desert in the summer, when it never rains. He did so because in Germany one cannot tell when it is going to rain, so carrying an umbrella is indeed act-rational; and it was too time-consuming and inconvenient to ascertain on each day in each place that he visited whether or not to carry an umbrella. This is an unusual case in which the rule of behavior, though not always leading to an act-rational decision, is adopted *deliberately,* so there is no need for a mechanism to bring about its adoption.

3.7. Probability Matching

In 1959, Siegel and Goldstein conducted an experiment that has come to be known as *Probability Matching.* Since then, it has been repeated hundreds of times; here we discuss just one version. A subject is seated in front of a device that emits, once in ten seconds, either a red or a green light at random. The probability of red is 1/4, that of green 3/4. Each time, the subject must predict the color of the light; if he succeeds, he is rewarded. Overwhelmingly, subjects predict red 1/4 of the time, green 3/4 of the time. That is not optimal, as the probability of success is then only 5/8, whereas *always* predicting green has a success probability of 3/4.

Before continuing, we observe[7] that this finding is an artifact of the experimental set-up; in the real world, it does not obtain. Many people have a choice of routes in getting to work; sometimes one route is faster, sometimes another—there could be a delay caused by a road accident, or a breakdown of the underground, or a visiting dignitary, or a host of other occurrences. If, say, Route A is faster 1/4 of the time, and Route B 3/4 of the time, then a probability matcher would choose Route A 1/4 of the time, and Route B 3/4 of the time. But that is not what people do. People take the same route to work every day; in this case, presumably Route B, which is precisely the optimal strategy.

Why, then, do they behave as they do in the experimental setup? For one thing, people are not used to sitting in front of devices that emit colored lights at random; they have not developed rules to deal with such situations. Therefore, they use a different rule, *social desirability*—that subjects want to be seen

[6] When personal circumstances forced him to abandon the practice.

[7] The observation was privately communicated by Professor Jacques Drèze.

in a favorable light—which has been observed by psychologists to apply in experiments.[8] In our case, subjects want to show their skill at "guessing right;" simply always making the same prediction would, they think, make them look obtuse, dull, obssessive. And, as noted, the situation in which they find themselves is unfamiliar; they have no experience. So they try to "look good;" in general, this is a good rule to apply in interactions with people.

As for people's behavior in getting to work: though this is in fact act-rational, it is unlikely to be the product of conscious maximization. On the contrary, it is a consequence of the general rule, "learn from experience, do what is best for you in general." In getting to work, "looking good" does not apply; one simply wants to get there asap. Thus both sides of the probability matching phenomenon—its occurrence in the laboratory, and its non-occurrence in the field—are attributable to rule-rationality.

3.8. Cooperation and the Gene for Altruism

In many—perhaps most—human interactions, cooperation is a good idea. Generally, when people help each other, all concerned are better off. Such cooperation may be act-rational when the sides enter into an enforceable agreement, like a contract. Or, it may be act-rational in a repeated interaction, as when people repeatedly do business with each other; see, e.g., Aumann [1981, 2006]. In such cases it may take the overt form of altruism: I help you today, ostensibly without any quid pro quo, and you help me tomorrow, also ostensibly without any quid pro quo. Or, we cooperate every day, even though on each day each agent separately would be better off acting selfishly (as in the Prisoner's Dilemma). In repeated interactions, such behavior is act-rational if each player reacts to selfish behavior on the part of the other by acting selfishly himself—or perhaps even "punishing" the other—in the future.

But it could—indeed does—occur also in one-time encounters, even when it is quite clear that the encounter is indeed one-time. What can account for this?

The answer is that acting altruistically (within limits)—i.e., truly without a quid pro quo—may be rule-rational. Rather than keeping accounts of who helped whom when, it may be simpler just to be generous, as a rule. Many human interactions are at least potentially repeated or long-term; in such cases, acting generously as a rule will work vis-a-vis others who also are generous as a rule, and also vis-a-vis others who *do* "keep accounts." It is not act-rational, because in an

[8] Communicated by Professor Maya Bar Hillel.

interaction that is one-time for sure—such as tipping in a far-off restaurant that will not be visited again—the decision maker could perhaps do better by acting selfishly.

What we are suggesting here is that altruism is a *mechanism* for achieving cooperation (in the absence of an explicit enforceable agreement), in much the same sense that pride, feelings of insult, self-respect and revenge are mechanisms for getting "reasonable" offers in the ultimatum game (Section 3.2 above). We intimated in that discussion that such traits evolved—genetically or memetically—because they usually, but not necessarily always, maximize fitness. Similarly here, altruism evolved, genetically or memetically, because it promotes cooperation, and so *usually* maximizes fitness.

As between genetic and memetic (i.e., social) evolution, the latter may seem more likely to account for altruism. But in fact, the opposite is true. In surprising and beautiful recent research, Knafo et. al. [2008] identified a mollecular basis for altruism—a real physiological *gene,* an identifiable part of the DNA! This was done in a laboratory experiment using the "Dictator Game," in which one player—the *dictator*—makes a unilateral decision regarding the distribution of a fixed sum of money between himself and another player, the *recipient.* It was found that dictators possessing the gene in question allocated significantly more to the recipient than other dictators. So we have here a direct biological basis for this form of rule-rationality.

As in Section 3.2, one could simply stop there: take altruism as given—a legitimate source of utility—just as some workers take revenge, insult, etc., as legitimate sources of utility and disutility. Indeed, workers in this area use the term "other-regarding preferences" to "explain" such behavior. But again as above, this is conceptually and methodologically not quite satisfactory; one might wish to delve deeper. Rather than taking altruistic behavior as given, one might wish to *account* for it in terms of more fundamental human needs. What purpose—evolutionary or otherwise—does it serve? What is its function?

That is the question addressed in the current treatment. And that question is particularly apt in view of the existence of a mollecular basis—a gene—for the behavior in question. It is all well and good to speak airily about "other-regarding preferences," but when you have a gene staring you in the face, you've got to ask yourself, from where did this come? How—and why—did this gene *evolve?*

4. Rule-Utilitarianism versus Act-Utilitarianism

The distinction between rule- and act-rationality is analogous to that between rule- and act-utilitarianism. *Utilitarianism* is a philosophical doctrine that judges the morality of behavior by the extent to which it advances the interests of Society as a whole; i.e., increases "social utility," defined as an aggregate (such as the sum) of individual utilities. Thus the most moral, ethical behavior is that which maximizes social utility. The idea may be traced to the ancient Greeks; apparently the first coherent formulation is that of Bentham [1789], and the term was coined by Mill [1861].

Bentham's original concept is often called *act*-utilitarianism. A related concept, called *rule*-utilitarianism, holds that one should not always necessarily act so as to maximize social utility; rather, one should follow rules of behavior that usually—but not necessarily always—increase social utility. Among the prominent promoters of rule utilitarianism was Harsanyi [1980].

A forceful illustration of the idea of rule utilitarianism was provided by Fyodor Dostoyevsky [1866] in his famous novel "Crime and Punishment." Raskolnikov, a penurious young student, murders a vicious, despicable old moneylender for her money. By all accounts, the murder increases social utility: For one thing, the world is much better off without the moneylender; for another, the transfer from her to Raskolnikov also increases social utility, since she does nothing at all with the money, while he starves.

But obviously, the murder cannot be considered moral. Why?

The reason is that Society cannot allow each individual to judge his actions on his own; inter alia, because of the moral hazard that that would entail. Society must develop rules, which apply to *classes* of acts. So Society has decided that murder—*all* murder—is to be considered amoral.

Rule- and act-rationality, and the distinction between them, are similar in form to rule- and act-utilitarianism, and the distinction between *them*. Both act-utilitarianism and act-rationality call for the individual to choose his acts in order to maximize some kind of utility: social in the case of utilitarianism, individual in the case of rationality. Both rule-utilitarianism and rule-rationality call for the individual to develop rules that maximize utility—social or individual, as the case may be—on the whole, but not necessarily in each individual case.

But in *substance,* utilitarianism and rationality are altogether different. For one thing, utilitarianism—of both kinds—is normative; it tells people how they *should* behave, if they want to be moral. Act-rationality, too, is normative; it tells

people how they should behave to advance their self-interest. But rule-rationality is a positive concept: it describes how people *do* behave. Indeed, when rule- and act-rationality conflict, a decision-theorist would almost always advise a decision-maker to act in accordance with act-rationality, not rule-rationality. For example, in the ultimatum game (Section 3.2), a decision-theorist would certainly advise the responder to accept an offer of DM 20—or even DM 1!

Another distinction between rule-utilitarianism and rule-rationality is that the former involves a deliberate choice on the part of the decision maker; a rule-utilitarian Raskolnikov would deliberately reject the idea of murdering the money-lender, *because* it violates rule-utilitarianism—i.e., is amoral.[9] In contrast, a rule-rational choice is almost never deliberate, in the sense of being made *because* it is rule-rational.[10] For example, in the ultimatum game, a responder who walks away from DM 20 will tell you that he is doing so "to teach the offerer a lesson," or something similar; he won't tell you that he's aware that the experimental setup precludes reputational effects, but nevertheless wants to follow a rule that enhances his reputation.

5. A Formal Framework

For clarity, a formal framework is useful. We now provide formal definitions of rationality, of both kinds; *mutatis mutandis,* they apply also to utilitarianism, of both kinds.

The elements of the formalism are

(1) a *chooser,*

(2) a set O (the chooser's *options),*

(3) a function u from O to the real numbers (the chooser's *utility function),* and

(4) an element c of O (the chooser's *choice).*

The choice is *rational* if it maximizes the chooser's utility over all his options; i.e.,

(5) $u(c) = \max\{u(o) : o \in O\}$.

The difference between act- and rule-rationality is not in the formalism, but in its interpretation. Act-rationality concerns a specific decision scenario; the

[9] *We* do not necessarily equate morality with rule-utilitarianism; but a rule-utilitarian Raskolnikov would, and we are discussing him. We ourselves take no position on this matter.

[10] Well, almost never. A possible exception is Selten's umbrella (Section 3.6 above).

chooser is a decision maker, and the options are possible acts—what the decision maker might do in that specific scenario.

For example, on a specific Sunday morning in Bonn, Professor Selten must decide whether or not to take an umbrella on his walk. There are only two options: taking—or not taking—the umbrella. The utility is determined by the convenience of having the umbrella if it rains, the inconvenience of having it if it does not rain, and Selten's estimate of whether it will rain, and how much, on that specific morning in Bonn.

Rule-rationality, on the other hand, concerns a whole class of decision scenarios. The options now are not acts—what to *do*—but *rules* for determining what to do in each specific scenario in the class. Formally, any function from specific decision scenarios to acts in such scenarios is a possible rule, though some such "rules" might be impractical. The utility of a rule is determined by its utility in each decision scenario in which it is applied, and also by the complexity of the rule itself, the informational requirements involved, and the resulting costs.

For example, in the case of Selten's umbrella, the rule must specify not only whether or not to take the umbrella on a specific Sunday morning in Bonn, but when to take it, and when not, on any day, anywhere in the world. One such rule could be, always take an umbrella; another, never take it; still another, always take it, except in Israel in the summer; yet another, decide act-rationally in each case.

Note that when referring to the choice of a rule, the term "decision maker"—which implies a deliberate decision process—is inappropriate. That's why in describing our formalism, we used the more general term "chooser," which allows for rules that are not deliberately chosen—as in most of the above examples.

In brief, act- and rule-rationality are determined by parallel processes: act-rationality chooses an act from a set of acts; rule-rationality, a rule from a set of rules. The point of this paper is that the term "rule-rational" applies also to acts: an act may well be rule-rational but not act-rational. That happens when the act, though itself not maximizing utility over all relevant acts, is prescribed by a rational rule—one that maximizes utility over all relevant rules. Indeed that is the case in the above examples.

The same formalism applies to act- and rule-utilitarianism, except that one must replace the chooser's personal utility function by *social* utility.

6. Discussion

6.1. Formalism and Reality

A word of caution: the formalism described in Section 5 should not be taken too literally, at least with rule-rationality. Applying the idea of rationality to rules rather than acts should be seen as a perspective—a way of looking at things—rather than as a fully laid-out theory.

Thus in the case of rule-rationality, items (2) and (3)—the set O of options (i.e., rules) and the utility function u—are often not very well specified. To define O, one should first specify the set of "decision scenarios" to which the chosen rule is meant to apply; and even after this set has been specified, it is not always clear what makes a rule feasible, when we consider it a member of the set O over which we wish to maximize. And even after the set O has been specified, it's not clear how to define the utility function u. Is this an average of the utilities of the acts that are engendered by the rule? Or is it the median, or the ninetieth percentile, or some other aggregate? How does one figure the utility of a *rule?*

Finally, when we say that the chosen rule c maximizes u, then in view of the fuzziness described in the previous paragraph, it does not really make sense to think of an absolute maximum. Rather, we should think of the rule as "doing very well" in the aggregate, and/or in general, but not necessarily achieving the absolute maximum.

6.2. Utility and Fitness

We return now to the above discussion (Section 2.2) of the relation between rationality and evolution, where we suggested that rationality has evolved because it maximizes fitness. The "missing link" there was that rationality maximizes utility, whereas evolution maximizes fitness. By the usual definition, utility expresses preferences—what a person likes, what he wants to do; he does not always singlemindedly want to increase fitness—the number of his offspring. Thus an obese person craving another piece of chocolate maximizes his "utility" by eating it, but surely not his fitness.

When utility is defined in this way—by preferences—the behavior characterized in the foregoing as "rule-rational" is, strictly speaking, in fact "act-rational:" it maximizes utility not only as a rule, but always. Indeed, one might argue that *all* behavior is then act-rational; that act-rationality is a tautology. Indeed, preferences are usually defined in terms of what one would do if faced with a choice

(see, e.g., Savage [1954], pp. 17 and/or 27-30). Thus by definition, the actual choice must be the preferred one; so if utility is defined by preference, and act-rationality maximizes utility, then the actual choice must be act-rational. From that viewpoint, then, much of the literature on bounded rationality, behavioral economics, and so on—as well as the current work—are off the mark.[11]

To make sense of this literature, one must define utility more substantively—in terms of basic desiderata like time, money, family welfare, life, health, food, and so on—which are indeed closely related to fitness. With such a definition, an act that maximizes utility is then indeed act-rational, and a rule that usually maximizes it, rule-rational.

6.3. The Literature

In the introduction we noted several bibliographic sources for the idea of "bounded" rationality or indeed irrationality. But the basic idea of rule-rationality—that much irrational behavior can, after all, be accounted for by the rational paradigm—is not really implicit in this literature. Closest, perhaps, is Milton Friedman's "as if" doctrine [Hetzel, 2007]. But that, too, comprises just one aspect—that behavior may be act-rational without any conscious attempt at maximization; but not the evolutionary genesis of rule-rationality, possible systematic failures of act-rationality, the concept of a rule and its maximality among all rules, and the matter of mechanisms.

As far as we know, the first published use of the term "rule rationality" is in Aumann [1997],[12] which has also a one-page account of the concept (Section 2.3).

7. References

Aumann, Robert J. (1981) "Survey of Repeated Games," in Böhm, V. (ed.), *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern,* Mannheim: Bibliographisches Institut, Wissenschaftsverlag, 11-42.

[11] Rather than pointing to specific choices that sppear irrational, some of the behavioral literature points to allegedly "inconsistent" choice patterns. These, too, can often be understood in terms of rule rationality—e.g., allegedly inconsistent choices may be covered by different rules.

[12] A write-up of the 1986 Nancy L. Schwartz Memorial Lecture at Northwestern University; a preliminary version is Aumann [1992].

— (1992) "Perspectives on Bounded Rationality," in Moses, Y. (ed.), *Theoretical Aspects of Reasoning about Knowledge, Proceedings of the Fourth Conference,* San Mateo: Morgan Kaufmann, 108-117.

— (1997) "Rationality and Bounded Rationality," *Games and Economic Behavior* 21, 2-14.

— (2006), "War and Peace," in Grandin, K. (ed.), *Les Prix Nobel 2005,* Stockholm: The Nobel Foundation, 350-358. Also, *Proceedings of the National Academy of Sciences (USA)* 103, 17075-17078.

Bentham, Jeremy (1789), *An Introduction to the Principles of Morals and Legislation,* London: T. Payne.

Dawkins, Richard (1976), *The Selfish Gene,* Oxford: Oxford University Press.

Dostoyevsky, Fyodor (1866), "Prestuplenie i nakazanie (Crime and Punishment)," *The Russian Messenger* (in twelve monthly installments).

Friedman, Milton (1953), "The Methodology of Positive Economics," in *Essays in Positive Economics,* Chicago: University of Chicago Press.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982), "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3, 367-388.

Harsanyi, John C. (1980), "Rule Utilitarianism, Rights, Obligations, and the Theory of Rational Behavior," *Theory and Decision* 12, 115-133.

Hegel, Georg W. F. (1807), *Phänomenologie des Geistes (Phenomenology of the Spirit),* Bamberg and Würzberg: Joseph Anton Gebhardt.

Hetzel, Robert L. (2007), "The Contributions of Milton Friedman to Economics," *Federal Reserve Bank of Richmond Economic Quarterly* 93, 1-30.

Knafo, Ariel, Salomon Israel, Ariel Darvasi, Rachel Bachner-Melman, Florina Uzefovsky, Lior Cohen, Esti Feldman, Elad Lerer, Efrat Laiba, Yael Raz, Lubov Nemanov, Inga Gritsenko, Christian Dina, Galila Agam, Brian Dean, Gary Bornstein, and Richard P. Ebstein (2008), "Individual Differences in Allocation of Funds in the Dictator Game associated with Length of the Arginine Vasopressin 1a Receptor RS3 Promoter Region and Correlation between RS3 Length and Hippocampal mRNA," *Genes, Brain, Behavior* 7, 266-275.

Maynard Smith, John, and G. R. Price (1973), "The Logic of Animal Conflict," *Nature* 246, 15-18.

Mill, John S. (1861), "Utilitarianism," *Fraser's Magazine* 64, 391-406, 525-534, 658-673.

Newell, Allen, C. Shaw, and Herbert A. Simon (1962), "The Process of Creative Thinking," in Grubert, H. E. and W. Wertheimer (eds.), *Contemporary Approaches to Creative Thinking,* New York: Atherton, 63-119.

O'Neill, Barry (1999), *Honor, Symbols, and War,* Ann Arbor: University of Michigan Press.

Savage, Leonard J. (1954), *The Foundations of Statistics,* New York: John Wiley.

Siegel, Sidney, and D. A. Goldstein (1959), "Decision-Making Behavior in a Two-Choice Uncertain Outcome Situation," *Journal of Experimental Psychology* 57, 37-42.

Simon, Herbert A. (1947), *Administrative Behavior,* New York: Macmillan.

Tversky, Amos, and Daniel Kahneman (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, 1124-1131.