Journal of Theoretical Biology 299 (2012) 172–179

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments

David G. Rand [a,b,c,n]

[a] Program for Evolutionary Dynamics, Harvard University, Cambridge MA 02138, USA
[b] Department of Psychology, Harvard University, Cambridge MA 02138, USA
[c] Berkman Center for Internet and Society, Harvard University, Cambridge MA 02138, USA

ARTICLE INFO

Available online 12 March 2011

Keywords:
Evolutionary game theory
Experimental economics
Cooperation
Internet
Economic games

ABSTRACT

Combining evolutionary models with behavioral experiments can generate powerful insights into the evolution of human behavior. The emergence of online labor markets such as Amazon Mechanical Turk (AMT) allows theorists to conduct behavioral experiments very quickly and cheaply. The process occurs entirely over the computer, and the experience is quite similar to performing a set of computer simulations. Thus AMT opens the world of experimentation to evolutionary theorists. In this paper, I review previous work combining theory and experiments, and I introduce online labor markets as a tool for behavioral experimentation. I review numerous replication studies indicating that AMT data is reliable. I also present two new experiments on the reliability of self-reported demographics. In the first, I use IP address logging to verify AMT subjects' self-reported country of residence, and find that 97% of responses are accurate. In the second, I compare the consistency of a range of demographic variables reported by the same subjects across two different studies, and find between 81% and 98% agreement, depending on the variable. Finally, I discuss limitations of AMT and point out potential pitfalls. I hope this paper will encourage evolutionary modelers to enter the world of experimentation, and help to strengthen the bond between theoretical and empirical analyses of the evolution of human behavior.

1. Introduction

Evolutionary game theory uses mathematics to formalize the process of evolution (Hofbauer and Sigmund, 1998; Nowak, 2006b). One area which has received a great deal of attention is the evolution of cooperation (Nowak, 2006a), with a particular emphasis on human cooperation. How can we explain the fact that people are often willing to help others, even at a cost to themselves? Answering this question is of great interest and importance for scientists across a wide range of disciplines. Evolutionary game theory has provided many deep insights in this area by combining biological and economic approaches.

When exploring the evolution of human cooperation, there are several possible levels of analysis. One extreme involves producing analytical solutions for a given model, providing deep understanding of the model's behavior and its dependence on parameter values. In order to generate analytical solutions, however, typically models must either be very simple or extreme limiting assumptions must be imposed. A second level of analysis involves computer simulations. Here models can be of arbitrary complexity and limiting assumptions may be relaxed, but there is a trade-off: the results are not general, and instead apply only to the particular parameter values simulated.

While the interplay between mathematical and computational models has been the focus of much of evolutionary game theory, a third level of analysis has recently been gaining in popularity: human behavioral experiments. Similarly to the way in which computer simulations add a layer of complexity to analytically tractable models, behavioral experiments add the element of human psychology. Using behavioral experiments built around economic games, researchers can have human subjects interact in precisely the same way agents interact in the theoretical models. By comparing the predictions of evolutionary models with observed behavior in the laboratory, we can gain insight into human evolution.

Yet few evolutionary game theorists conduct behavioral experiments. This fact is not particularly surprising: conducting experiments requires a very different set of resources and skills from those required for theoretical studies. To run behavioral experiments, researchers need access to a laboratory (usually with a large number of networked computers) in which to run the experiments, an active

[n] Correspondence address: Program for Evolutionary Dynamics, Harvard University, Cambridge MA 02138, USA.
E-mail address: drand@fas.harvard.edu

0022-5193/$ - see front matter & 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.03.004

and well maintained subject pool from which to recruit participants and a large research budget from which to pay them, as well as the wherewithal to physically interact with participants, read instructions aloud, pass out instructions, count out and distribute cash payments, etc. These practical barriers to entrance have largely prevented theoreticians from crossing into the world of behavioral experiments, and have discouraged many of those with a theoretical background from developing experimental designs to testably differentiate between model predictions.

With the advent of online labor markets, however, many of these barriers have been dramatically reduced. Online labor markets use the internet to connect employers with potential workers, perform some vetting of worker credentials and facilitate easy transfers of payments (Horton et al., in press). One of most popular online labor markets as of 2011 is Amazon Mechanical Turk (AMT), where most jobs are short (less than five minutes) and pay small amounts of money (less than $1). Researchers can easily use AMT to recruit subjects for incentivized behavioral experiments using economic games, where earnings depend on the subjects' decisions in the game. The entire process occurs over the computer, making the experience a familiar one for theorists—conducting experiments over AMT feels very similar to running computer simulations. Thus online labor markets open the world of experimentation to theorists.

In Section 2, I discuss the interaction between theory and experiments to date in the context of human cooperation. In Section 3, I introduce online labor markets and describe how they can be used to conduct such cooperation experiments. In Section 4, I provide an overview of evidence that data from online labor markets is comparable to data collected in the physical laboratory, and present two new experiments exploring the accuracy and consistency of self-reported demographic data from Amazon's Mechanical Turk. In Section 5, and give practical advice for running experiments using the internet. In Section 6, I conclude.

2. Theory, experiments and human cooperation

Evolutionary game theoretic models give insight into what behaviors can be favored by natural selection. Behavioral experiments explore how people actually behave. The integration of models and experiments goes in both directions, with models motivating particular experimental designs, and data from experiments being used to refine theories and inspire novel models.

In some cases, general insights from evolutionary modeling influence the design and interpretation of experiments. For example, a great deal of theoretical work has emphasized the importance of reputation (Kandori, 1992; Kreps and Wilson, 1982; Nowak and Sigmund, 2005; Ohtsuki and Iwasa, 2006) and repetition (Axelrod, 1984; Axelrod and Hamilton, 1981; Fudenberg and Maskin, 1986; Fudenberg and Maskin, 1990; Nowak and Sigmund, 1993; Nowak and Sigmund, 1992) in the evolution of human cooperation. Yet most experiments on cooperation (usually performed by economists) have focused on one-shot anonymous games (see Camerer, (2003) for an overview). A number of more recent studies have been undertaken re-evaluating the conclusions from one-shot experiments in light of the salience of reputation and repetition. These studies have revealed that often conclusions from one-shot games either (i) do not hold in games where there are future consequences for your actions today, for example involving the effectiveness of punishments and rewards for promoting cooperation (Dreber et al., 2008; Milinski et al., 2002; Nikiforakis, 2008; Rand et al., 2009b; Rockenbach and Milinski, 2006) (see Appendix A for a detailed discussion of this issue), or (ii) are better interpreted in the context of reciprocity and reputation, for example subjects' extreme sensitivity to subtle cues suggesting that they are being watched (Burnham and Hare, 2007; Burnham, 2003; Haley and Fessler, 2005).

Models can also motivate specific experimental designs to test theoretical predictions. Theoretical work on direct reciprocity (Axelrod, 1984; Axelrod and Hamilton, 1981; Fudenberg and Maskin, 1986; Fudenberg and Maskin, 1990; Nowak and Sigmund, 1993; Nowak and Sigmund, 1992) led to a number of experiments showing the power of (stochastic) repetition for promoting cooperation and exploring the specific strategies used in repeated games (Dal Bó, 2005; Dal Bó and Fréchette, 2011; Fudenberg et al., in press; Wedekind and Milinski, 1996). Indirect reciprocity models inspired experiments showing how reputation can promote cooperation and exploring the social norms used by human subjects (Milinski et al., 2002; Milinski et al., 2001; Seinen and Schram, 2006; Semmann et al., 2005; Ule et al., 2009; Wedekind and Milinski, 2000). Models exploring cooperation in optional public goods games (Hauert et al., 2002) motivated experiments demonstrating the cyclic dominance predicted theoretically (Semmann et al., 2003). The connection between evolutionary models and experiments can be quantitative as well as qualitative: stochastic evolutionary game theory models can sometimes quantitatively reproduce the human behavior observed in experiments where classical game theoretic approaches cannot (Rand et al., 2009a).

Experiments do not always confirm the predictions of theoretical models. In such cases, which are often the most valuable for theorists, experiments can inspire new modeling directions. For instance, recent experiments have raised questions about the ability of spatial structure to promote cooperation in the laboratory (Grujić et al., 2010; Suri and Watts, 2010; Traulsen et al., 2010). Given that evolutionary analysis typically involves steady-state/equilibrium behavior, one might wonder whether the mismatch between these (or any other) experimental and theoretical results is caused by the experiments not having been run for a sufficiently long time. One way to gain insight into this issue is to examine trends over time: if behavior in the experiments has stabilized, as is the case in these experiments on spatial structure, it seems unlikely that running the experiments for longer would change the outcomes. Instead, it has been suggested that this discrepancy is caused by the fact that learning and strategy updating within an experimental session may be characterized by extremely high rates of experimentation, which undermine the effects of spatial structure (Traulsen et al., 2010). These results have emphasized the importance of considering evolutionary dynamics with high mutation rates (i.e. 'exploration dynamics') (Traulsen et al., 2009), as well as devising experimental methods for distinguishing between simple strategies 'mutating' and more complex strategies that are probabilistic and/or conditional.

Another example is given by a collection of recent experiments that have raised questions about the role of costly punishment in promoting cooperation. While a large body of literature in evolutionary game theory has explored the co-evolution of punishment and cooperation, almost all of these models have assumed that only cooperators might punish defectors. 'Anti-social punishment' targeted at cooperators has been excluded *a priori*. Yet experiments show that defectors may retaliate when punished in repeated games (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Dreber et al., 2008; Nikiforakis, 2008; Wu et al., 2009), that defectors may punish cooperators even in the absence of the possibility for retaliation (either in one-shot games or games where identities are shuffled between rounds) (Gächter and Herrmann, 2009; Gächter and Herrmann, in press; Herrmann et al., 2008) and that people will even pay to punish each other in

the absence of any motivations whatsoever (for example, in the so-called 'joy of destruction' game) (Abbink and Sadrieh, 2009; Abbink and Herrmann, 2011). In response, new models have been introduced re-examining the ability of punishment to promote cooperation when the strategy is not arbitrarily reduced to exclude anti-social punishment (Janssen and Bushman, 2008; Rand et al., 2009a; Rand et al., 2010). Developing a more well-rounded understanding of punishment that includes spite, revenge and dominance as well as the more 'traditional' focus of norm enforcement is an important direction for subsequent modeling work.

Experiments may also suggest entirely new paradigms not previously considered by evolutionary theorists. An example is given by fascinating experiments that have explored endogenous choice, where participants pick one of several games to play (Gurerk et al., 2006; Rockenbach and Milinski, 2006; Sutter et al., 2010). For example, a public goods game with or without sanctioning opportunities: here sanctions have a dual purpose. Because of the endogenous choice, they serve to sort subjects by type as well as creating incentives. A first step in the direction of studying evolution in this type of scenario is represented by a recent model of pool punishment (Sigmund et al., 2010), a form of institutional punishment introduced early in the experimental study of punishment (Yamagishi, 1986). Further models, where participants in each institution are insulated from one another (as in the endogenous choice experiments), are a promising direction for future research.

Because of this interplay between theoretical and empirical investigations of human behavior, there is great value in having theorists also conduct experiments. In the next section, I introduce online labor markets and explain how they can facilitate this cross-over between theory and experimentation.

Before proceeding, however, it is important to point out that when integrating evolutionary models and human behavioral experiments, one must be mindful of the complex range of strategies employed by humans. Most models from evolutionary biology tend to consider unconditional agents who either always cooperate or always defect, and who then update their choice through an evolutionary process. While these extremely simple strategies may be reasonable descriptions of the behavior of some non-human animals, human cognition almost certainly involves more complex strategies. For reviews of animal versus human cognition in the context of cooperation, see Brosnan et al. (2010), Melis and Semmann (2010). In games with repetition or reputation, these more complicated strategies explicitly condition their behavior on previous events (Dal Bó and Fréchette, 2011; Fudenberg et al., in press; Wedekind and Milinski, 1996; Wedekind and Milinski, 2000). But even in one-shot anonymous games, human agents might employ conditional strategies that depend on their expectations and beliefs about how others will act. An often discussed example from the experimental economics literature is the idea of people being 'conditional cooperators' in one-shot settings, such that they will cooperate as long as they expect the other player(s) to also cooperate (Fischbacher et al., 2001). Such players could be considered 'altruistic,' but nonetheless might wind up defecting much of the time depending on their beliefs. Similarly, selfish individuals might behave prosocially depending on their beliefs about others: for example, 'fair' offers in the Ultimatum Game can largely (or perhaps entirely) be explained by the belief that lower offers will be rejected (Roth et al., 1991). Thus examining behavior is often not enough to characterize people's actual altruistic versus selfish preferences and tendencies.

There are several approaches employed by social scientists to address this issue. One method is to explicitly assess beliefs as well as behaviors in the experimental design. Here, subjects are asked what they expect others to do, as well as making one or more behavioral decisions. These decisions can then be interpreted in the context of the associated beliefs. Eliciting beliefs is potentially challenging, as subjects may not give truthful reports. A solution is to add additional incentives, such that subjects are paid extra if their reported beliefs are correct (Blanco et al., 2010; Croson, 2000; Gächter and Renner, 2010; Prelec, 2004).

Another approach is to remove the role of beliefs by allowing subjects to condition their behavior on the behavior of others (Fischbacher et al., 2001), typically using the 'strategy method.' Here subjects indicate their behavior in each of a number of possible situations, and are paid based on the decision indicated for the situation that actually occurs. For example, conditional cooperation can be assessed using an alternating Prisoner's Dilemma where subjects indicate their decision (C or D) if (i) they are the first mover, (ii) they are the second mover and the first mover chooses C and (iii) they are the second mover and the first mover chooses D. All subjects are then randomly assigned a role, and second mover decisions are determined based on the behavior of the respective first mover. Without such a design, one cannot tell whether a first mover's C is motivated by altruism or a self-interest coupled with the belief that C will elicit reciprocal cooperation from the second mover; or whether a first mover's D is motivated by self-interest or the belief that the second mover will not reciprocate.

A third possibility is to compare behavior across multiple games (Dreber et al., 2011; Harbaugh and Krause, 2000). For example, to determine the extent to which behavior in a given setting is motivated by altruistic concerns, one might have subjects play the game of interest followed by a Dictator Game (Player 1 chooses how to split a sum of money between herself and an anonymous recipient). In the Dictator Game, giving nothing is payoff maximizing regardless of one's beliefs about the recipient; thus giving away any money is clearly indicative of some other-regarding preferences. If behavior in the game of interest is correlated with giving in the Dictator Game, that suggests an altruistic motivation.

3. Online labor markets

To conduct experiments using economic games, researchers must be able to recruit a sufficiently large number of subjects; to provide them with instructions, make sure they understand the rules of play and then collect their decisions (usually using a large number of networked computers); and to pay them according to their earnings in the study. Accomplishing this in the physical world can be quite arduous, time-consuming and expensive. Many schools have numerous full-time staff members whose sole responsibility is managing the logistics necessary for running behavioral experiments; and many researchers spend a large portion of their time applying for grants to fund behavioral experiments.

Thanks to the internet, it is now possible for theorists to satisfy all of these requirements and conduct experiments as easily as running a computer simulation. In recent years, a number of online labor markets have arisen. These labor markets use the internet to connect employers with potential workers, who are paid to complete tasks on the computer. As in traditional labor markets, payment is conditional on satisfactory completion of the job, and workers often receive bonus pay based on how well they complete the task. The labor market typically handles all payment details—employers make a lump-sum transfer of money into an account, and then indicate how much each worker should receive. The labor market website takes care of the rest.

Thus online labor markets are ideal for conducting incentivized behavioral experiments. Researchers act as employers,

hiring workers to participate in experiments. The baseline payment corresponds to the 'show-up fee' typically paid to subjects just for coming to the experiment, and then workers earn additional bonus payments based on their actual decisions in the game (and the decisions of the others they interact with). There are few logistical issues for the experimenter to be concerned with. Furthermore, much lower payments can induce online workers to participate compared to subjects in the traditional lab, because the time investment costs are much lower (no need to physically come to the lab space, spend time and money on transit, etc.). An additional benefit of online experiments is that by logging subjects' IP addresses, one can get accurate information about certain interesting demographic factors that subjects themselves might not even be aware of, such as housing density, crime rates and rainfall in the area where each subject is located.

As the entire process takes place over the computer, running experiments using online labor markets is an easy transition for theorists to make. The experience is quite similar to running a computer simulation on a supercomputing cluster: the researcher designs the experimental instructions (analogous to planning out the program), creates a survey website or web applet through which participants read instructions and indicate their decisions (analogous to writing the code) and uploads the experiment to the labor market as a job posting (analogous to submitting a run to the cluster). The experiment is then automatically advertised to (typically) thousands of workers, and as workers complete the experiment, data accumulates and is downloaded by the researcher. Once the experiment is over, the researcher calculates each participant's earnings, and uploads the payment information to the labor market.

The most straightforward experimental designs to implement in this way are those that do not require feedback. With such designs, researchers can collect the decisions of all subjects, and then once the experiment is over, match subjects up to determine payoffs (''ex-post matching''). Thus subjects who interact and affect each other's payoffs need not be present at precisely the same time, and no sophisticated software for simultaneous interaction is needed. For example, in one-shot symmetric games such as the Prisoner's Dilemma, each subject makes their decision (cooperation or defect). Then once all data is collected, subjects are randomly paired, and payoffs are calculated based on what each subject indicated. For more complicated games, the 'strategy method' can be used, in which subjects indicate how they would act at each node in a decision tree. These designs take advantage of the fact that in many online labor markets, the employer has a grace period of at least several days from the completion of the job before bonuses must be paid. This gives time to implement the ex-post matching.

As of 2011, the most active online labor market for conducting behavioral experiments is Amazon Mechanical Turk (AMT). On AMT, workers are usually paid small amounts for short tasks (generally less than $1 for less than 5 minutes of work). AMT workers are from all over the world, with the majority of workers being from either the United States or India. Spending less than $1 per person is it possible to collect data from over 1000 subjects in only one or two days using AMT. For a discussion of the details of running studies on AMT, see Mason and Suri (2010).

4. Replications using mechanical turk

When using the internet to conduct experiments, researchers have much less control over their subjects than in the physical lab. Participants might not be paying close attention; multiple people might be making a single set of decisions together; participants can easily leave in the middle of the experiment for whatever reason; participants may answer self-report questions untruthfully; and a single subject might have multiple online identities and thus be able to participate multiple times in a single experiments (although note that online labor markets go to lengths to prevent multiple IDs). One might also be concerned that the subjects recruited through AMT differ dramatically from those recruited in the physical laboratory. Proponents of AMT maintain that the ability to quickly recruit much larger numbers of subjects can compensate for the noise induced by these factors. But before AMT can be used with confidence as an experimental platform, it must be shown that this is true and that these potential issues do not compromise data gathered on AMT. In this section, I survey an initial series of replication studies showing that data from AMT is consistent with data from the physical lab. I also present two new studies exploring the accuracy of self-reported data from AMT. The accumulation of even more replications and validation experiments will play an important role in establishing online experiments as an important part of the empirical toolkit.

Most compelling are direct replication studies in which the exact same experiment is run both in the physical lab and on AMT. Horton et al. (in press) conducted a one-shot Prisoner's Dilemma experiment both offline ($N=30$) and online ($N=155$). Both experiments used identical instructions, except for stake size: subjects in the physical lab earned between $3 and $10 in the game, while AMT subjects earned between $0.30 and $1.00. The results emphasize the importance of making sure subjects understand the instructions. After reading the instructions, but prior to making their decisions, subjects were asked a series of questions about the payoff structure. This allows a comparison of those AMT subjects who did and did not understand the experimental setup. They found that the cooperation level among the 74 AMT subjects who got all comprehension questions correct was almost identical to what was observed in the physical lab (physical lab, 37% C; online, 39% C; Chi^2 test $p=0.811$). Among the 81 AMT subjects who did not get all the questions correct, however, cooperation was much higher, 54%, and roughly equal to chance. Furthermore, Horton et al. asked two different kinds of comprehension questions: qualitative questions about what decision was better or worse for each player, and quantitative questions where participants had to perform detailed payoff calculations. Their data show that nearly all subjects (88%) who answered the qualitative questions correctly also performed the payoff calculation correctly, and the results of the experiment are virtually identical if only excluding based on the qualitative questions. Thus qualitative comprehension questions may be sufficient to ensure subjects understand the essence of the game. Horton et al. also conducted two qualitative replications that reproduce classical lab results. The first demonstrated the framing effect by showing that the language with which a decision problem is posed can change behavior. The second demonstrated priming by showing that exposure to an unrelated stimulus prior to the decision problem can change behavior (in this case, reading a religious passage about charity increased cooperation in a Prisoner's Dilemma among subjects that believe in God).

A second direct replication using economic games was conducted by Suri and Watts (2010). By designing a more complex, specially-built set of web software tools, they were able to implement repeated play online. They recruited a large number of subjects simultaneously from AMT, and redirected them to their external game website. In their replication experiment, they had $N=96$ subjects play the same linear repeated public goods game of Fehr and Gächter (2000). Subjects played a 10-round game in groups of 4, and received information after each round about the decisions of the other group members. Suri and Watts find

quantitative agreement between contribution levels among their AMT subjects and the $N=40$ subjects in the physical lab study, within each of 10 rounds of play. They also show that on AMT, halving the stakes from 1 cent per unit earned in the game to 0.5 cents per unit earned does not significantly affect behavior.

A third direct replication used various social psychology tests and compared $N=318$ AMT subjects to both $N=141$ students from a Midwestern University and $N=137$ visitors to an online discussion board (Paolacci et al., 2010). They found no difference across experiments in either attentiveness (measured by correctly answering a trivial question) nor basic numeracy skills (measured by the Subjective Numeracy Scale (Fagerlin et al., 2007)). They also found very similar effect sizes in a task examining framing effects (as in Horton et al. (in press)), and in two classic psychological effects, the conjunction fallacy and outcome bias.

A large number of qualitative replications, demonstrating that well-known psychological effects are also present among AMT workers, have been conducted and presented online in various AMT-related blogs. It has also been shown that AMT subjects are much more diverse and representative of the US population than the usual convenience sample of college undergrads, and that AMT subjects display a similar level of consistency across a battery of personality test questions as typically seen in other experiments (Buhrmester et al., 2011).

4.1. New experiments on self-report demographics

To explore the truthfulness of answers to self-report survey/demographic questions, I conducted two new AMT experiments for this paper. In the first experiment, I took advantage of the fact that a subject's location can be independently verified by logging their IP address. I recruited $N=176$ AMT workers to complete a short demographic questionnaire, including a question asking their country of residence. I also logged each subject's IP address, and determined the country corresponding to that IP. I found that the self-reported country of residence matched the country implied by the IP address in 97.2% of subjects. Thus at least for country of residence, self-report accuracy was very high on AMT.

For self-report questions other than country of residence, it is not possible to directly verify responses. However, one can get some insight into reliability by examining the consistency of responses of a given worker across multiple studies. In the second experiment, I examined the data from two separate studies conducted some time apart, one of which involved 1920 workers and the other of which involved 1222 workers. Among the $N=100$ workers who had participated in both studies, I found that 96% of subjects reported the same gender in both studies; 93% of subjects reported within 1 year of the same age in both studies; 98% of subjects reported the same country of residence in both studies; 81% of subjects reported the same education level in both studies; 82% of subjects reported within one bracket of the same yearly income level in both studies (with brackets of $<\$5$k, \$5–\$10k, \$10–\$15k, \$15–\$25k, \$25–\$35k, \$35–\$50k, \$50–\$65k, \$65–\$80k, \$80–\$100k and $>\$100$k); and 84% of subjects reported within one point of the same strength of belief in God in both studies (using a 10 point Likert scale). Thus there is some variation in reliability across questions, but even the less reliable questions are fairly consistent, and clearly indicate that most subjects are not merely making random selections.

5. Limitations, potential pitfalls and words of caution

AMT offers an extremely powerful new tool for conducting behavioral experiments. Yet as with any experimental method, there are various limitations of AMT that are important to keep in mind. Most obvious is the limitation on what kinds of experimental designs can be implemented using AMT. Firstly, only experiments conducted entirely over the computer are possible. For example, experiments that correlate behavior with hormone levels (Apicella et al. 2008) or genes (Dreber et al. 2009) cannot be conducing on AMT. It is also technically somewhat difficult, although not impossible (Suri and Watts, 2010), to conduct repeated games on AMT, or any other type of game, which requires multiple participants interacting with real-time feedback. Because you cannot be sure about exactly what subjects are doing while completing the experiment, designs that require complete control of subjects' attention (such as those using cognitive load manipulations) are also not practical on AMT. And because subjects are not all physically together experimenters cannot create absolute confidence in common knowledge (i.e. completely convince subjects that all participants receive the same information). A related issue is general participant trust in the experimental instructions. In economic game experiments in particular, it is critical that subjects believe they will be paid as described by the experimenter. To explore this issue, Horton et al., in press ran a survey of participant attitudes, and found that AMT subjects were only slightly less trusting than physical lab subjects.

A potential issue that might not be immediately obvious is non-random attrition. On AMT, it is very easy for subjects to quit mid-experiment. Thus if some treatments are more difficult or unpleasant than others, subjects may be more likely to drop out. In this case, a confound is introduced into the experiment: the two treatments differ not only in the experimental manipulation, but also in the pool of subjects who participated. Researchers must be vigilant to ensure drop-out rates are similar across treatments. For example, a potentially off-putting or time-consuming manipulation could be included in both experimental treatments—in one treatment before the decision task of interest, and in the other treatment after the decision. Thus attrition rates will be similar, but the manipulation will only affect the decision in one treatment.

As described in the previous section, ensuring that subjects understand the instructions is critical. This issue poses a larger challenge on AMT than in the physical lab, for several reasons. First, the experimenter is not present and cannot answer any questions subjects might have. Second, many subjects are paying substantially less attention on AMT compared to the lab. And third, many subjects are not native English speakers, or generally have lower English competence than college undergraduates. AMT compensates for these comprehension issues by allowing experimenters to recruit very large numbers of subjects. One easy solution to this issue is to include detailed comprehension questions, and then exclude from analysis (and deny payment to) subjects who do not answer correctly. Also, of course, researchers should make every effort to make their instructions as simple and intelligible as possible.

Statistical analysis of experimental data relies on the assumption that observations are independent (i.e. each observation comes from a different person). AMT makes an effort to prevent workers from having multiple accounts, and when a job is posted on AMT, the default setting is that any given worker can only complete that task once. Thus the multiple ID issue is not a substantial problem on AMT. However, researchers must be aware that should they re-post the same task twice, they are likely to get a sizeable fraction of repeat participants. Should you need to re-post a task, the job description can inform workers that they are not eligible if they have completed other tasks for you recently, and you can then reject the work (and discard the data) from repeat participants. Additionally, subjects who complete a job when it is first posted may differ systematically from later subjects. Therefore it is very important to plan out all experimental treatments ahead of time, such that subjects are uniformly randomized across conditions.

One general concern for running experiments, which theorists may be unfamiliar with, is the need for ethics approval. Most institutions have an Institutional Review Board (IRB), which must approve any study involving human subjects before it can be performed. Be sure to contact your IRB before conducting any experiments on AMT. An ideal arrangement would involve a general approval for experiments that fall within the range of demands and payments typical of other jobs posted on AMT.

A final issue relates to false positives and multiple testing (Ioannidis, 2005). The generally accepted threshold for statistical significance in experiments is a 5% chance of a false positive (i.e. $p < 0.05$). Thus if a given experiment in which no true effect is present was run many times, on average 1 out of 20 experimental runs would return a significant result just due to chance. Using AMT, it becomes very easy to run the same experiment many times. This presents a danger that researchers may (most likely subconsciously) continue to re-run an experiment until a falsely positive significant result is obtained, and then publish only the significant result. Although this danger is also present in the physical lab, it is much more acute using AMT precisely because of AMT's speed and low cost. This same speed and ease of AMT experiments, however, also allows for more rapid replications and follow-up studies (Pfeiffer et al., 2009), such that false positives are likely to come to light much more quickly.

6. Conclusion

Integrating behavioral experiments with evolutionary models is a promising approach for understanding the evolution of human behavior. Experiments generate important insight into the applicability of different models, and inspire new modeling directions. It is particularly important for researchers designing experiments to have a full grasp of the relevant theoretical models, in order to properly connect the two methods of inquiry. Yet thus far, relatively few evolutionary theorists have chosen to conduct behavioral experiments, due at least in part to the practical difficulties involved in laboratory research.

In this paper, I outline how online labor markets such as Amazon Mechanical Turk have removed many of these practical difficulties. I have also provided evidence that data collected on AMT is valid, as well as pointing out limitations and potentially problematic issues to be aware of when conducting AMT experiments. AMT makes it easy for theorists to run experiments in much the same way as they would run computer simulations. This opens up a new, exciting world of investigation, and promises to produce a wealth of fascinating new discoveries about the evolution of human behavior in the coming years.

Acknowledgments

I think Benjamin Allen, Anna Dreber and Martin Nowak for helpful comments on the main text, and Manfred Milinski and Bettina Rockenbach for helpful comments on the appendix. I gratefully acknowledge financial support from a grant from the John Templeton Foundation.

Appendix A. Reward, punishment and cooperation in games with reciprocity

While costly punishment has received the lion's share of attention, costly rewarding also plays an important role in human prosociality. Costly rewarding, where one subject pays a cost to give a benefit to another, is directly analogous to the Prisoner's Dilemma, and thus many such rewarding opportunities exist in daily life. Allowing subjects to reward other group members in a public goods game (second-party reward) is as effective as second-party punishment in promoting cooperation in all but the final period of fixed length public goods games (Sefton et al., 2007; Sutter et al., 2010), and in indefinitely repeated public goods games (Rand et al., 2009b), while neither reward nor punishment promotes cooperation in a single one-shot public goods game (Walker and Halloran, 2004). Rand et al. (2009b) also find that when both reward and punishment are available in an indefinitely repeated public goods game, a group's probability to reward high contributors is positively correlated with contributions and payoff, while no such correlations exist for the probability to punish low contributors; although surprisingly, the opportunity for rewarding does not decrease the average frequency of punishment. In a two-player proposer game where the responder can either reward, punish, do both, or do neither, the average proposal is highest when both reward and punishment are possible, and is higher for reward only than for punishment only (Andreoni et al., 2003). Taken together, these studies clearly demonstrate that subjects have a taste for second-party rewarding, and that rewards can be an important force for promoting cooperation.

In addition to evidence for the effectiveness of second-party rewarding, numerous experiments demonstrate a willingness to engage in third-party rewarding. Even in one-shot anonymous interactions where subjects observe the outcome of a dictator game and can then pay to reward or punishment the dictator, subjects are as likely to reward fair or generous behavior as they are to punish selfish behavior (Almenberg et al., 2011). In the presence of reputation, evolutionary game theoretic models show that cooperation can spread through indirect reciprocity, where my actions towards you depend on your previous actions towards others (Nowak and Sigmund, 2005; Ohtsuki and Iwasa, 2006). Consistent with these theoretical models, experiments show that under various reputation systems, subjects will frequently reward others, and in particular will preferentially reward those with a past history of being cooperative (Milinski et al., 2001; Seinen and Schram, 2006; Semmann et al., 2005; Wedekind and Milinski, 2000). A very elegant experiment shows that this tendency for third-party rewarding can be harnessed to avert the tragedy of the commons in a repeated public goods game (Milinski et al., 2002). After each round, each of the 6 public goods game group members is a donor for another randomly selected member as recipient, but direct reciprocity is excluded. The donor has full knowledge of the recipient's past behavior, in both the public goods game and the indirect reciprocity game, and then chooses whether or not to incur a cost to confer a benefit on the recipient. This setting is a mix of second-party rewarding for past actions toward the group (which includes the rewarder) and third-party rewarding for past actions towards others in the indirect reciprocity game. Here rewarding is common and leads to stable high levels of contribution in the public goods game. In another experiment (Semmann et al., 2005) the donor is either a member of the recipient's or of another public goods group. Rewarding occurs at the same level in either treatment and induces the same level of contribution to the public good.

There is also a synergistic interaction between this rewarding setup and the opportunity to self-select into an institution with costly punishment (Rockenbach and Milinski, 2006). Groups of eight subjects each played 20 periods of the public goods game. In treatment 'PUN&IR', before each period, each player can choose between joining a group in which the public goods game is followed by both costly punishing and an indirect reciprocity game, and a group in which the public goods game is followed solely by an indirect reciprocity game. In treatment 'PUN', before

each period, each player can choose between joining a group in which the public goods game is followed by costly punishing and a group in which the public good game is not combined with any other option. In both treatments the punishment groups achieve higher contributions than the no-punishment groups. Interestingly, the subjects in the PUN&IR treatment prefer the punishment opportunity group despite the presence of an alternative group offering only the reciprocity option, and the combination of indirect reciprocity and punishment results in the highest contributions and the highest efficiency. Additionally, the availability of indirect reciprocity changes how subjects chose to punish. In the indirect PUN&IR treatment, fewer punishments occur, but those that do are more focused on heavy free-riders. It is interesting to note two important differences between these results and those of the experiment with second party reward and punishment discussed above (Rand et al., 2009b): adding reward to punishment resulted in no increase in contributions or payoffs in the second party setup, and no decrease in punishment use; whereas the opposite is true in the Rockenbach and Milinski setup. The source of these differences merits further study, but may lie in the endogenous choice of punishment institution used here—in this setup, punishment works as a sorting tool in addition to its role in actually sanctioning low contributors.

In a related study, subjects have a choice each round between a standard public goods game and a setting with reward and punishment opportunities (Gurerk et al., 2006). Identities are shuffled from round to round, and reward has a 1:1 technology while punishment has a 3:1 technology. Almost all subjects eventually switch to the reward+punishment institution, and achieve much higher contributions than the game without targeted interaction. The frequency of reward use decreases over time, however. Both the lack of persistent identities and the 1:1 reward technology may contribute to the instability of rewarding in this experimental setup.

For a more detailed discussion of reward and punishment in public goods games, see Milinski and Rockenbach (this issue).

References

Abbink, K., Sadrieh, A., 2009. The pleasure of being nasty. Economics Letters 105, 306–308.

Abbink, K., Herrmann, B., 2011. The moral costs of nastiness. Economic Inquiry 49 (2), 631–633.

Almenberg, J., Dreber, A., Apicella, C.L., Rand, D.G., 2011. Third Party Reward and Punishment: Group Size, Efficiency and Public Goods, Psychology and Punishment. Nova Science Publishers, Hauppauge, NY.

Andreoni, J., Harbaugh, W.T., Vesterlund, L., 2003. The carrot or the stick: rewards, punishments and cooperation. American Economic Review 93, 893–902.

Apicella, C.L., Dreber, A., Campbell, B.C., Gray, P.B., Hoffman, M., Little, A.C., 2008. Testosterone and Financial Risk Preferences. Evolution and Human Behavior 29 (6), 384–390.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

Blanco, M., Engelmann, D., Koch, A., Normann, H.-T., 2010. Belief elicitation in experiments: is there a hedging problem? Experimental Economics 13, 412–438.

Brosnan, S.F., Salwiczek, L., Bshary, R., 2010. The interplay of cognition and cooperation. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2699–2710.

Buhrmester, M.D., Kwang, T., and Gosling, S.D., 2011. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 3 (6), 13-5.

Burnham, T., Hare, B., 2007. Engineering human cooperation. Human Nature 18, 88–108.

Burnham, T.C., 2003. Engineering altruism: a theoretical and experimental investigation of anonymity and gift giving. Journal of Economic Behavior and Organization 50, 133–144.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ.

Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? Experimental Economics 9, 265–279.

Croson, R.T.A., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. Journal of economic behavior & organization 41, 299–314.

Dal Bó, P., 2005. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. American Economic Review 95, 1591–1604.

Dal Bó, P., Fréchette, G.R., 2011. The evolution of cooperation in infinitely repeated games: experimental evidence. American Economic Review 101, 411–429.

Denant-Boemont, L., Masclet, D., Noussair, C., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic Theory 33, 145–167.

Dreber, A., Fudenberg, D., Rand, D.G., 2011. Who cooperates in repeated games? Available at SSRN: 〈http://ssrn.com/abstract=1752366〉.

Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A., 2008. Winners don't punish. Nature 452, 348–351.

Dreber, A., Apicella, C.L., Eisenberg, D.T.A., Garcia, J.R., Zamore, R., Lum, J.K., Campbell, B.C., 2009. The 7R polymorphism in the dopamine Receptor D4 gene (DRD4) is associated with financial risk-taking in men. Evolution and Human Behavior 30 (2), 85–92.

Fagerlin, A., Zikmund-Fisher, B.J., Ubel, P.A., Jankovic, A., Derry, H.A., Smith, D.M., 2007. Measuring numeracy without a math test: development of the subjective numeracy scale. Medical Decision Making 27, 672–680.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994.

Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397–404.

Fudenberg, D., Maskin, E., 1986. The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54, 533–554.

Fudenberg, D., Maskin, E.S., 1990. Evolution and cooperation in noisy repeated games. American Economic Review 80, 274–279.

Fudenberg, D., Rand, D.G., Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. American Economic Review, in press.

Gächter, S., Herrmann, B., 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 791–806.

Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public goods experiments. Experimental Economics 13, 364–377.

Gächter, S., Herrmann, B. The limits of self-governance when cooperators get punished: experimental evidence from urban and rural Russia. European Economic Review, in press.

Grujić, J., Fosco, C., Araujo, L., Cuesta, J.A., Sánchez, A., 2010. Social experiments in the mesoscale: humans playing a spatial Prisoner's Dilemma. PLoS ONE 5, e13749.

Gurerk, O., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312, 108–111.

Haley, K.J., Fessler, D.M.T., 2005. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evolution and Human Behavior 26, 245–256.

Harbaugh, W.T., Krause, K., 2000. Children's altruism in public good and dictator experiments. Economic Inquiry 38, 95–109.

Hauert, C., De Monte, S., Hofbauer, J., Sigmund, K., 2002. Volunteering as red queen mechanism for cooperation in public goods games. Science 296, 1129–1132.

Herrmann, B., Thoni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge.

Horton, J.J., Rand, D.G., Zeckhauser, R.J. The online laboratory: conducting experiments in a real labor market. Experimental Economics, in press, doi:10.1007/s10683-011-9273-9.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med 2, e124.

Janssen, M.A., Bushman, C., 2008. Evolution of cooperation and altruistic punishment when retaliation is possible. Journal of theoretical biology 254, 541–545.

Kandori, M., 1992. Social norms and community enforcement. The Review of Economic Studies 59, 63–80.

Kreps, D.M., Wilson, R., 1982. Reputation and imperfect information. Journal of economic theory 27, 253–279.

Mason, W., and Suri, S., 2010. Conducting behavioral research on Amazon's Mechanical Turk, Available at SSRN: 〈http://ssrn.com/abstract=1691163〉.

Melis, A.P., Semmann, D., 2010. How is human cooperation different? Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2663–2674.

Milinski, M., Semmann, D., Krambeck, H.J., 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415, 424–426.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.-J.r., 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proceedings of the Royal Society of London Series B: Biological Sciences 268, 2495–2501.

Milinski, M., Rockenbach, B. On the interaction of the stick and the carrot in social dilemmas. Journal of Theoretical Biology, this issue, doi:10.1016/j/jtbi.2011.03.014.

Nikiforakis, N., 2008. Punishment and counter-punishment in public goods games: can we still govern ourselves? Journal of Public Economics 92, 91–112.

Nowak, M., Sigmund, K., 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature 364, 56–58.

Nowak, M.A., 2006a. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M.A., 2006b. Evolutionary Dynamics: Exploring the Equations of Life. Belknap press of Harvard University Press, Cambridge, MA and London, England.
Nowak, M.A., Sigmund, K., 1992. Tit for tat in heterogeneous populations. Nature 355, 250–253.
Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.
Ohtsuki, H., Iwasa, Y., 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. Journal of Theoretical Biology 239, 435–444.
Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making 5, 411–419.
Pfeiffer, T., Rand, D.G., Dreber, A., 2009. Decision-making in research tasks with sequential testing. PLoS ONE 4, e4607.
Prelec, D.e., 2004. A Bayesian truth serum for subjective data. Science 306, 462–466.
Rand, D.G., Ohtsuki, H., Nowak, M.A., 2009a. Direct reciprocity with costly punishment: generous tit-for-tat prevails. Journal of Theoretical Biology 256, 45–57.
Rand, D.G., Armao, Iv, J.J., Nakamaru, M., Ohtsuki, H., 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. Journal of Theoretical Biology 265, 624–632.
Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., Nowak, M.A., 2009b. Positive interactions promote public cooperation. Science 325, 1272–1275.
Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444, 718–723.
Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. The American Economic Review 81, 1068–1095.
Sefton, M., Schupp, R., Walker, J.M., 2007. The Effect of rewards and sanctions in provision of public goods. Economic Inquiry 45, 671–690.
Seinen, I., Schram, A., 2006. Social status and group norms: indirect reciprocity in a repeated helping experiment. European Economic Review 50, 581–602.
Semmann, D., Krambeck, H.-J., Milinski, M., 2003. Volunteering leads to rock-paper-scissors dynamics in a public goods game. Nature 425, 390–393.
Semmann, D., Krambeck, H.-J., Milinski, M., 2005. Reputation is valuable within and outside one's own social group. Behavioral Ecology and Sociobiology 57, 611–616.
Sigmund, K., De Silva, H., Traulsen, A., Hauert, C., 2010. Social learning promotes institutions for governing the commons. Nature 466, 861–863.
Suri, S., Watts, D.J., 2011. Cooperation and contagion in web-based, networked public goods experiments. PLoS ONE 6 (3), e16836, doi:10.1371/journal.pone.0016836.
Sutter, M., Haigner, S., Kocher, M.G., 2010. Choosing the stick or the carrot? endogenous institutional choice in social dilemma situations. Review of Economic Studies 77, 1540–1566.
Traulsen, A., Hauert, C., De Silva, H., Nowak, M.A., Sigmund, K., 2009. Exploration dynamics in evolutionary games. Proceedings of the National Academy of Sciences 106, 709–712.
Traulsen, A., Semmann, D., Sommerfeld, R.D., Krambeck, H.-J., Milinski, M., 2010. Human strategy updating in evolutionary games. Proceedings of National Academy of Sciences of the United States of America 107, 2962–2966.
Ule, A., Schram, A., Riedl, A., Cason, T.N., 2009. Indirect punishment and generosity toward strangers. Science 326, 1701–1704.
Walker, J.M., Halloran, M., 2004. Rewards and sanctions and the provision of public goods in one-shot settings. Experimental Economics 7, 235–247.
Wedekind, C., Milinski, M., 1996. Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus generous tit-for-tat. Proceedings of the National Academy of Sciences of the United States of America 93, 2686–2689.
Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.
Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D., Cressman, R., Tao, Y., 2009. Costly punishment does not always increase cooperation. Proceedings of the National Academy of Sciences 106, 17448–17451.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51, 110–116.