Are Individuals Fickle-Minded?

Mathew D. McCubbins and Mark Turner

Abstract

Game theory has been used to model large-scale social events—such constitutional law, democratic stability, standard setting, gender roles, social movements, communication, markets, the selection of officials by means of elections, coalition formation, resource allocation, distribution of goods, and war—as the aggregate result of individual choices in interdependent decision-making. Game theory in this way assumes methodological individualism. The widespread observation that game theory predictions do not in general match observation has led to many attempts to repair game theory by creating *behavioral* game theory, which adds corrective terms to the game theoretic predictions in the hope of making predictions that better match observations. But for game theory to be useful in making predictions, we must be able to generalize from an individual's behavior in one situation to that individual's behavior in very closely similar situations. In other words, behavioral game theory needs individuals to be reasonably consistent in action if the theory is to have predictive power. We argue on the basis of experimental evidence that the assumption of such consistency is unwarranted. More realistic models of individual agents must be developed that acknowledge the variance in behavior for a given individual.

1 Introduction

Methodological individualism focuses on individual agents. It views large-scale social phenomena as the result of individual mental states that lead to actions. Specifically, in this book, methodological individualism is defined as the view that all explanations within the social sciences should be centered around individuals, their actions, beliefs, preferences, and the like. Accordingly, social phenomena such as the French revolution, an increase in the crime rate, residential segregation, the government's decision to lower the taxes, and the occurrence of unions are to be explained in terms of individuals, their actions, and the like.

Game theory analyzes outcomes as the aggregate result of choices that players make during interdependent decision-making. These choices are viewed as grounded in individual cognition by the players about the players involved, the actions they take, the information they possess, the strategies available to them, the outcomes they anticipate, and the equilibria that can be achieved across them.

Game theory makes a core set of assumptions about individual agents, namely, that they have consistent beliefs and preferences; that their actions result consistently from those preferences and beliefs; that these preferences, beliefs, and actions remain consistent across equal choice moments; and that the basic mental

processes and inference procedures by which preferences and beliefs lead to actions remain the same under all conditions.

It is widely established experimentally that subjects do not in general follow the predictions of game theory. Accordingly, behavioral game theorists have stepped in with new assumptions about consistent *deviations* from classical rationality and assigned to subjects consistent dispositions to account for these deviations — dispositions having to do with risk preference, cognitive abilities, social norms, etc. All of these theories are fundamentally cognitive theories, making claims about how individual human minds work when choosing.

In this article, we assess the game-theoretic modeling of individual agents. We argue that these game-theoretic models assume a consistency in these agents that is, as yet, unwarranted. We argue that more realistic models of agents must be developed on the basis of systematic empirical research.

2 Generalizing from experimental data

The conception in game theory that consistent preferences drive decisions has had extraordinary influence lately. Literature in the social sciences is replete with publications reporting choices by subjects engaged in economic games in laboratory settings. Typically, these articles draw macroscopic inferences for real behavior from the behavior of these individuals in the laboratory. For example, here is the thesis of a highly-cited 1995 paper on general human behavior in transactions:

> We designed an experiment to study trust and reciprocity in an investment setting. … Observed decisions suggest that reciprocity exists as a basic element of human behavior and that this is accounted for in the trust extended to an anonymous counterpart. … Is trust a primitive in economic models of behavior? What factors increase (or decrease) the likelihood of trust in economic transactions? We provide answers to these questions in a specific experimental setting, the investment game. By guaranteeing complete anonymity and by having subjects play investment game only once, we eliminate mechanisms which could sustain investment without trust; these mechanisms include reputations from repeat interactions, contractual precommitments, and potential punishment threats. We then show that positive investments still occur, suggesting that trust is an economic primitive. (Berg et al. 1995, pages 122-123.)

The reasoning in this article—leading to the conclusion that trust is an economic primitive for individual agents—follows a common path: some subjects

play a single game in a laboratory setting; their behavior is interpreted as the reliable outward sign of how they think and decide; accordingly their behavior in this one game in an experimental setting is used as a principle for modeling them as individual agents *tout court*; and this model of individual thinking and individual deciding is then generalized to human cognition and behavior in the world.

The ambition to use economic games run in laboratory settings as a microcosm of the world is understandable. Science prefers when possible to reduce vast complexity to simpler principles, to smaller pictures congenial to human thought. It would indeed be most useful if this reduction of social phenomena to a summation of actions by individual subjects in economic games proved to be scientifically legitimate.

But we should be cautious: the history of human ideas is replete with reductions that have turned out to be wrong, often to the surprise of generations of people who relied on them. In many cases, these misguided reductions are still powerfully with us. Vedic astrology continues to exert strong influence on decision-making among Hindus. The Tarot deck provides a remarkable microcosm for understanding the future and for planning accordingly, but it has not been shown to have any scientific value. Haruspication of entrails, augury, cartomancy, palmistry, pyromancy, and I Ching divination are similar reductions. These reduction strategies are often parodied—once they are discarded by a culture—for their vacuity.

In this paper, we present experimental evidence indicating that the drawing of macroscopic inferences about human behavior from the behavior of individuals in individual economic games in laboratory settings is not yet warranted.

3 What is Going on in the Laboratory?

Laboratories in the social sciences are unlike laboratories in the physical sciences. Laboratories in the physical sciences are constructed under the view that the physical conditions in the laboratory can be designed so as to match exactly the relevant conditions of interest in the world. There is in this conception nothing, so to speak, about the bench scientist's laboratory bench *per se* that stops the scientist from generalizing from what happens on the bench to the rest of reality, inclusive of those real situations that take place far from the scientist's bench. This happy conception of seamless generalizability allows the physical sciences, with the right care and nuance and adjustments, to claim that causal relationships detected in the laboratory for the most part generalize automatically to the world.

The case is utterly different for experiments on human beings, who belong to a social species evolved for behavior under certain conditions, and those conditions are not laboratory conditions. In principle, the burden is on the experimenter to show that the conditions in the laboratory do indeed match the relevant conditions of interest in the world in all the right ways. This can be a heavy burden, for several reasons.

We must assume as a beginning point that laboratory experiments involving economic games should fail to generalize to the world. The first reason we must make this defensive, defeasible assumption is that there are powerful and well-known "experimenter effects." For a physical or chemical system, or most biological systems, such as algae, no one imagines that the system is thinking, consciously or unconsciously, about the experimental situation and the experimenters. But human subjects are thinking, consciously and unconsciously, about both the experimental condition and the experimenters. Accordingly, absent compelling proof to the contrary, although one can assume that the data from such an experiment, if the experiment is impeccable on all other methodological scores, reveals something about how the subject behaves in that experimental condition, one cannot in principle assume that it reveals anything about how the individuals behave outside of that experimental condition.

The second reason for doubting that data from experiments involving economic games will generalize is that human beings and human cognition are evolved for messy environments. Conditions of the laboratory are sparse—which means that they are not the conditions for which human beings are evolved. The fact that the conditions of the laboratory are sparse requires us to be skeptical that generalization from experimental data to ecological human behavior is legitimate. Vision, for example, is evolved for conditions of white light, which are quite messy. Tests in the laboratory on human vision using simple, clean, monochromatic light do indeed show something important—namely how the human vision system operates under those experimental conditions—but it does not generalize to normal human vision. Color constancy, for example, an indispensable feature of human vision and inference, does not work under monochromatic light the way it does under normal conditions. An experiment under conditions that have not been demonstrated to match those in the real-world situations of interest is called "ecologically invalid," or just "invalid."

In sum, methodological rigor requires that we begin from the default assumption (albeit a defeasible default assumption) that laboratory experiments involving economic games do not generalize to human behavior. We are warranted

in giving up that assumption only where high covariance has been reliably demonstrated between the behavior in the economic game and in the normal human setting. It cannot in general be assumed that inferences established from the microcosm of the economic games in laboratory settings generalize to the human macrocosm.

4 Classical Economics and "Playing Nash"

Research using economic games in experimental settings begins with the baseline assumption—taken from classical economics, as in (Morgenstern & von Neumann 1947)—that subjects will optimize their payoffs within the strip of interdependent decision-making called the "game," and do so by assuming that other agents will optimize their own payoffs. On these assumptions, interaction in interdependent decision-making must follow an equilibrium path. In this article, we will say, without summarizing the well-known details, that a game-player is "playing Nash" when she is following a rule for play (a "solution concept") that will give her the optimal payoff that she can achieve through her own unilateral choices in a game where she is assuming both that all the other players are playing Nash and that all the players know each other's equilibrium strategies. More generally, we will say that the classical paradigm proposes to explain human behavior through closed-form analytic models as a function of the Players involved, the Actions they can take, the Information they possess, the Strategies available, the Payoffs for actions, the Outcomes for the players, and the Equilibria that can be achieved across players—PAISPOE, for short.

Experiments with subjects playing economic games show that in general they do not play Nash. This is the oldest news on the planet, and we have nothing to add to that consensus, except that our batteries of experiments show the same thing. Details are available in (McCubbins & Turner 2012; McCubbins, Turner, & Weller 2012a, 2012b, & 2012c).

5 Epicycles

Interpreters of data often guess why the players do not play Nash. To have scientific weight, these guesses would need to survive being tested as new hypotheses against out-of-sample data. Treating these guesses as knowledge would be "adding epicycles"—a slang term for "bad science." The term refers to the supposed penchant of Ptolemaic astronomers to preserve the underlying theory by adding cycles-upon-cycles-upon-cycles as needed to erase the divergence between the theory and the known data.

The need to avoid epicycles in scientific investigation is well understood. Gigerenzer (2004: 602) offers what he calls "Feynman's conjecture":

> To report a significant result and reject the null in favor of an alternative hypothesis is meaningless unless the alternative hypothesis has been stated before the data was obtained.

6 "Bounded Rationality" as an epicycle

The first and still the best-known patch for PAISPOE models is "Bounded Rationality," a term coined by Herbert Simon. According to Simon, rationality of the PAISPOE variety is limited because people lack information or have cognitive limits, including limits imposed by inability to think fast enough in the time available. Without a doubt, as every cognitive scientist knows, cognitive limits often make it impossible for people to do full PAISPOE calculation. Also without a doubt, lack of information can impede PAISPOE reasoning. Work by scholars such as Herbert Simon on "satisficing" and Gerd Gigerenzer on "heuristics" has contributed to our understanding of alternatives to PAISPOE reasoning.

Asserting that bounded rationality accounts for the mismatch between data and PAISPOE models is not in principle scientifically illegitimate. Quite the contrary. But the assertion is merely an epicycle if it is presented as an explanation for the mismatch, in the absence of a demonstration that a particular limit is indeed the cause of the mismatch.

When economists guess that subjects are failing to play Nash because bounded rationality impedes their ability to understand the structure of the game and its payoffs, the economists sometimes train the subjects on the game through "trial" rounds before they begin gathering the data that will be the basis of their conclusions. Training utterly stops any possibility of generalizing the behavior in the game to ecological behavior, for two reasons: (1) training creates an absolute difference between crucial conditions in the experiment and the ecological situations in which human beings have not been trained; (2) it is well-known in cognitive science that human beings can be trained to a frame that is contrary to their own patterns, and trained to it so well that it no longer seems alien; the benefits of such training are widely known in the martial arts, navigation, mathematics and scientific reasoning, diplomacy, and so on; and there is no reason to assume that behavior under training to a frame will generalize to normal human behavior—indeed, the mismatch was the very reason for training the human being.

Although it is indisputable that human rationality is bounded, adding "bounded rationality" to PAISPOE models as an epicycle has not provided us with

any better models of human behavior than were provided by classical economics of the Morgenstern & von Neumann variety.

7 Framing

Since framing can influence decision, it is often proposed that deviation from Nash is accounted for by framing. The classic example of such a framing analysis is Kahneman and Tversky's "Prospect Theory," which proposes that differences in the framing of a choice can bias the choice one way or another despite the fact that the framing is immaterial to the consequences of the action with respect to the payoff matrix (Kahneman & Tversky, 1979 & 2000). If we frame an action as a trade, then, since every trade is both a loss and a gain, it is possible for us to frame the action so as to emphasize loss or gain. Prospect theory proposes that there is a bias depending on this framing (Tversky & Kahneman, 1992; and Tversky and Fox, 2000): it is assumed within expected utility theory that choosers are in general risk-averse, but, on the contrary, ceteris paribus, there is, according to prospect theory, a four-fold pattern of risk attitudes: risk-seeking for gains of low probability and for losses of high probability; risk-aversion for gains of high probability and for losses of low probability (Tversky & Fox, 2000, p. 94). Accordingly, choosers will tend to make different choices depending on how the choice is framed—as loss or gain—despite the fact that the expected values of the alternative choices are identical. Kahneman & Tversky focus on framing effects in the decision-theoretic problem of choosing between alternative lotteries. Economists have since expanded this line of research into game-theoretic contexts, showing for example that framing affects the choice to contribute to a public good or impose externalities on others (see for example Andreoni, 1995; Cookson 2000). McDermott, Fowler, & Smirnov (2008) argue that "context-dependent" attitudes toward risk have a basis in evolutionary psychology. Post, Van den Assem, Baltussen, & Thaler (2008) show this same sensitivity to framing in the high-stakes choices of contestants on the game show "Deal or No Deal," a decision environment decidedly far-removed from the foraging of our evolutionary ancestors.[1]

8 Character type

[1] Unlike experimental studies of framing, Post et al. rely on observational data, in which the frame (previous earnings) is not controlled by an experimenter but generated endogenously by the subject.

It is often proposed that subjects have a certain character or psychological type that accounts for their deviation from Nash. For example, it is purported that people vary in the extent to which they are "self-regarding" versus "other-regarding." Purportedly, people vary in their "risk preference." Purportedly, people vary in their tendency to forego personal gain when doing so delivers a comparably much larger gain for other players. Purportedly, people vary in their preference for "fair" outcomes. And so on. It is also proposed that different players have different "level-k" signatures in particular settings. The idea behind "level-k" signatures is simple, and often used in films and novels. Consider, for example, this passage from *The Princess Bride:*

> The Sicilian smiled and stared at the wine goblets. "Now a great fool," he began, "would place the poison in his own goblet, because he would know that only another great fool would reach first for what he was given. I am clearly not a great fool, so I will clearly not reach for your wine."
>
> "That's your final choice?"
>
> "No. Because you knew I was not a great fool, so you would know that I would never fall for such a trick. You would count on it. So I will clearly not reach for mine either."
>
> "Keep going," said the man in black.
>
> "I intend to." The Sicilian reflected a moment. (Goldman 1973, 139-140.)

The Sicilian, to make his decision, is thinking that the man in black is thinking that the Sicilian is thinking that the man in black is thinking that the Sicilian is thinking that . . . . In theories of Level-k reasoning, we begin with Level Zero. It is not clear to us from the literature what Level Zero is thought to be, but it is described as "unstrategic thinking," so perhaps a Level-0 thinker (say Ann) simply shoots straight for the maximum payoff for herself in the payoff matrix, without any thought that the other player (say Paul) might have preferences of his own and so, strategically, interfere by making choices that move the Level-0 player, Ann, toward a different outcome, one that is better for Paul. "Unstrategic thinking" might mean that Ann chooses as if she is playing against random, non-intentional events—perhaps a role of the dice. She is then playing "against nature," in the economic parlance, where, oddly, "nature" does not include intentional human cognition. Paul is a Level-1 thinker if he is playing so as to interact optimally with a Level-0 player. And so on up the line: a Level-2 thinker is imagining what a Level-1 thinker is

thinking and responds optimally to a Level-1 thinker's strategy. And a Level-3 thinker responds optimally against a Level-2 thinker, and so on.

The Sicilian, a self-assessed genius ("Never go in against a Sicilian when death is on the line!") is many k-levels beyond everyday human subjects. He is even careful to prepare against potential adaptive behavior by the man in black (e.g, the man in black, an unequaled fighter, might try to kill the Sicilian, who is holding a large knife to Buttercup's throat to prevent this adaptive behavior): the Sicilian distracts the man in black, managing to get him to look away from the goblets for a second, during which brief interval of time the Sicilian switches the goblets. After they have drunk and swallowed, the man in black announces that the Sicilian has guessed incorrectly. The Sicilian crows, "You only *think* I guessed wrong . . . That's what's so funny. I switched glasses when your back was turned." Of course, the Sicilian dies in the next second from the iocane powder poison: the man in black, like human beings everywhere, has behaved adaptively rather than strategically in the game. As he explains to Buttercup, "They were both poisoned. I've spent the past two years building up immunity to iocane powder."

Guesses can be hypotheses, but not explanations. Adding terms or factors to a theory that has failed tests, for the sake of making the theory fit the data, is methodologically acceptable if these changes are regarded as new, untested hypotheses.

9 A Battery of Experiments to Test Epicycles

To test whether behavior in economic games can be generalized to the world at all, we look for the most likely candidate, that is, behavior to which the generalization is most likely to apply. We take it that behavior in one economic game is most likely to apply to behavior

1. *by the same subject*
2. *under identical experimental conditions*
3. *in closely similar economic games*
4. *very near in time*

Accordingly, we must put the identical subject through a continuous battery of such games under the identical experimental conditions. Furthermore, in running this battery, we should

1. avoid training subjects, as discussed above, yet
2. test that they understand the payoff matrices and strategies of other players by quizzing them;
3. make framing as spare and general as practical; and

4. hold framing as consistent as possible across this battery of measures so as to avoid variation in behavior owing to variation in framing.

This is what we have done. Our battery included many economic games.

10 Can we generalize from behavior in economic games?

Subjects in our experiments are told that they are randomly paired at the beginning of every task with someone else in another room and that all their behavior is anonymous and private and that all the subjects have the same information. They receive no feedback on their play or indeed any information except in the few interactive games that they play at the end of the battery (e.g. Trust), where they must be told what the other player sent. They know that they are paid for every task according to how they perform, and that they will be paid at the end of the experiment in private by an assistant who will know only their number and the envelope of cash to be given to the person with that number. The following analysis considers data from 190 subjects for four economic games in our battery: Trust, Dictator, Donation, and Majority Public Goods. Of course, we do not use these misleading names in describing the experiment to subjects. These names are only for ease of reference.

Let us begin with the Trust game. Player 1 and Player 2 both begin with $5. Player 1 can send any integer dollar amount to Player 2, including $0. Whatever Player 1 transfers is tripled and given to Player 2. Then Player 2 can return any integer dollar amount to Player 1, and the game ends. Notice that if Player 1 sends anything but $0, then Player 2's pot of money becomes at least twice as large as Player 1's, maybe much larger. For example, if Player 1 sends $1 to Player 2, then Player 1 is left at that point with $4, but the $1 sent is multiplied by 3, so Player 2 has $8. "Nash" for Player 1 is to send $0. Do our 190 subjects play Nash as Player 1 in Trust? Hardly. 105 of 190, or 55.3%, send money as Player 1. This is just an example of the well-established fact that human subjects, informed that they are playing in an economic game with other human subjects, cannot be relied upon to play Nash. It is this fact, as we discussed above, that induces the proposal of epicycles.

105 subjects out of 190 received money as Player 2 in the Trust Game. Can we count on them to play Nash and return $0? No. Given that they had every reason to view themselves as having been placed in an advantageous situation, can we count on them to be generous and not play Nash? No. 64 of these 105 subjects, or 61%, play Nice and return money, and 41 of these 105 subjects, or 39%, play Nash.

At this point, we can all feel the great temptation to "explain" these events by reducing the causality in the decision-making to personality: 61% of these Subjects

are "Nice" (or "other-regarding") and 39% are not nice (or "self-regarding," or whatever)—we call them "Nash."

Our central point is that such a conclusion depends upon assuming that human beings are consistent in their preferences and methods of making choices and that if they make a different choice it is because they face different conditions, yet it is just this assumption that most needs to be tested empirically.

Our battery of experiments was designed so as to let us test the hypothesis that "personality type" generalizes, in other words, that people are consistent. Does a subject's supposed "Nice" versus "Nash" type generalize to even the identical subject's behavior in identical settings and identical conditions during the same span of a few hours in a similar economic game? Our battery was designed so that subjects played both Player 2 in the Trust and Player 1 in Dictator *under the identical payoff conditions and, going forward, the identical game structure*. In the Dictator game, there are two players: the Dictator (Player 1) and the Receiver (Player 2). The Dictator has an endowment and chooses what part of it, if any, to send to the Receiver. The Receiver receives the amount sent and the Dictator keeps that part of the endowment the Dictator chose not to send, and that is the end of the game. The Receiver's role is entirely passive. We arranged our Dictator game so that the Dictator has the same endowment he or she has in the role of Player 2 in Trust, and that the Receiver has the same endowment he or she has in the role of Player 1 in Trust. These endowments are common knowledge. Accordingly, our Dictator game is identical to the second half of our Trust game. In effect, each individual subject plays the second half of the Trust game twice. Formally, there was no mathematical or economic difference in any individual subject's conditions as Player 2 in Trust and Player 1 in Dictator.

Specifically, for any specific subject S*, S* was in the role of Player 2 in Trust at one point in the battery and Player 1 in Dictator at another point in the battery. We introduce the label t(S*) for the other subject with whom S* was randomly paired in Trust. We introduce the label d(S*) for the other subject with whom S* was randomly paired in Dictator. In Trust, where both t(S*) and S* begin with $5, t(S*) sends an amount (perhaps $0), which is tripled and added to the $5 endowment which S* had at the beginning of the game. At that point in the Trust game, S* has *a* dollars and t(S*) has *b* dollars. Later in the battery of experiments, S* plays Player 1 in Dictator, and we arranged the experiment so that in Dictator, the endowment for that particular subject S* is exactly *a* and the endowment for specific subject d(S*) is exactly *b*. That is, the endowments that a given subject S* faces in the two games are identical at these two points in the two games. Here is a table:

Money that players have after the P1 send in Trust:

S* has *a* | t(S*) has *b*

Money that players have as endowments in Dictator:

S* has *a* | d(S*) has *b*

At this point in each of the two games, S* (for the 105 subjects who received money as Player 2 in Trust) has at least twice as much money as the other person. Accordingly, in Dictator, for these 105 subjects, *a* is always at least twice as large as *b*, and sometimes much bigger.

In both games, at this point, there is only one choice left to make, and that choice is identical in both games: S* must choose how much, if any, of *a* to send to the subject with whom S* is randomly paired in that game. So at this point in the two games, going forward, the two games have the identical structure and payoffs.

Do the purported Nice v. Nash personality types we might think we see when subjects play Trust Player 2 generalize even to the identical economic situation with the identical choice to make, now in Dictator?

No. 41 of the 64 Nice types as Trust Player 2 are Nice in Dictator, but 23 are Nash. So the generalization works for only 64% of the Nice subjects. 37 of the 41 subjects who are Nash during Trust Player 2 are Nash as Dictator Player 1; the generalization on "Nash" personality type holds (at this point) for 90% of subjects, making the "Nash" generalization look (at first blush) better than the "Nice" generalization, but still not a reliable generalization, since 10% of the Nash-types in Trust Player 2 are Nice as Player 1 in Dictator.

Next, we compare the behavior of each of these subjects in Trust to the behavior by the same subject as Player 1 in the Donation game. In Donation, both players begin with $5. Player 1 can pass any amount of the $5 to Player 2. The amount is multiplied by 4 before it is given to Player 2. Then the game ends. In this case, any amount of Niceness by Player 1 results in a fourfold level of Niceness received by Player 2, as measured in money. For example, if Player 1 passes $1, Player 1 is left with $4, but Player 2 now has $9. If Player 1 passes $5, Player 1 is left with $0, but Player 2 now has $25. Do the Nice versus Nash personality types we think we might see in Trust Player 2 and Dictator generalize subject by subject to the Donation Game?

No. Of the 41 most completely confirmed Nice types in Trust and Dictator, 27% suddenly are Nash in Donation. That is, 41 of the 190 subjects receive money as Player 2 in Trust, send money as Player 2 in Trust, and send money as Player 1 in Dictator. But 27% of those 41 are Nash in Donation! Now consider the 23 Ss who receive money as Player 2 in Trust, play Nice as Trust Player 2, but play Nash as

Dictator Player 1. What do they do in Donation? Half (12) of those 23 play Nice and half (11) play Nash. Consider the 37 subjects of 105 with the clearest Nash character type: they receive money as Trust Player 2 yet return nothing, and also send nothing as Dictator Player 1. Can we at least count on this confirmed 35% of the pool of 105 subjects who received money as Trust Player 2 to be rock solid Nash? No. In Donation, 30% of them play Nice. And so on.

Next, we compare what these specific subjects did when they played the Majority Public Goods game. In this game, each subject is assigned randomly to a group of 10 subjects (about whom they know nothing and with whom they cannot communicate) and given $5. The subjects can each keep the $5 or move the $5 to a group pot. If at least 6 of the 10 do so, then the pot is tripled and each subject in the group receives a 10% share of the pot. If fewer than 6 of the 10 do so, then nothing from the pot is given back to the subjects. This game is not a perfect distinguisher between Nice and Nash, because there is one place where they overlap. A subject who plays Nash will not contribute if the subject thinks that 0, 1, 2, 3, 4, 6, 7, 8, or 9 of the other players in the group will contribute. But if the subject believes that exactly 5 of the others will contribute to the pot, then the subject believes that joining the group of givers would raise its membership to exactly 6, in which case the subject receives $9 by contributing but $5 by not contributing. Can we rely on the purported Nice types to play Nice in the Majority Public Goods game?

No. For example, of the 30 subjects out of 190 who receive money as Player 2 in Trust, return money as Player 2 in Trust, send money as Player 1 in Dictator, and send money as Player 1 in Donation, 13, or 43%, do not contribute in the Majority Public Goods Game. Similarly, of the 26 subjects who receive money as Player 2 in Trust, return 0 as Player 2 in Trust, send 0 as Player 1 in Dictator, and send 0 as Player 1 in Donation, 5, or 19%, contribute in the Majority Public Goods Game. Similarly, for other sub-sub-sub-subcategories of the subjects, we find that a putative "personality" signature is unreliable in predicting behavior in the Majority Public Goods Game.

In summary, of the 105 subjects who received money as Player 2 in Trust, only 17, or 16%, keep a consistently "Trusting" or "Cooperative" or "Generous" signature, and only 21, or 20%, keep a consistently "Ungenerous" signature.

But what about those 105 subjects who *send* money as Player 1 in Trust? Surely they were Nice. In the Trust game, both players do much better if they trust each other: if Player 1 sends the full $5 available, then Player 2 has $20 and can send $10 back to Player 1. Both players then have doubled their initial endowment. Husbands and wives in community property states, or any two people under a

trusted contract according to which they split the benefit, should, under Nash, immediately send everything as Player 1 in Trust, because the contract means that you do not have to rely on the generosity of the other person: you own a 50% share of all assets, and so does the other player.

Let us compare behavior in the Trust Game with behavior in the Prisoner's Dilemma game (PD). Prisoner's Dilemma was another part of our battery. It is always set up so that a strategy of cooperating is strictly dominated by a strategy of defecting: Whichever choice the other player makes, the subject is always better off in choosing to defect. A Nash player, of course, must defect. Yet, if both players cooperate, they are better off than if both players defect.

The data are complicated at this point by the fact that we tested four different kinds of framing of the identical choice and payoff structure. The subjects did not all receive the same framing. Methodologically, we might prefer then not to lump them together, but this raises an interesting point: we often see in the literature data lumped together from different experiments with different protocols, run at different times by different experimenters, moreover using a between-subjects design. In our case, we have a within-subjects design, in the sense that the same subjects played both Trust and Prisoner's Dilemma. Under all four Prisoner's Dilemma framings, the payoffs were identical and everything else was held constant, except for the four framings. In all four versions, each subject had the choice to cooperate or to defect. Lumping these four groups together is not methodologically clean, but it is fairly conservative relative to the practices we often find in the literature, and our overarching point here is that one should doubt generalizations assumed in the literature. Can generalizations hold up over this lumping? Here are the results.

We start with the 105 Ss who play Nice as Player 1 in Trust. Did they play Nice in Prisoner's Dilemma? No. 34 of 105, or 32%, play Nice. But 68% play Nash.

Now let us look at the other 85 subjects. 85 subjects play Nash in sending $0 as Player 1 in Trust. Do they play Nash in Prisoner's Dilemma? Not so much. 65 of 85, or about 76%, play Nash. But 20, or about 24% play Nice.

11 Characterizations Do Not Hold Up

Perhaps there are other ways to use economic games as laboratories in which behavior can be generalized to the world. Perhaps there are other signatures, other reductions, in the form of characterizations. Perhaps there are other ways in which economic games can serve as a microcosm from which we can learn about the macrocosm of human behavior. But the ways we review here—all of which are forms of characterizing actors as having stable preferences and stable ways of

making choices—do not withstand our tests of their validity, and conclusions from them should be held in abeyance until science develops more realistic models of individual agents on the basis of systematic empirical research.

12 Conclusion

Common sense tells us that people have beliefs and desires, or beliefs and preferences; that they are aware of them; and that they act according to them. But cognitive science has undermined commonsense notions of the mind. What we take for granted about human thought has proved to be unimaginably more complex than anyone had expected; to be profoundly misrepresented by our supposedly bedrock, commonsense, intuitive notions; and to be conducted almost entirely in the backstage of cognition, invisible to consciousness. The cartoons of consciousness are highly useful, and there is no evolutionary advantage in building consciousness so that it can see through them. Human beings are awesomely effective, but for the most part clueless about how they work.

The basic assumptions about the human mind made in PAISPOE models may seem unassailable, sheer common sense, but that cuts no scientific ice. There status as common sense is no reason to accept them. Rather, they are hypotheses, and, as such, must be tested to have weight.

More than thirty years ago, Lee Ross (Ross 1977) coined the term "fundamental attribution error" for the excessive tendency of everyday "intuitive psychologists"—that is, everybody—to explain other people's behavior by attributing dispositions to them. Jocularly but also aggressively, and certainly influentially, he proposed that this error was the main basis for the field of social psychology.

There are many assertions made in economics that depend upon the folk-psychology assumptions we see in PAISPOE models, such as that what players are doing under a set of beliefs must be in equilibrium. Where the data diverge from the PAISPOE models, it is tempting to deploy the fundamental attribution error to explain that divergence between model and data, by adding an epicycle that consists of characterizing the actors by attributing stable dispositions. But perhaps the PAISPOE assumptions, and the assertions that depend upon them, are wrong in the first place. In cognitive science, commonsense notions of how vision works, how language works, how memory works, how categorization works, how inference works, and so on have all fallen by the wayside.

We propose that the future of economics lies not in drawing further conclusions from PAISPOE assumptions but rather in testing them scientifically in order to recast the foundations of the field.

If we accept the game-theoretic assumption that people have consistent preferences and consistent methods of deciding and that different choices are the result of different conditions, what are we to make of the data from our battery of experiments? One hypothesis would be that we somehow fielded a group of fickle-minded people, alien individuals, who gave us data that we must throw out as bizarre, or that the experiments are corrupt, or that the design involved confounds, and so on—and all of these possibilities should be considered.

But there is another logical possibility that we must also consider: perhaps the assumption that individual agents have consistent preferences and consistent methods of decision-making that run across different situations and contexts is wrong. After all, it is not clear on evolutionary or cognitive grounds that individual agents should be expected to work in this way. There is room for doubt. These assumptions of game theory may be reductions that we must discard. Before we launch centuries of research on the assumption that the solar system is geocentric, we should collaborate to take that assumption and test it to destruction. Before we launch centuries of research on the assumption that individual agents are to be modeled as consistent modulo circumstances, we should collaborate to take that assumption and test it to destruction. Our point is not at all that, with a little data, we prove that these assumptions are clearly wrong, but rather that we can now see that they are assumptions. We cannot base science on untested assumptions. If we are to build a house, we must build it on rock rather than sand, and if these principles are what we mean to build our house upon, we must first prove that they are rock and not sand. We have not done that.

This paper questions one prominent example of a theory committed to the thesis of methodological individuals, that is, game theory. We have argued that science needs a more adequate model of the individual actors than the one espoused by game theory. We need more realistic conceptions of agents—conceptions which must be developed on the basis of systematic empirical research. We do not suggest that a more adequate model of agents would automatically serve as support for the strong view of methodological individualism. No matter where the debates between methodological individualists and holists may land, there will be ample room in any workable social science for accounts that refer to actors.

References

Andreoni, J. 1995. Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments. *Quarterly Journal of Economics* 110(1): 1-21.

Berg, Joyce, Dickhaut, John, and McCabe, Kevin. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10: 122-142.

Cookson, R. 2000. Framing Effects in Public Goods Experiments. *Experimental Economics* 3: 55-79.

Gigerenzer, Gerd. 2004. Mindless Statistics. *The Journal of Socio-Economics* 33: 587-606.

Goldman, William. 1973. *The Princess Bride, S. Morgenstern's Classic Tale of True Love and High Adventure, The 'good parts' version, Abriged by William Goldman*. New York: Ballantine Books.

Kahneman, D., and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263-291.

Kahneman, D., and Tversky, A. (eds.). (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

McCubbins, Mathew D. & Turner, Mark. 2012. Going Cognitive: Tools for Rebuilding the Social Sciences. In *Grounding Social Sciences in Cognitive Sciences*, ed. Sun, Ron, Chapter 14, 387-414. Cambridge MA: MIT Press.

McCubbins, Mathew D., Mark Turner and Nicholas Weller. 2012a. The Theory of Minds Within the Theory of Games. *Proceedings of the 2012 International Conference on Artificial Intelligence*.

McCubbins, Mathew D., Mark Turner and Nicholas Weller. 2012b. The Mythology of Game Theory. *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*. Springer Lecture Notes in Computer Science.

McCubbins, Mathew D., Mark Turner and Nicholas Weller. 2012c. The Challenge of Flexible Intelligence for Models of Human Behavior. *Technical Report of the Association for Advancement of Artificial Intelligence Spring Symposium on Game Theory for Security, Sustainability and Health*.

McDermott, R., Fowler, J, H., & Smirnov, O. 2008. On the evolutionary origin of prospect theory preferences. *Journal of Politics* 70(2): 335-350.

Morgenstern, O. & von Neumann, J. 1947. *The theory of games and economic behavior*. Princeton: Princeton University Press.

Post, T., Van den Assem, M. J., Baltussen, G. & Thaler, R. H. (2008). Deal or no deal? Decision making under risk in a large-payoff game show. *American Economic Review* 98(1): 38-71.

Ross, L. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology* 10: 173-220.

Tversky, A. & Fox, C. R. 2000. Weighing risk and uncertainty. In Kahneman & Tversky 2000, p. 93-117.

Tversky, A. & Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297-323.