The Theory of Minds Within the Theory of Games

M.D. McCubbins[1], M. Turner[2], and N. Weller[3]
1 Marshall School of Business, University of Southern California, Los Angeles, California, USA
2 Department of Cognitive Science, Case Western Reserve University, Cleveland, Ohio, USA
3 Department of Political Science, University of Southern California, Los Angeles, California, USA

Abstract - *Classical rationality as accepted by game theory assumes that a human chooser in a given moment has consistent preferences and beliefs and that actions result consistently from those preferences and beliefs, and moreover that these preferences, beliefs, and actions remain the same across equal choice moments. Since, as is widely found in prior experiments, subjects do not follow the predictions of classical rationality, behavioral game theorists have assumed consistent deviations from classical rationality by assigning to subjects certain dispositions—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. All of them are fundamentally wrong in assuming one kind of consistency or another. Or at least, all of the proposals for consistency in belief, preference, and action with which we are aware turn out to be wrong when tested experimentally.*

Keywords: Behavioral game theory, experiments, cognition, Trust, Dictator, Donation

1 Introduction

Game theoretic models are utilized for behavioral predictions across a variety of domains such as allocation of security forces [1], allocation of health care services [2], and the design of political, social and market institutions [3; for relevant surveys see: 4, 5, 6, 7]. Despite the widespread use of game theoretic models to explain human behavior, we often observe behavior, from voting, to the divergence of political parties platforms, to market bubbles and crashes that do not readily accord with the predictions derived from game theory [8].

To address the discrepancy between predicted and actual behavior, scholars have proposed four common patches: (1) cognitive biases or errors in how people make decisions [9, 10, 11, 12]; (2) a mismatch between the experimenter's defined payoffs and an individual's actual utility [13, 14]; (3) the effects of uncertainty, bounded search ability, limited time for learning and equilibration, or limits in the ability of thinking about others' likely behavior [15]; and (4) newer and more clever equilibrium refinements that capture the folk psychology of different game theorists [16, 17, 18, 19].

While it is common to report that experimental subjects do not make choices that comport with Nash equilibrium strategies (or even von Neumann-Morgenstern utility maximization), we should not infer that human reasoning is thus somehow flawed. It is perhaps the case that our existing, deductive, models of human reasoning are too limited. Humans are able to solve many tasks that are quite difficult [20, 21]. Like vision, taste and smell, human intelligence and behavior are varied and flexible, creating an enormous diversity of beliefs and choices, but the models that we use to predict behavior do not and cannot capture this diversity. To build a better theory of human behavior, we must start with an appreciation for how we actually reason. As cognitive science has shown, intuitive notions of how the mind works (vision, language, memory, etc.) may be very useful for humans to hold as scaffolding for consciousness, but they are comprehensively wrong and simplistic. Intuitive notions of how we reason are not a basis for science. How we reason must be discovered, not assumed, and certainly not borrowed from intuition

One of the principal problems stems from the core solution strategy for noncooperative games of Nash equilibrium. While it is mathematically elegant, it requires agents in the game to have correct and consistent beliefs [22]. To have "correct beliefs" we assume that the agents, or the players in a game, to regard all other players as being "Nash players" and to predict that they all follow Nash equilibrium (NE) strategies. It is also typically required that players have "common knowledge" that they are all Nash players, that is, that they know that other players know that they themselves are following Nash equilibrium strategies, and so on, ad infinitum. Others have pointed out that "Common Nash refinements have similar attributes. Although these refinements differ in what they allow players to know and believe, they continue to require that actors share identical conjectures of other players' strategies" [23, p. 106]. If players do not believe that other players will adopt NE strategies, however, it is

Prepared for the 2012 International Conference on Artificial Intelligence.
Corresponding author: Nicholas Weller.

no longer true that players' best response to each other will be to follow a NE strategy. The natural, biological, or cognitive means by which this comes about are not specified, merely that, given enough time and effort, players can all learn what behaviors to adopt and when to adopt them, or, baring this eventuality, that societies will adopt rules, laws or norms to restrict and channel behavior to more efficient forms. Prior work on subjects' beliefs in experimental settings suggest that subjects possess non-equilibrium beliefs [24, 25] and that in at least some settings their behavior can be reasonable, given their beliefs [26].

In what follows, rather than trying to stitch together measurements from different experiments, with different protocols, run under different conditions and at different times, we use a within-subjects design, run over a single academic term, to investigate choices in a large battery of single-shot games. Within our battery of tasks we elicit subjects' beliefs about the other subject's actions, plus we elicit *recursive* beliefs about other subjects' *beliefs* in these games. We demonstrate that subjects' actions and beliefs are consistently inconsistent, deviating from one person to another and for each person from one task to another. A given subject is often not consistent in action across nearly identical choice moments at different times in the same within-subject battery of tasks and a given subject is often not consistent even at the same time in their beliefs, preferences and actions. We also see great variation across subjects for a given task. For our specific battery, this refutes not only Nash equilibrium but also the patches that have been designed to explain deviation from it.

2 Experimental Design

We report on a number of tasks here related to the well-known Trust Game [27]. In our experiments, subjects know that their choices are always private and anonymous, even to the experimenters at the time of the experiment (i.e., double blind). Subjects receive no feedback during the course of the experiment about the consequences of their choices, except for quizzes related to the given tasks (subjects may, for some of our tasks, be able to infer the consequences of their choices). For each task, subjects are randomly matched to another subject. Thus, to the extent possible, every task is a single shot, separate from the prior and future choices. Subjects are divided into two rooms, in groups of ten in each room. We ensure that no subject knows anyone else in either of the two rooms of the experiment. As much as possible, then, this environment creates a situation in which subjects derive their utility solely from the payoffs in the experimental tasks and not from concerns about reputation, signaling for future games, experimenter demand, or other actions that are not related to the immediate monetary payoffs we present.

The Trust Game (sometimes called the investment game) involves two players. Each player begins with a $5 endowment. The first player (Player 1 or the "Investor" [27]) chooses how many dollars, if any, to pass to an anonymous second player (Player 2 or "Trustee" [27]). In our experimental protocols, we use no labels other than "the other person(s)" (to avoid a gaming frame). To avoid suggesting an investment or reciprocity frame we label each action as a "transfer." The first player keeps any money he does not pass. As in [27], the money that is passed is tripled in value and the second player receives the tripled amount. The second player at that point has the original endowment of $5 plus three times the amount the first player passed, and decides how much, if any, of that total amount to return (i.e., transfer) to the first player. The second player at the moment of choice in the Trust Game is in a role that is equivalent to the role of Dictator in the classic Dictator Game [see 8]. As in the Dictator Game, the dominant strategy equilibrium, which is trivially the subgame perfect Nash equilibrium (SPNE), is that Player 2 will return $0. By backward induction, then Player 1 in the Trust Game will send $0. This is also a dominant strategy. As such, any amount sent by Player 1 to Player 2 should be viewed by Player 1 as a donation, and we label the first half of the Trust Game as what we call a Donation Game.

These equilibrium strategies derive from the assumption is that all players maximize their monetary payoff and that they believe that all other players do the same. In the Trust Game, a Player 1 with these beliefs would conclude that Player 2 will return nothing and so, as a maximizer, Player 1 sends nothing. The beliefs that players hold about other players lead to the belief at every level of recursion that all players will send $0, and that they will guess that others will send $0, and they will guess that others will predict that everyone will send $0, and so on, ad infinitum.

But what happens if a subject with these NE beliefs finds himself off the equilibrium path? In the Trust Game, only Player 2 could make a choice after finding himself or herself presented with an off-the-equilibrium-path choice. If Player 2 is gifted with anything more than his or her $5 endowment, the subgame perfect Nash equilibrium strategy is still to send $0 back.

Every subject makes decisions first as Player 1. They are then randomly paired with someone from the other room and they make choices as Player 2. So everyone gets to be Player 1 first, then Player 2, about 90 minutes later. To limit learning, even if it is just learning about the actions of a randomly assigned partner in another room, we defer the choices for all subjects as Player 2 to the end of the experiment. They thus play Trust twice, but in different roles. Player 1 never learns the consequences of any of his or her choices in the Trust Game. Player 2 can of course infer the consequences of his or her own choices.

We add elements to the basic Trust Game to tap into subjects' beliefs. Our belief elicitation mechanism borrows

from the idea of a prediction market [28], which in experimental settings such as those described here have been referred to as "scoring rules" [29, 30, and for a brief survey see 8]. We do not ask subjects to report their expectations, as some experimenters have done, in order to prod strategic thinking, rather, we ask them to "guess" other subjects' choices, or to guess other subjects' "predictions." As with all of our protocols, we try to provide little or no framing of the experimental tasks offered to our subjects. Only after Player 1 makes his choice about how much to transfer, do we ask him to guess how much Player 2 will return. We then elicit Player 1's recursive beliefs about Player 2. So, we next ask Player 1 to guess what Player 2 will later predict how much Player 1 is transferring. We further ask Player 1 to guess Player 2's prediction of Player 1's guess of how much Player 2 will return. We also elicit each subject's recursive beliefs when they are in the role of Player 2. Before Player 2 learns Player 1's choice, and thus before Player 2 knows how much they have available to them, we ask Player 2 to guess how much money Player 1 transferred. We also ask Player 2 to guess how much Player 1 predicted that Player 2 would guess that Player 1 transferred. After Player 2 learns Player 1's transfer, we ask Player 2 to guess how much Player 1 predicted she would return. All subjects know that all subjects earn $3 for each correct guess and earn nothing for a guess that is wrong. All subjects in our experiments know this. We also allowed subjects to, in essence, "double-down," on each guess, adding a second "bet" equal to $3 if they are correct in their guess and $0 otherwise. We also always quizzed our subjects with respect to the instructions, paying them for correct answers.

In calibrating these prediction questions prior to the launch of our experiments we learned two things: (1) that there does not exist an easy language for eliciting recursive beliefs, so we made use of generic cartoon "heads" to represent what subjects are predicting and "$" sign and arrow icons to represent actions and the object of their current attention; and (2) subjects laughed and failed to answer our queries, even when diagrammed in cartoon form, with written explanations. These two preliminary findings suggested to us that people really do not have the recursive beliefs required by NE.

The questions we ask vary slightly for each task, but as an example, here is the exact question we ask Player 2: "How much money do you guess the other person transferred to you? If you guess correctly, you will earn $3. If not, you will neither earn nor lose money." We add similar incentivized prediction tasks to various experimental tasks. Players do not learn whether their predictions were right or wrong and subjects never have any information about other subjects' guesses.

Subjects also make decisions in a variety of other games, including the already mentioned Dictator Game and what we call the Donation Game. In both these games, each subject is randomly paired with yet other subjects in the other room. In the Dictator Game, The Dictator (Player 1) and the Receiver (Player 2) have endowments identical to those the subjects had when they were in the role of Player 2 and Player 1 in Trust (although they have been randomly rematched and they known they've been randomly rematched). Accordingly, the Dictator Game was identical right down to the specific endowments to the second half of the Trust Game. In effect, each subject replayed the second half of the Trust Game, but now without the reciprocity frame. The SPNE is for the Dictator to send $0 to the Receiver. The Donation Game is identical, except that each player begins with a $5 endowment and the amount Player 1 chooses to send is quadrupled before it is given to Player 2 (making it roughly similar but not identical to the choice faced by Player 1 in the Trust Game, without the possibility of reciprocity). The dominant strategy and SPNE is again for the Donor to send $0.

The subjects in our experiment completed the tasks using pen and paper in a controlled classroom environment. Subjects were recruited using flyers and email and text messages distributed across a large public California university and were not compelled to participate in the experiment, although they were given $5 in cash when they showed up and signed in. A total of 180 subjects participated in this experiment. The experiment lasted approximately two hours, and subjects received on average $41 in cash. The experiment was followed sometime later with a questionnaire, for which subjects were also paid.

3 Result: Subjects' Beliefs in the Trust Game

Common patches to help explain the commonly observed departures to NE strategies (other-regarding preferences, cognitive constraints, decision-making biases, or equilibrium refinements) usually continue to maintain the assumption that players deviate from game-theoretic expectations in consistent ways. For example, if players prefer to reduce inequality, that preference should be stable across all manner of economic games [10]. Or, if players cannot perform backward deletion of dominated sub-games, as game theory requires, then this handicap should operate in all game environments of equal difficulty [17, 18, 19]. In this section we focus both on whether subjects have beliefs that are consistent with SPNE and whether their beliefs are consistent across tasks (regardless of alignment with SPNE). To date, there has been little focus on identifying the extent to which players have consistent beliefs or behavior across games.

Cognitive science gives us considerable reason to doubt that players will behave identically across different

environments, because changes in environment lead to changes in mental activation, which affects beliefs and behavior. As Sherrington famously wrote, the state of the brain is always shifting, "a dissolving pattern, always a meaningful pattern, though never an abiding one" [32]. If the particular tasks, and order of those tasks, induce different mental activations, then belief and behavior should vary accordingly. Our experiment is designed to shed light on whether subjects have consistent beliefs and make consistent choices.

In many of our tasks, we ask subjects to make guesses about other players' actions and predictions. Do subjects believe what game theory assumes they believe? The answer is that there is huge variance across what subjects believe in a single game and also huge variance within subjects from one task to another.

The SPNE in the Trust Game is that neither Player 1 nor Player 2 will send any money to the other. All should believe that all others will predict that no one will send money, and all such beliefs should be infinitely recursive, so that Player A believes Player B believes Player A believes Player B will send no money, and so on for any number of steps and for any subject in any role A or B.

But we see quite the contrary in our experiments: only 68 of 180 subjects as Player 2 believe that Player 1 will send nothing. In other words, 62% of subjects have "incorrect" beliefs, that is, beliefs contrary to those that support SPNE strategies.

Next we examine the guesses made by Player 1 of the amount Player 2 will return. In what follows, we include even the Player 1s who sent nothing. (Since Player 2 begins with a $5 endowment, Player 2 can transfer money even if Player 1 sent nothing.) Ninety-two of the 180 subjects guess that Player 2 will return $0, but 88, or 49%, believe that Player 2 will return some money. This means that 49% of these subjects have "incorrect" beliefs. Their beliefs diverge broadly from SPNE, across a large span of possible returns.

We can compare subjects' beliefs about others in one part of the Trust Game with their choices in that same part of the Trust Game. For example, we can examine the difference between what a subject choose to do as Player 1 in the Trust Game, and what they believe as Player 2 that Player 1 will do (and recall, when they are Player 2, they've already made choices as Player 1). The modal category is subjects believe that other subjects will play like them: 109 of the 180 subjects guess that the choice of the Player 1 with whom they are randomly matched will be the same as their own choice when they were Player 1. For these subjects, theory of mind might equal theory of self, or this may simply represent the "false consensus" effect in which people think others are more like them than they actually are [31], or it might be akin to the curse of knowledge, but we can't really tell. Perhaps most surprising, there is a large variance, with 71 subjects (39%) making guesses that differ from their own choices.

We can also examine the number of subjects who have beliefs consistent with NE across tasks. In the Trust Game, subjects make predictions as Player 1 about the behavior of Player 2 and as Player 2 about the behavior of Player 1. We already demonstrated that in either single task, a great many subjects do not have SPNE beliefs. If Player 1 has NE beliefs, it means that this subject guessed that Player 2 would return nothing. If Player 2 has NE beliefs, it means that the subject guessed that Player 1 would send nothing. Overall, out of 180 subjects in our analysis, only 63 subjects made guesses as both Player 1 and 2 that were consistent with NE beliefs. In other words, only 35% of our subjects have consistently "NE beliefs" *even inside this one game*.

There were 83 subjects who lacked NE beliefs in both part of the Trust Game, 29 subjects who possessed "NE beliefs" as Player 1 but not as Player 2, and only 5 subjects who possessed "NE beliefs" as Player 2 but not as Player 1. Our experiment does not allow us to identify why players' beliefs diverge from the NE beliefs, but it is clear that most subjects deviate from NE beliefs during at least one of the experimental tasks.

There were 60 subjects who were "fully Nash actors" in the Trust Game, that is, the subjects whose actions as both Player 1 and 2 were consistent with SPNE strategy. We examine whether these 60 subjects have beliefs that are "fully Nash" in the Trust Game. The answer is no. First, let us consider these 60 subjects in the role of Player 1 in Trust. Of these 60 subjects, 56 guessed as Player 1 that Player 2 would return nothing, which is consistent with SPNE. Only 40 of the 60 "fully Nash" Trust players (66%) guessed that Player 2 predicted that they would transfer $0. The other 20 of the 60 "fully Nash" Trust players (1/3rd) lacked that SPNE belief. 49 of the 60 also guessed Player 2's prediction of Player 1's guess of what Player 2 will return to be $0. These results show that even the 60 "fully Nash" Trust subjects hold beliefs whose degree of consistency with SPNE principles varies question by question even when we look at only those questions asked of them when they are in the role of Player 1. Beliefs show flexibility.

We next turn to the beliefs of those 60 "fully Nash" Trust subjects when they are in the role of Player 2 in Trust. Of the 60, 44 guess that Player 1 will transfer nothing; that is, 16 of 60 (27%) lack SPNE beliefs. Of the 60, 35 guess that Player 1 predicts that they will return nothing; that is, for this question, 42% of these 60 "fully Nash" Trust subjects have beliefs that are inconsistent with SPNE. Overall, non-SPNE beliefs are quite common even among the 60 "fully Nash" actors in the Trust Game. Beliefs show flexibility and refute the assumed beliefs of Nash equilibrium and the four patches often applied to NE.

4 Result: Inconsistency of Behavior in Trust, Donation, and Dictator

We turn next to examine the actions of subjects across a number of similar tasks to see if individual subjects behave consistently. In particular, we look at a set of tasks, all of which involve choosing how much money to transfer to another person and in some tasks the decision is not contingent on the other player. In the Trust Game, subjects play the role of Player 1 and 2 during the course of the experiment.

One way to investigate consistency of behavior is to examine the choices of subjects who as Player 2 in Trust received money from Player 1. Of the 100 subjects who received money as Player 2 in Trust, only 62 returned any of the money to Player 1. Additionally, of those 62, only 40 sent money in the Dictator Game, which is identical to the percentage of subjects who sent nothing in other "double blind" versions of the Dictator Game [13]. This shows that many subjects do not behave consistently in these two identical choice situations, in which their actions could reduce inequality. Further, of the 40 who sent money in Dictator, only 29 also send money in the Donation Game. This means that of the 100 subjects who received money as Player 2 in Trust, only 29 sent money to the other player in all three related tasks. This shows that the same subjects do not behave consistently even in their violations of SPNE.

Another example of apparently inconsistent behavior comes from examining the subjects who passed \$0 out of \$5 in the Donation Game, suggesting they are not concerned with others' earnings. Of the 87 subjects who passed \$0 in the Donation Game, 63 of them also passed nothing as Player 1 in the Trust Game, both behaviors of which are consistent with standard game theoretic behavior. However, the other 24 subjects passed \$0 in the Donation Game and some amount greater than \$0 in the Trust Game. What model predicts this behavior? Why would a player who passes nothing in the Donation Game pass money in the Trust Game? One possibility is that the player believes that passing money in Trust will result in greater earnings, because Player 2 will return enough money to make the choice to pass money financially beneficial. However, among these 24 subjects some guess they will earn money in the Trust Game and others guess they will lose money. The lack of consistency in these two very similar settings is further evidence that assumptions of consistent behavior are at odds with much human action.

Subjects deviate remarkably from NE strategies. We have reported how subjects' beliefs deviate from those necessary to support equilibrium strategies. We have also shown that these deviations are not consistent. Accordingly, it is doubtful that proposals to explain deviation from NE strategies will succeed if they presume a consistent mental or behavioral signature.

Now, we ask whether actions are minimally rational, that is, do subjects' actions accord with their beliefs? To begin, we investigate whether action and belief accord in the Trust Game. In our experiment, over one-half of subjects in the role of Player 1 (100 out of 180) pass a nonzero amount of money to Player 2 (this differs from [27]), which is inconsistent with a SPNE strategy, and on average subjects pass \$1.44 (which is not significantly different than the findings in [27]). Of the 100 subjects who receive money as Player 2, 62 of them return some money to Player 1. On net, Player 1 loses money (our results here are almost identical to those by [27]).

To look at the relationship between beliefs and actions we examine the difference between the amount Player 1 sends to Player 2 and the amount Player 1 guesses Player 2 will return. Recall that any money sent by Player 1 is tripled before it is sent to Player 2 and added to Player 2's initial \$5, (e.g., if Player 1 sends all \$5, then Player 2 has \$20, and if Player 2 splits that money, then Player 1 and Player 2 end with \$10 each, and we would say that each has "earned" \$5 through their actions). Overall, there are only a few players who guess that they will lose money by sending money to the other player. Mostly, players expect to break even or benefit slightly from their decision. The beliefs held by these players imply not only that they do not expect others to play consistently with SPNE strategies, but also that they expect, on average, to profit from their non-SPNE strategy to send money. But again, beliefs are not consistent across subjects.

There are 100 subjects who as Player 1 in Trust chose to send a positive amount to Player 2, and 20 of those players guess they will not receive anything in return. These 20 players guess that Player 2 will follow a SPNE strategy. These 20 subjects cannot simultaneously be maximizing their payoffs and hold the belief that Player 2 will follow a SPNE strategy of returning \$0 so it is hard to see how their choices accord with their own beliefs. We must either conclude that they are not payoff maximizers or relax the assumption that subjects act according to beliefs. One possible response is to give up the assumption that believing, preferring, deciding, and acting are coordinated mental events. Perhaps subjects act without fully activating their decisions, or believe without activating the consequences of those beliefs for action, or act without activating beliefs, and so on.

These results make it clear that we may not be able to simply observe behavior and then make correct inferences about the underlying beliefs that generated the behavior. When we observe a Player 1 in the Trust Game pass money what beliefs do we presume preceded that behavior? Is this a subject who is motivated by other-regarding preferences who does not expect to benefit financially? Or, is this a player who sincerely believes that

the 2nd player in the Trust Game will return enough money to make the initial decision profitable?

There were 60 subjects who were "fully Nash actors" throughout the game; that is, they chose SPNE strategies (i.e., $0) as both Player 1 and Player 2., We ask whether these 60 "fully-Nash" actors in Trust are "fully Nash" in the related Donation and Dictator Games. Of these 60 subjects, 57 pass $0 in the Dictator Game and 50 of the 60 pass $0 in the Donation Game. If we focus on those 57 subjects who are "fully Nash actors" as both Player 1 and Player 2 in Trust and also as Dictator in the Dictator Game, we find that 48 of the 57 pass nothing in the Donation Game. Therefore, across our entire subject pool, only 48 (27%) have consistent NE behavior in three related games of Trust, Donation, and Dictator.

Although deviations in a single game have been widely recognized, research has not focused on how behavior across games is related. We have shown in a variety of ways that NE-consistent behavior in one task does not guarantee similar behavior in another setting. Therefore, even accurately predicting a subjects' action in one setting is no guarantee that it is possible to predict accurately the subjects' action in another setting. We demonstrate thisusing game theoretic environments that are exceedingly similar, which would seem to stack the deck in favor of finding consistent behavior across games. However, the results from our battery of experimental tasks demonstrate that subjects regularly deviate from SPNE in both their beliefs and behavior, that the deviations are themselves inconsistent, and that there is variation in the degree to which behavior accords with belief. These deviations are so pervasive and the variation so large, even among subjects taking actions in similar or identical strategic settings, that is seems unwarranted to refer to them as "deviations." On the contrary, even though our subject pool is derived from subjects with very high math SATs, who were on the whole in the top 2% of high school graduates, consistent "NE behavior and beliefs" appear to be remarkable deviations from human cognitive patterns and human behavior.

5 Discussion

Games are defined by seven characteristics: players, actions, information, strategies, payoffs, outcomes, and equilibria, including equilibrium refinement [33]. Equilibria must be mutually consistent, indeed, "[T]he Nash equilibrium (NE) concept . . . entails the assumption that all players think in a very similar manner when assessing one another's strategies. In a NE, all players in a game base their strategies not only on knowledge of the game's structure but also on *identical conjectures about what all other players will do.* The NE criterion pertains to whether each player is choosing a strategy that is a best response to a shared conjecture about the strategies of all players. A set of strategies satisfies the criterion when all player strategies are best responses to the shared conjecture. In many widely used refinements of the NE concept, such as subgame perfection and perfect Bayesian, the inferential criteria also require players to have shared, or at least very similar, conjectures" [23: 103-104]. NE is, at its core, a cognitive theory.

In cognitive science, "theory of mind" refers to our amazing disposition to attribute mindedness to other human beings. Classical economics takes theory of mind for granted, and extends it to the view that all of those minds are driven by consistent preferences and beliefs to consistent actions. Behavioral game theory applies patches to come up with a conception of individual minds as consistently deviant from classical rationality. Hence the phrase "predictably irrational." The difference between classical rationality and behavioral game theory is not about consistency: classical rationality assumes that everyone is consistent in the same way, while behavioral game theory assumes that each person is consistent in a certain "deviant" way. While classical rationality and behavioral game theory are often taken to be opposed, we seem them as uniformly based on an assumption of consistency that does not stand with experimental test.

Some scholars justify consistency as a mathematical shortcut that is meant to represent the result of some unspecified learning, evolutionary adjustment process, or the adoption of social norms, laws or institutions [34]. These processes, however, are rarely defined. This line of reasoning also implies that beliefs and choices will not be consistent if players do not have time to learn or evolve.

As in many related experiments [for a survey see 8], subjects in our experiments do indeed deviate from SPNE predictions, both in their actions and in their beliefs. We also demonstrate that subjects' recursive beliefs (beliefs about the beliefs of other players) are often inconsistent with NE predictions, a result that has not been widely appreciated. We show further that the variance in actions and beliefs is very large.

Even the common patches of behavioral game theory do not predict the tremendous diversity that we observed for *individual subjects* and the variance across subjects' beliefs and behavior. Subjects' reported beliefs and their behavior are regularly incongruent – subjects who play consistently with NE prediction do not always possess the assumed game theoretic beliefs. Furthermore, subjects' beliefs differ across settings even when there are no changes in the cognitive complexity of the setting. Taken together these results suggest that game theoretic models and common modifications employed to make them explain the deviations from straightforward NE, do not accurately predict the variations in behaviors we observe in a laboratory setting. Therefore, research into decision-making should turn to discovering the cognitive patterns of decision-making.

Acknowledgments. McCubbins acknowledges the support of the National Science Foundation under Grant Number 0905645. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Turner acknowledges the support of the Centre for Advanced Study at the Norwegian Academy of Science and Letters.

References

[1] Pita, J., Kiekintveld, C., Tambe, M., Steigerwald, E. and Cullen, S. 2011. "GUARDS - Innovative Application of Game Theory for National Airport Security." In *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI).

[2] Roth, A.E. 1990. "New Physicians: A Natural Experiment in Market Organization," *Science*, 250, pp. 1524-1528.

[3] Kagel, J.H. and Roth, A.E. eds. 1997. *The Handbook of Experimental Economics*. Princeton University Press.

[4] Fudenberg, D., and J. Tirole. 1991. *Game Theory*. MIT Press.

[5] Ordeshook, P. C. 1986. *Game Theory and Political Theory*. Cambridge University Press.

[6] Nisan, N., Roughgarden, T, Tardos, E., and Vazirani, V. V. 2007. *Algorithmic Game Theory*. Cambridge University Press.

[7] Tirole, J. 1988. *The Theory of Industrial Organization*. MIT Press.

[8] Camerer, C. 2003. *Behavioral Game Theory*. Russell Sage Foundation, New York, New York/Princeton University Press, Princeton, New Jersey

[9] Kahneman, D., and Tversky, A. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, Vol. 47, No. 2. pp. 263-292.

[10] Rabin, M. and Thaler, R.H. 2001. "Anomalies: Risk Aversion," Journal of Economic Perspectives. vol. (1), pages 219-232

[11] Ainslie, G. 2001. *Breakdown of will*. Cambridge: Cambridge University Press.

[12] Elster, J. 1999. *Strong Feelings: Emotion, Addiction, and Human Behavior*. MIT Press.

[13] Hoffman E., McCabe, K., Shachat, K. and Smith, V. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7(3): 346-380

[14] Rabin, M. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review, LXXXIII*. 1281-1302.

[15] Simon, H. 1957. "A Behavioral Model of Rational Choice", in *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*. New York: Wiley.

[16] Camerer, C.F., T.H. Ho and K. Chong. 2004. "A cognitive hierarchy model of behavior in games," *Quarterly Journa of Economics.* Vol 119, No. 3, p. 861.

[17] Stahl, D. and Wilson, P.1994. "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior and Organization*, 25, 309–327.

[18] Costa-Gomes, M. A., and Crawford, V. P. 2006. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*. vol. 96(5), pages 1737-1768.

[19] McKelvey, R. and Palfrey, T. 1998. "Quantal response equilibria for extensive form games." *Experimental Economics*, 1, 9-41.

[20] Gigerenzer, G. 2000. *Adaptive thinking: Rationality in the real world.* New York: Oxford University Press.

[21] Turner, M. 2009. "The Scope of Human Thought." http://onthehuman.org/humannature/

[22] Rasmusen, E. 2006. *Games and Information.* Oxford: Blackweller Publishers, Fourth Edition.

[23] Lupia, A., Levine, A. S., and Zharinova, N. 2010. "Should Political Scientists Use the Self Confirming Equilibrium Concept? Benefits, Costs and an Application to the Jury Theorem." *Political Analysis* 18:103-123.

[24] Kuhlman, D. M., and Wimberley, D. L. 1976. "Expectations of choice behavior held by cooperators, competitors, and individualists across four classes of experimental game." *Journal of Personality and Social Psychology, 34*, 69-81.

[25] Croson, R. 2007. "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry,* Vol. 45, pp. 199-216.

[26] McKenzie, C. R. M., and Mikkelsen, L. A. (2007). "A Bayesian view of covariation assessment." *Cognitive Psychology*, 54, 33-61

[27] Berg, J. E., Dickhaut, J., and McCabe, K. 1995. "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, Vol. 10, 1995, pages 122-142.

[28] Wolfers, J. and Zietzwitz, E. 2004. "Prediction Markets." *Journal of Economic Perspectives.*

[29] McKelvey, R. and Page, R. T. 1990. "Public and private information: An experimental study of information pooling. *Econometrica*, 58, 1321-39.

[30] Camerer, C. and Karjalainen, R. 1994. "Ambiguity-aversion and non-additive beliefs in non-cooperative games: Experimental evidence." In Munier, B. And Machina, M. (eds.), *Models and Experiments on Risk and Rationality*. Dordrecht: Kluwer, 325-58.

[31] Ross, L., Green, D., and House, P. 1977. "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes." *Journal of Experimental Social Psychology*. 13, pp. 279-250.

[32] Sherrington, C.S. Sir. [1941] 1964. *Man on his Nature*. [The Gifford Lectures, Edinburgh, 1937–1938. New York: The Macmillan Co.; Cambridge: The University Press, 1941]. New York: New American Library.

[33] Rasmusen, E. 1989. *Games and Information: An Introduction to Game Theory*. Blackwell Publishers.

[34] Denzau, A. and North, D. C. 2000. "Shared mental models: Ideologies and institutions." In Lupia, A., McCubbins, M., and Popkin, S. *Elements of reason: Cognition, choice and the bounds of rationality*. Cambridge Univesity Press.