Nashbots:
How political scientists have underestimated human rationality, and how to fix it

Daniel Enemark (corresponding author)
Independent Research Consultant
Enemark Consulting
4534 Maryland St
San Diego, CA 92116
(650) 636-3224
denemark@gmail.com

Mathew D. McCubbins
Ruth F. De Varney Professor of Political Science and Professor of Law
Duke University
208 Gross Hall, Box 90204
Durham, NC 27708
(919) 660-4324
mathew.mccubbins@duke.edu

Mark Turner
Institute Professor and Professor of Cognitive Science
Case Western Reserve University
Crawford Hall 607
10900 Euclid Ave
Cleveland, OH 44106-7063
(216) 368-4753
turner@case.edu

Abstract: Political scientists use experiments to test the predictions of game-theoretic models. In a typical experiment, each subject makes choices that determine her own earnings and the earnings of other subjects, with payments corresponding to the utility payoffs of a theoretical game. But social preferences distort the correspondence between a subject's cash earnings and

her subjective utility, and since social preferences vary, anonymously matched subjects cannot know their opponents' preferences between outcomes, turning many laboratory tasks into games of incomplete information. We reduce the distortion of social preferences by pitting subjects against algorithmic agents ("Nashbots"). Across 11 experimental tasks, subjects facing human opponents played rationally only 36% of the time, but those facing algorithmic agents did so 60% of the time. We conclude that experimentalists have underestimated the economic rationality of laboratory subjects by designing tasks that are poor analogies to the games they purport to test.

Laboratory experiments in political science and economics have provided tremendous insight into human decision-making, and experimental results often call into question rational-choice theories. For example, experiments have shown that subjects use focal points rather than mixed-strategy equilibria when coordinating (Schelling 1960) and avoid losses in ways that violate expected utility theory (Allais 1953; Kahneman and Tversky 1979). Experiments play an increasingly important role in political science because they provide important evidence of causal effects, and can support or undermine the predictions of formal models. Many experiments confirm theoretical predictions, but some have contradicted prominent theories of resource management (Ostrom, Walker, and Gardner 1992), voting (Forsythe et al. 1996), committee decision-making (Guarnaschelli, McKelvey, and Palfrey 2000), legislative bargaining (Diermeier and Morton 2005), and international conflict (Tingley and Walter 2011; Quek 2016).

Camerer (1997) argues that these contradictions should lead us to "think of plausible explanations for [observed behavior] and extend formal game theory to incorporate these

explanations." But the problem with using experimental results to extend theory is that there are *two* possible reasons that those results might diverge from theoretical predictions. On one hand, game theory may lack predictive validity, in which case an extension is required. On the other hand, a laboratory task may lack construct validity—that is, it may not have the same players, actions, payoffs, or information as the theoretical model it is designed to test—in which case its results should not be used to modify or extend theory.

Perhaps the most common way laboratory tasks fail as analogies to theoretical games is in the construction of payoffs. Experimenters typically use cash payments (or lotteries) to represent the utility payoffs of a theoretical game, and ask subjects to make choices that determine their own earnings and the earnings of others. The problem with this method is that the utility payoffs in a game represent a complete accounting of all a player's motivations, whereas cash payments are not the only thing that motivates a laboratory subject—no matter how much money is on the line (Darai and Grätz 2010). When making choices that affect other people, most subjects have "social preferences" like the desire for fairness or generosity, and these social preferences change the payoff structure of the game.[1] Moreover, since humans vary in their social preferences, subjects cannot know their anonymous partners' payoffs with certainty, which in turn changes the information structure of the game.

We argue that the use of algorithmic opponents can mitigate the distorting effect of social preferences on the payoffs and information of experimental tasks. We pit human subjects against "Nashbots"—algorithmic agents programmed to employ equilibrium strategies—and this

[1] Subjects may also have preferences regarding cognitive effort, risk-taking, and compliance with perceived experimenter expectations. We take a number of measures to minimize these confounds in our experiments. (See Appendix A.)

reduces the distortion of social preferences in two ways. First, subjects should not be concerned about their opponents' earnings, since algorithms cannot actually own or use money. Second, subjects should not believe that there is any possibility that their opponents hold social preferences, since we explain clearly how the Nashbots maximize their own earnings and expect human players to do the same. Our experiments show that pitting subjects against algorithmic opponents increases by 66% the frequency with which their behavior matches game-theoretic predictions.

1. Using experiments to test rational-choice theory

Since game theory purports to model human behavior (Rasmusen 2006; Myerson 2013), it should be subject to experimental tests. Indeed, testing theory is one of the chief roles of experiments in political science (Roth 1995; Druckman et al. 2006; Palfrey 2008). Fiorina and Plott (1978) offered one of the first tests of game-theoretic predictions in their experimental study of decision-making under majority rule. Their finding—that game-theoretic models outperformed sociological theories of behavior—helped popularize the rational-choice paradigm.

Experiments that test formal theories pose this question: to what degree do individuals engage in the sophisticated, self-interested, strategic reasoning prescribed by game-theoretic solution concepts? To answer this question, we must observe individuals' behavior in an environment in which we perfectly understand the defining features of the game: players, actions, payoffs, and information (Rasmusen 2006). The greatest challenge in creating such an environment is that humans are influenced by a multitude of motivations and accounting for them is a daunting task. Moreover, to evaluate subjects' economic rationality, the experimenter must know precisely what information they have about the motivations of other players.

In game theory, predicting behavior is simplified by the assumption of utility payoffs that are both all-encompassing and (in games of complete information) common knowledge. A player's utility payoffs perfectly summarize his subjective experience of relative satisfaction with the various outcomes of a game. By definition, a player prefers any outcome of higher utility to any outcome of lower utility; if a player faces a decision that determines the outcome of a game and understands the consequences of that decision, then whatever choice he makes *must be* the choice of highest utility to him.[2]

If we design an experimental representation of the Ultimatum Game (UG) using oatmeal cookies as our operationalization of utility, then subjects who don't like oatmeal cookies aren't actually playing an UG. Similarly, in the cash-incented UG with a human opponent, subjects who value fairness over making a few extra bucks aren't playing an UG. Moreover, subjects who *do* like oatmeal cookies or prefer to maximize dollars but don't know whether their opponents have the same preferences are playing a game of incomplete information, so even these subjects aren't playing an UG.

When an experimenter simply translates utility into dollars, he sidesteps his responsibility to ascertain his subjects' preferences among all the possible outcomes of a task. For example, Frechette et al. (2000) use cash payments to represent the utility payoffs of Baron and Ferejohn's (1989) bargaining model of legislative equilibrium—essentially an iterated, multiplayer UG. They conclude "proposers consistently fail to allocate themselves anything close to what the

---

[2] This is even true if the player's decision requires him to burn cash. If a subject chooses to thwart his own advantage in an existentialist attempt to disprove rational egoism, the only flaw with the model is this: it fails to account for the fact that the utility of disproving rational egoism exceeds the utility of the cash he burns.

theory predicts." But a proposer in their cash-incentivized task may actually prefer taking less, because she values fairness or generosity; or she may believe that taking less will increase her expected cash payment, because her opponent might reject a low offer. (This belief is justified; in our experiments, an ultimatum proposer's expected earnings were 70% higher when he offered $2 rather than $1.) The experimental task Frechette et al. devised differs in both payoffs and information from Baron and Ferejohn's model, so it cannot test that model.[3]

Unfortunately, it is difficult to identify the effect of subjects' social preferences on their subjective utility. Psychologists have long understood that different individuals view the interests of others with greater or lesser concern (Van Lange 1999), and small changes in the framing of a task strongly influence subjects' apparent social preferences (Bardsley 2008; Lazear et al. 2012).

Because social preferences and beliefs are so difficult to measure, the best way to replicate the payoffs and information of a theoretical game in the lab is to remove social considerations entirely. To do this, we replace subjects' human opponents—towards whom they may feel sympathy or antipathy—with computer algorithms that cannot receive any actual money but are programmed to maximize their utility functions. Because humans are indifferent to the objectives of computer algorithms, utility when interacting with an algorithm should be free from social motives. That is, subjects should respond to their own financial incentives (a

---

[3] A colleague posed this concern: "if we are interested in how humans play a game against each other, if the models we use cannot predict that, they need to be changed." But most political scientists *aren't* interested in how undergraduates distribute cash among their peers—they're interested in statecraft and conflict. Theorists devise games to model these problems, and if a subject faces the same decision described by the model, it doesn't matter if his opponent is human. After all, many rational-choice models have non-human players like states or parties. And remember, if a subject has different payoffs or information than those in the model, he is playing a different game. No solution concept, applied to one game, can be expected to predict behavior in another game.

dollar equivalent of the theoretical payoffs), taking into account their opponents' likely actions but without sympathy for their plight. Subjects pitted against Nashbots are informed that on each task, the algorithm is designed to follow exactly one objective: maximizing its earnings for that single task, independent of any other task, while assuming that the subject will do the same.

2. Method

We designed an experiment to evaluate the construct validity of cash exchange between humans as a measurement of utility, and identify the proportion of subjects whose deviations from equilibrium behavior can be explained by the failure to account for social preferences. Of course, social preferences cannot explain all deviation from equilibrium; many strategic problems are simply too complex for untrained subjects to be able or willing to solve in the laboratory.[4] For this reason we present our subjects with simple tasks for which—absent social preferences—they are likely capable of finding an equilibrium. (We include paid quizzes on the instructions. Subjects answer correctly more than 90% of the time.) Our analytic strategy is to compare the behavior of subjects interacting with humans to the behavior of those interacting with algorithmic agents.[5]

In our experiment, 230 subjects participated in economic-exchange tasks—190 matched with human players sitting in another room, and 40 matched with algorithms running on a

---

[4] Ostrom et al. (1992), for example, report that even after subjects discussed their task for 10 minutes in groups of eight, "no group found the optimal solution."

[5] Johnson et al. (2002) used a similar method to argue that subjects do not use subgame-perfect equilibrium strategies, but their findings do not shed much light on the proportion of subjects whose deviations from equilibrium behavior can be explained by social preferences, for four reasons: they only studied one task, they picked a very difficult task, their computer treatment was confounded by learning, and they only report means, which are theoretically uninformative. As a result, their results are mostly uninformative in regard to this study.

computer in another room. We report on 11 experimental tasks set within five strategic scenarios. Each scenario served as an analogy to a theoretical game. Some scenarios had a second stage in which the subject would be informed of his opponent's choice before choosing an action. This paper excludes second stages (to preclude learning effects) and tasks for which there is no deterministic game-theoretic prediction (e.g. games with mixed-strategy equilibria).

In each of the following scenarios, subjects are paired with partners in another room—either a human or computer. Human opponents are randomly reassigned after each activity, and algorithmic agents retain no memory of previous activities. To measure beliefs, we pay subjects to predict their opponents' choices and allow them to bet on the accuracy of their predictions. Below in parentheses, we provide the relevant solution concept and its prediction. See Appendices for the full protocols, handouts, and game solutions.

1. **Donation:** The subject and her partner both begin with $5; the subject may transfer any amount of her $5 to the partner; the subject loses her transfer and her partner receives quadruple this amount. (Payoff maximization: transfer $0.)

2. **Prisoner's Dilemma:** The subject chooses whether to take $2 for himself at a cost of $3 to his partner. His partner faces the same choice. (Dominant strategy: take $2.)

3. **Ultimatum (Stage 1):** The subject splits $10 between herself and her partner. If the partner accepts the subject's split, both parties receive the amounts chosen by the subject. If the partner rejects, the $10 is lost. (Subgame perfect Nash equilibrium strategy: offer $1.)

4. **Sequential Chicken (Stage 1):** The subject chooses STOP or GO; his partner is informed of his choice, then chooses STOP or GO. If both pick STOP, both earn $4; if both pick GO

both earn nothing; if one picks STOP and the other GO, they earn $3 and $5 respectively. (Subgame perfect Nash equilibrium strategy: GO.)

5. **Trust (Stage 1):** The subject and her partner both begin with $5; the subject may transfer any amount of her $5 to her partner; the subject loses her transfer and her partner receives triple this amount. The partner will find out how much he received from the subject, and may return any amount to the subject. (Nash equilibrium strategy: transfer $0.)

    5.1. **Trust Predictions:** Subjects are asked to predict opponents' choices. Correct predictions earn $3, incorrect $0. T1 is the subject who acts first, choosing an amount to transfer; T2 is the one who acts second, choosing an amount to return. (Nash strategy for all three tasks: $0.)

        A. T2 indicates the amount he predicts T1 will transfer.

        B. T1 predicts the amount T2 predicts T1 will transfer.

        C. T2 predicts T1's prediction of the amount T2 predicts T1 will transfer.

    5.2. **Trust Confidence Bets:** Subjects may bet on the accuracy of their predictions. If they bet, they earn $2 if they were correct and lose $1 if incorrect. If they do not bet, they earn $0. (Nash strategy for all three tasks: having chosen $0 in the relevant Prediction, place bet.)

        A. T2 bets on the accuracy of his response to (5.1.A).

        B. T1 bets on the accuracy of his response to (5.1.B).

        C. T2 bets on the accuracy of his response to (5.1.C).

4. Results

Figure 1 shows that as expected, across all 11 tasks, subjects were more likely to follow game-theoretic predictions when playing against algorithmic agents. For nine tasks, the difference is statistically significant. Figure 2 shows the proportion of subjects making equilibrium choices in the two treatment conditions. On average, subjects make equilibrium choices 60% of the time when faced with algorithmic opponents, but only 36% of the time with human opponents. Removing social preferences increases by 66% the probability that a subject will act in accordance with the theory's predictions. The take-away point is that against Nashbots, subjects' behavior confirms game-theoretic predictions a majority of the time; against human opponents, their behavior contradicts those predictions a majority of the time.

[Figures 1 & 2]

The results support our claim that using algorithmic opponents strengthens the relationship between cash payments and utility. The Donation and Prisoner's Dilemma tasks test whether a subject is willing to maximize her dollar earnings at the expense of her opponents; the subject can choose unilaterally between outcomes, so if she doesn't choose to maximize her cash payment, either she doesn't understand the task or her payoffs diverge from the cash incentives. Thus, *every* subject who understands the task and values the outcomes as the model describes will choose the cash-maximizing outcome. When subjects are matched with Nashbots instead of humans, the proportions of subjects who violate theoretical predictions decrease by 43% and 64%, respectively.

The results also support our claim that using algorithmic opponents improves the relationship between experimental and theoretical information. Trust Prediction tasks A, B, and

C test whether a subject can predict his opponent's behavior based on the information he has about that opponent's incentive structure. When subjects are matched with Nashbots, the proportions of subjects who violate theoretical predictions decrease by 59%, 47%, and 55%, respectively. (Subjects were also more than twice as likely to make *correct* predictions in each of these three tasks.)

The remaining tasks are second-order consequences of the preferences and beliefs discussed above. If a subject believes her opponent will reliably maximize payoffs in the Trust Game, she should be willing to bet on her prediction (Trust Confidence Bets A, B, and C). If a subject believes her opponent will maximize payoffs *and* she prefers to maximize her own payoffs, she will play equilibrium strategies in stage one of the two-stage games (Ultimatum, Sequential Chicken, and Trust). The data show a decrease in non-equilibrium choices across all six tasks. The differences for Trust Confidence Bets A and B are significant, but the difference for Bet C is barely insignificant ($p = .06$). The differences for stage one of Trust and Ultimatum are significant, but the difference for Chicken is insignificant ($p = .85$). The slightly weaker evidence for these tasks, which require second-order reasoning, is not a surprise, since factors other than social preferences—such as cognitive limitations—may play a larger role in preventing subjects from making equilibrium choices (Simon 1955; Camerer et al. 1993).

Even without Nashbots, the frequency of equilibrium choices is higher in our experiments than in some of the classic papers. In Trust Stage 1, 45% of our subjects made equilibrium choices when facing a human opponent; in Berg, Dickhaut, and McCabe (1995) 6% did. In Ultimatum Stage 1, 6% of our subjects made equilibrium offers to humans; in Forsythe et al. (1994) none did. Given the variety of ways in which two experimental setups might differ, it's

difficult to identify the reasons for these discrepancies. But the fact that our subjects were more predisposed to employ equilibrium strategies even against humans makes the magnitude of the Nashbots effect especially impressive.

5. Conclusions

By ignoring social preferences, experimentalists have significantly underestimated the rationality of their subjects. Pitting humans against profit-maximizing algorithms eliminates social preferences and makes laboratory tasks better analogies to the games they purport to test, and we show that this approach causes a much larger portion of subjects to play as theory predicts. We recommend that experimenters employing economic exchange tasks include algorithmic opponents, at least as a robustness check, to account for the potential distortions caused by social preferences. One avenue for future research is to replicate with algorithmic opponents experiments that have been used to confirm or contradict formal theories of politics.

Just as experimenters have a responsibility to identify precisely the game that their subjects are playing, theorists have a responsibility to address the experimental evidence that social preferences distort utility functions. Many subjects do not value their financial self-interest enough for cash payments to induce an experimenter's intended preference ordering. And many subjects are not confident enough in the self-interest of *others* for a task with cash payments to induce the intended *beliefs*. In light of these experimental findings, what kinds of incentives can theorists convincingly claim are important enough to reliably induce preferences and beliefs? Or for what kinds of individuals is an incentive sufficient to induce preferences and beliefs? One of the most important tasks for a theorist is to specify the domain of his theory. Theorists attempting

to model politics need to ensure that their formal models have well specified and clearly justified domains.

Aldrich and Lupia (2011) have called for "research agendas that integrate experimental and formal modeling pursuits" to "improve the applicability and relevance of formal models." We agree, but when testing formal models, experimentalists must ensure that the tasks they design match the models they claim to test. Experiments showing altruism and envy may shed light on human psychology, but when experimenters cannot confidently identify players' preferences over the outcomes and beliefs about the other players' preferences and beliefs, they shed no light on formal theory. We believe that matching human subjects with algorithms is not just a useful manipulation for uncovering social preferences; it is the correct way to translate games to the lab—the only condition in which cash payments are a valid operationalization of utility.

6. References

Aldrich, J., and A. Lupia. 2011. "Experiments and game theory's value to political science." In *Cambridge Handbook of Experimental Political Science*, ed. Druckman, Green, Kuklinski, and Lupia. 88.

Allais, M. 1953. "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine." *Econometrica* 21 (4): 503–546.

Bardsley, N. 2008. "Dictator game giving: altruism or artifact?" *Experimental Economics*, 11 (2): 122-133.

Baron, D. P., and J. A. Ferejohn. 1989. "Bargaining in legislatures." *American Political Science Review*, 1181-1206.

Berg, J., J. Dickhaut, and K. McCabe. 1995. "Trust, reciprocity, and social history." *Games and Economic Behavior*, 10 (1): 122-142.

Brañas-Garza, P. 2007. "Promoting helping behavior with framing in dictator games." *Journal of Economic Psychology*, 28 (4): 477-486.

Camerer, C.F. 1997. "Progress in behavioral game theory." *Journal of Economic Perspectives*, 11 (4): 167-188.

Darai, D. and S. Grätz. 2010. "Golden Balls: A Prisoner's Dilemma Experiment. Socioeconomic Institute." University of Zurich, Working Paper No. 1006.

Diermeier, D., and R. Morton. 2005. "Experiments in majoritarian bargaining." In *Social Choice and Strategic Decisions,* eds. Austen-Smith and Duggan. Berlin: Springer, 201-226.

Druckman, J. N., D. P. Green, J. H. Kuklinski, and A. Lupia. 2006. "The growth and development of experimental research in political science." *American Political Science Review*, 100 (4): 627.

Fiorina, M. P., and C. R. Plott. 1978. “Committee decisions under majority rule: An experimental study.” *American Political Science Review*, 72 (2): 575-598.

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton. 1994. “Fairness in simple bargaining experiments.” *Games and Economic Behavior,* 6 (3): 347-369.

Forsythe, R., T. Rietz, R. Myerson, and R. Weber. 1996. “An experimental study of voting rules and polls in three-candidate elections.” *International Journal of Game Theory*, 25 (3): 355-383.

Frechette, G. R., J. H. Kagel, and S. F. Lehrer. 2003. “Bargaining in legislatures: An experimental investigation of open versus closed amendment rules.” *American Political Science Review*, 97 (2): 221-232.

Guarnaschelli, S., R. D. McKelvey, and T. R. Palfrey. 2000. “An experimental study of jury decision rules.” *American Political Science Review*, 94 (2): 407-423.

Johnson, E. J., C. Camerer, S. Sen, and T. Rymon. 2002. “Detecting failures of backward induction: Monitoring information search in sequential bargaining.” *Journal of Economic Theory*, 104 (1): 16-47.

Kahneman, D., and A. Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica*, 57, 263-291.

Lazear, E. P., U. Malmendier, and R. A. Weber. 2012. “Sorting in experiments with application to social preferences.” *American Economic Journal: Applied Economics*, 4 (1): 136-163.

Myerson, M. 2013. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Ostrom, E., J. Walker, and R. Gardner. 1992. “Covenants with and Without a Sword.” *American Political Science Review,* 86, 404–17.

Palfrey, T. R. 2008. “Laboratory Experiments.” In *Oxford Handbook of Political Economy,* eds. D. A. Wittman and B. R. Weingast. Oxford: Oxford University Press, 915-936.

Quek, K. 2016. “Rationalist experiments on war.” *Political Science Research and Methods*, 4 (2).

Rasmussen, E. 2006. *Games and Information*. Hoboken, NJ: Wiley-Blackwell.

Roth, A.E. 1995. “Introduction to Experimental Economics.” In *Handbook of Experimental Economics,* eds. J. H. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press, 3-98.

Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Simon, H. A. 1955. “A Behavioral Model of Rational Choice.” *The Quarterly Journal of Economics*, 69 (1): 99-118.

Smith, V. L. 1976. “Experimental economics: Induced value theory.” *The American Economic Review,* 66 (2): 274-279.

Tingley, D. H., and B. F. Walter. 2011. “Can cheap talk deter? An experimental analysis.” *Journal of Conflict Resolution*, 55 (6): 996-1020.

Van Lange, P. A. 1999. “The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation.” *Journal of Personality and Social Psychology*, 77 (2): 337.

Figure 1. Subjects more often comport with theoretical expectations under the Algorithmic-Opponent condition than under the Human-Opponent condition

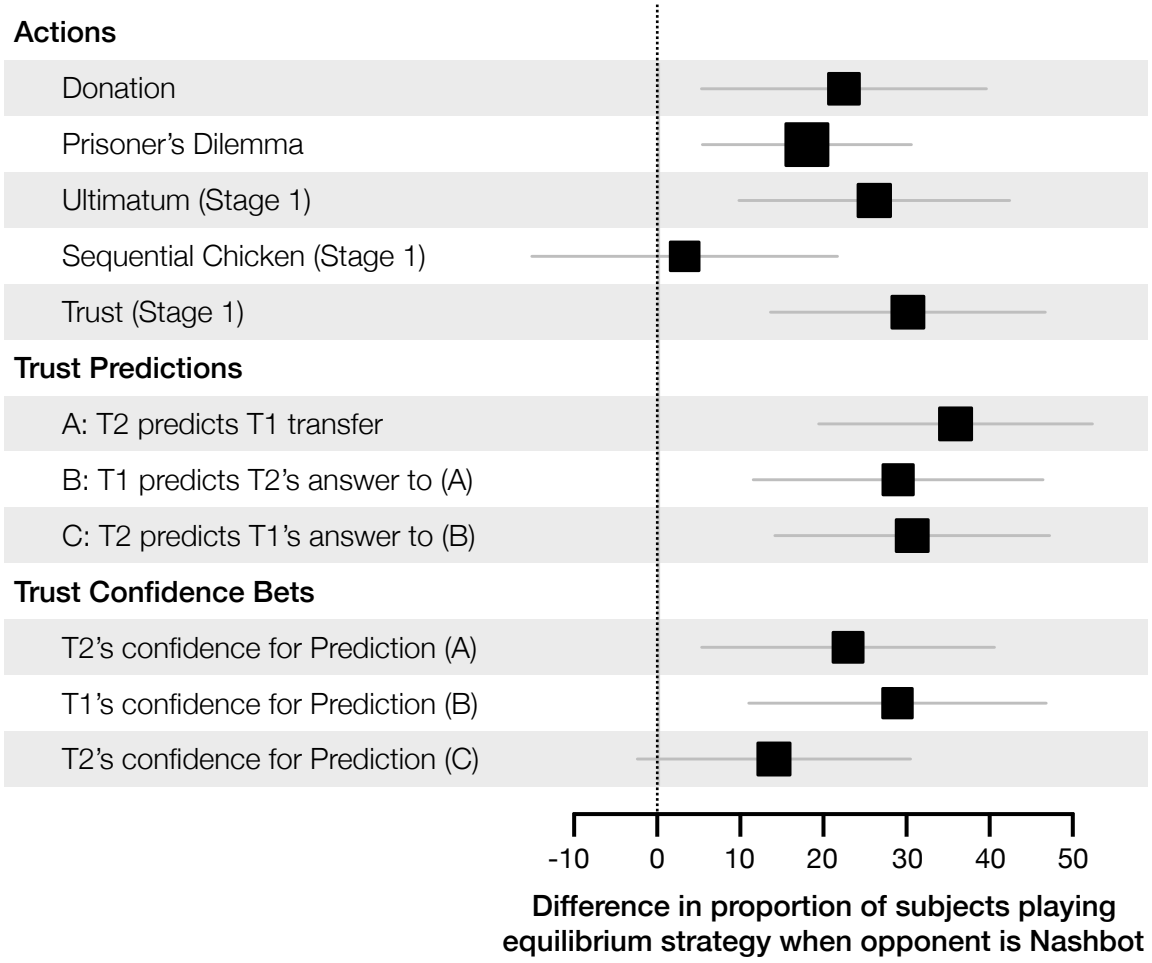

Figure 2. Proportion of subjects making equilibrium choices