Journal of Economic Behavior & Organization 90S (2013) S57–S75

Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

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

Information, irrationality, and the evolution of trust

Michael L. Manapat[a,b,c,*], Martin A. Nowak[a,b,c,d], David G. Rand[a,c,e]

[a] Program for Evolutionary Dynamics, Harvard University, One Brattle Square Suite 6, Cambridge, MA 02138, USA
[b] School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA
[c] Department of Mathematics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA
[d] Department of Organismic and Evolutionary Biology, 26 Oxford Street, Harvard University, Cambridge, MA 02138, USA
[e] Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 19 July 2012
Received in revised form 18 October 2012
Accepted 30 October 2012
Available online 8 November 2012

JEL classification:
C70
C72
C73

Keywords:
Trust
Evolutionary game theory
Cooperation
Reputation

ABSTRACT

Trust is a central component of social and economic interactions among humans. While rational self-interest dictates that "investors" should not be trusting and "trustees" should not be trustworthy in one-shot anonymous interactions, behavioral experiments with the "trust game" have found that people are both. Here we show how an evolutionary framework can explain this seemingly irrational, altruistic behavior. When individuals' strategies evolve in a context in which (1) investors sometimes have knowledge about trustees before transactions occur and (2) trustees compete with each other for access to investors, natural selection can favor both trust and trustworthiness, even in the subset of interactions in which individuals interact anonymously. We investigate the effects of investors having "fuzzy minds" and making irrationally large demands, finding that both improve outcomes for investors but are not evolutionarily stable. Furthermore, we often find oscillations in trust and trustworthiness instead of convergence to a socially optimal stable equilibrium, with increasing trustworthiness preceeding trust in these cycles. Finally, we show how "partner choice," or competition among trustees in small group settings, can lead to arbitrarily equitable distributions of the game's proceeds. To complement our theoretical analysis, we performed a novel behavioral experiment with a modified version of the trust game. Our evolutionary framework provides an ultimate mechanism—not just a proximate psychological explanation—for the emergence of trusting behavior and can explain why trust and trustworthiness are sometimes stable and other times unstable.

1. Introduction

Trust and trustworthiness are essential characteristics of successful human societies. We can study these phenomena game-theoretically using the trust game, an interaction between an "investor" and a "trustee" (Berg et al., 1995; Glaeser et al., 2000; Bohnet and Zeckhauser, 2004; Bohnet and Huck, 2004; Malhotra, 2004; Cox, 2004; King-Casas et al., 2005). The investor begins with an initial stake of one monetary unit and can either keep the stake or transfer it to the trustee. To represent the value created by interactions based on trust, the stake is multiplied by a factor $b > 1$ if the transfer is made. The trustee then chooses how much to return to the investor.

* Corresponding author at: Program for Evolutionary Dynamics, Harvard University, One Brattle Square Suite 6, Cambridge, MA 02138, USA. Tel.: +1 617 871 9656.
E-mail address: mlmanapat@gmail.com (M.L. Manapat).

0167-2681/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jebo.2012.10.018

In a one-shot anonymous trust game, there is no reason for a self-interested trustee to return anything. Hence, there is no reason for a self-interested investor to make the transfer. The potential gains of trust and exchange are lost. In essentially all behavioral experiments with the trust game, however, investors make transfers with high probability and trustees return a substantial amount (Berg et al., 1995; Glaeser et al., 2000; Bohnet and Zeckhauser, 2004; Malhotra, 2004; Kosfeld et al., 2005; Fehr, 2009; Johnson and Mislin, 2011). These results are inconsistent with the predictions of classical economic theory and have generated a great deal of interest across numerous fields. Economists have suggested proximate psychological motivations for such behavior, including preferences for fairness (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) and reciprocity (Rabin, 1993; Levine, 1998; Sethi and Somanathan, 2001; Cox, 2004; Dufwenberg and Kirchsteiger, 2004; Falk and Fishbacher, 2006). But such models leave open the question of how these preferences arose and how they are maintained (Wilson and Gowdy, 2012). Our evolutionary analysis provides an ultimate and not just proximate explanation for trust. Rather than restating the observed behavior in terms of a utility function with other-regarding preferences, our evolutionary approach proposes a specific mechanism that could have lead to the evolution of the observed preferences (Wilson and Gowdy, 2012).

A key aspect of many evolutionary approaches is to realize that behavior observed in a laboratory experiment reflects strategies that people developed in broader contexts outside the laboratory (for direct empirical evidence, see Rand et al., 2012). Here we apply this perspective to the problem of trust. Our model accounts for the fact that preferences evolved in the context of daily life in which investors sometimes have foreknowledge of trustee behavior because of the existence of reputation systems. We show how this potential access to information fundamentally changes the nature of the game and leads to the evolution of trust and marginal levels of trustworthiness. We then show that competition among trustees in small group settings, together with this information about trustees, can lead to trustworthiness that is more than just marginal. These results help us understand why humans exhibit the social preferences that promote successful economic exchange—these preferences are adaptive in the more ecologically valid setting we consider and can lead to trust and trustworthiness even in one-shot anonymous interactions. Our evolutionary model also gives insight into how one might design modern-day institutions, such as online markets, to promote trust and trustworthiness.

We consider an evolving population of investors and trustees. Each investor has a strategy p_0, which is the probability the investor makes the transfer when the interaction is anonymous and the investor has no information about the trustee. Each trustee has a strategy r, which is the fraction of the transfer that the trustee returns to the investor (in all settings, as the trustee does not know whether the investor has information in any particular interaction). We call p_0 the investor's "trust" and r the trustee's "return."

For any interaction, there are two possible scenarios in our model. With probability $1-q$, the investor finds herself in an anonymous interaction where she does not have any information about the trustee and makes the transfer with probability p_0. In this case, the investor's expected payoff is $1-p_0+p_0br$ and the trustee's expected payoff is $p_0b(1-r)$. With probability q, on the other hand, the investor learns information about the trustee before the interaction occurs and thus knows the fraction r that the trustee will return before deciding whether to make the transfer (Frank, 1987; Sethi and Somanathan, 2001; Dekel et al., 2007). The parameter q is thus a measure of the availability of information about trustees as it spreads in the population of investors (Nowak and Sigmund, 1998, 2005; Wedekind and Milinski, 2000; Milinski et al., 2001, 2002; Panchanathan and Boyd, 2004; Brandt and Sigmund, 2005; Ohtsuki and Iwasa, 2006; Ohtsuki et al., 2009; Pfeiffer et al., 2012). When an investor has information about the trustee she is facing, she can condition her behavior on that information. Trustees are not aware if investors have information in any given interaction and therefore have a fixed strategy across all interactions. Our basic model of information spread is simple. It assumes that an investor has a uniform probability q of knowing the trustee's r. In Section 3.2, we show that a more realistic mechanism for the spread of information yields similar results.

Adding information to the trust game fundamentally alters the nature of the interaction. In the fraction q of instances in which the investor has information about the trustee, players are no longer in the domain of trust. Instead, they are bargaining as in the ultimatum game (Güth et al., 1982): the trustee is effectively offering a split of the pot b, and the investor can accept the split by making the transfer or reject it by withholding. The trustee thus becomes the "ultimatum game proposer" and the investor the "ultimatum game responder." Adding information effectively reverses the order of play (Nowak et al., 2000; McNamara and Houston, 2002). Here we study how this linking of trust and bargaining leads to the evolution of pro-social behavior in both domains.

We will see that information leads to the evolution of fully trusting and marginally trustworthy behavior. But for true trustworthiness to evolve, something more is needed. To this end, we introduce "partner choice," a mechanism in which an investor uses limited information about trustees to make a semi-informed decision about the trustee with whom she should interact. We find that partner choice can lead to the robust evolution of trust and trustworthiness. Thus, because humans developed in contexts in which information about partners was available and, importantly, partner choice was possible, our evolved intuitions about whether to trust and to be trustworthy—even in situations in which information is not present—induce us to act in a pro-social way.

2. Evolutionary process

To explore the evolution of trust, we allow the strategies of investors and trustees to change through an evolutionary process. This process can be interpreted as genetic evolution or as social learning (Nowak and Sigmund, 2004). In either case,

higher payoff strategies become more common while lower payoff strategies die out. In every round, each investor interacts once with each trustee. Players accumulate payoffs across interactions. Investors can learn from other investors, and trustees can learn from other trustees. Mutation is also possible. The fidelity of learning (intensity of selection) and the mutation rate are parameters of the process. We now discuss the details of this model when investors are rational and payoff-maximizing in the fraction q of cases in which they have information about the trustee. Later we will relax this assumption, but changes to the investors' decision rule do not affect the nature of the evolutionary process.

We begin with a well-mixed population split evenly between investors and trustees. Each investor has her own p_0, the probability that she makes the transfer when she has no information about the trustee. When she does know the trustee's return fraction r, she always makes the transfer if $r > 1/b$ and never makes the transfer if $r \leq 1/b$. Each trustee has his own r, the fraction of what he receives that he returns to the investor.

Suppose that an investor with strategy p_0 and a trustee with strategy r play the game a large number of times (without changing their strategies). Let $p^*(r)$ be the transfer probability when the investor knows the trustee's r, so

$$p*(r) = \begin{cases} 0, & \text{if } r \leq 1/b, \\ 1, & \text{if } r > 1/b. \end{cases} \tag{1}$$

With probability $1 - q$, the investor does not know the trustee's r. In those cases, she transfers 0 with probability $1 - p_0$ and 1 with probability p_0. Thus, the average payoff to the investor is

$$(1 - p_0) + p_0 br \tag{2}$$

and the average payoff to the trustee is

$$p_0 b(1 - r). \tag{3}$$

With probability q, the investor knows the trustee's r. In those cases, she transfers 0 with probability $1 - p^*(r)$ and 1 with probability $p^*(r)$. Thus, the average payoff to the investor is

$$(1 - p*(r)) + p*(r)br \tag{4}$$

and the average payoff to the trustee is

$$p*(r)b(1 - r). \tag{5}$$

We compute the payoff for the investor and trustee by assigning weights of $1 - q$ and q (respectively) to the payoffs described above.

Initially, each p_0 and r is selected uniformly from the interval [0, 1]. In each round of the evolutionary simulation, every investor "plays" every trustee, and the payoffs are computed as above. (Put another way, each "interaction" can be thought of as a very large number of games with the interaction payoff computed as the average payoff over all of those games.) After this round-robin tournament, both populations evolve as follows. Two investors, call them A and B, are selected uniformly at random. Let $\overline{\pi}_A$ be the average per-game payoff of A, $\overline{\pi}_B$ the average per-game payoff of B, and

$$\rho = \frac{1}{1 + e^{-\beta(\overline{\pi}_A - \overline{\pi}_B)}}. \tag{6}$$

B is replaced by a copy of A with probability $\rho(1 - \mu)$ and by a random mutant (with a strategy chosen uniformly at random from [0, 1]) with probability μ. With probability $(1 - \rho)(1 - \mu)$ no one changes his or her strategy. The same strategy update then occurs for the trustees. This evolutionary process is known as the "pairwise comparison process" (Traulsen et al., 2007). The parameter μ is the mutation rate. The parameter β is the intensity of selection. The larger β is, the more likely it is that a player will imitate the strategy of someone doing better (and not imitate the strategy of someone doing worse).

This process is repeated for a large number of rounds, and the stationary values of the strategy parameters are found by averaging over the last 80 percent of rounds. Evolutionary processes such as this one result in feedback between the individual and the population: as the distribution of strategies in the population changes, so too do the strategies chosen by individuals in that population. Thus, causality in economic change occurs both upwards from the individual to the super-indiviudal level and downwards from the super-individual level to the individual. This bidirectional causality is not usually a feature of classical economic theory (Gowdy et al., 2012). Furthermore, instead of just being a "marginal analysis of self-regarding individuals in a near-to-equilibrium system" (Gowdy et al., 2012), the evolutionary approach gives us insight into the full history of the development of the traits being studied, and that history can have significant economic and public policy implications (Wilson and Gowdy, 2012).

Later, we will model investor "irrationality" by having the transfer probabilities $p^*(r)$ in the presence of information be either translated step functions or sigmoids. The simulations in those cases proceed exactly as described above except that p^* is replaced with one of these other functions.

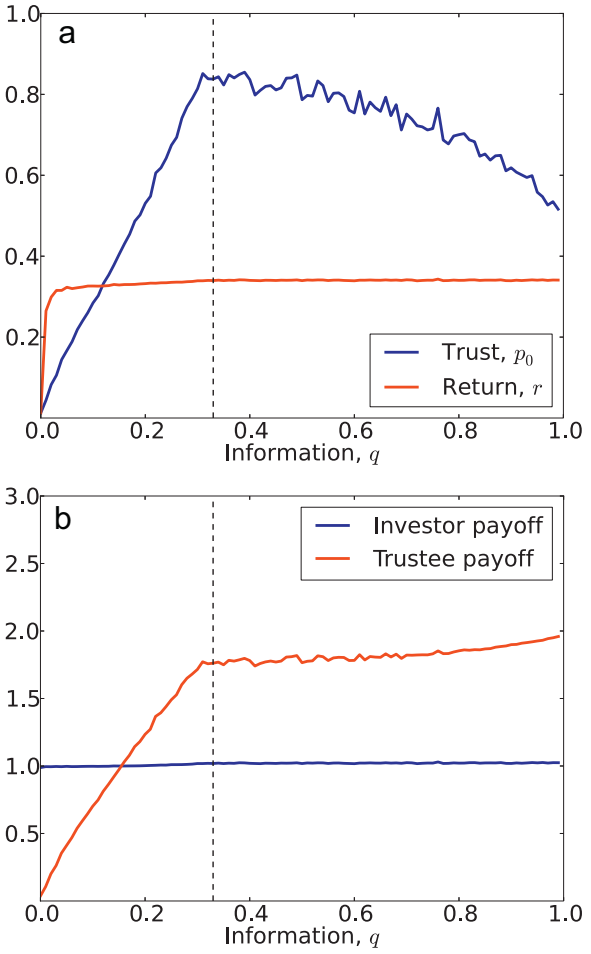

Fig. 1. Giving investors information about trustees benefits trustees. (a) The average investor trust p_0 and trustee return r as functions of the information q when $b = 3$. When q is 0, p_0 and r are both 0. Once $q \geq 1/b$ (dashed line), p_0 is large (and in fact would be constantly 1 in the limit of infinite selection intensity) and r is just slightly more than $1/b$. (b) Average payoffs of investors and trustees as functions of q. When information is present, trustees capture almost all the profits. Giving information about trustees to investors benefits trustees more than investors. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, and the selection intensity $\beta = 20$. Results are averaged over 50 simulation runs, each run consisting of 50, 000 rounds.

3. Rational investors

3.1. Payoff-maximizing investor decision rule

Suppose first that, as described in the previous section, investors are perfectly rational and always act to capture any profit. When such an investor is aware of the trustee's r, she makes the transfer if $r > 1/b$ and obtains a payoff of $rb > 1$. Conversely, she does not make the transfer if $r < 1/b$. When investors have no information at all, $q = 0$, then evolution leads to the classical trust game Nash equilibrium: investors never transfer, $p_0 = 0$, and trustees return nothing, $r = 0$. The situation changes markedly, however, when investors sometimes have information about trustees, $q > 0$. To explore the effect of information, we begin by using agent based simulations, studying stochastic evolutionary game dynamics in finite populations (Nowak et al., 2004; Fudenberg and Imhof, 2006; Manapat et al., 2012). Fig. 1(a) shows the investor's trust p_0 and trustee's return r for different values of q.

When the probability that an investor knows the trustee's return is sufficiently large, $q \geq 1/b$, the benefits of trust are realized. Trustees return a fraction just slightly greater than $1/b$ on average. Thus, investors almost always make the transfer when they have information. Even when they do not have information, they usually trust and make the transfer (i.e., p_0 is large). Interestingly, the average trust p_0 does not increase to 1 even though $p_0 = 1$ and $r = 1/b + \epsilon$[1] is a Nash equilibrium for $q \geq 1/b$ (see Section 4.5). If q is close to 1, investors usually know trustees' return values and rarely have to make blind

[1] Henceforth, $v + \epsilon$ will represent the smallest possible amount, in monetary units, strictly greater than v.

decisions. Hence, their values of p_0 are largely irrelevant. As q increases, the selection pressure on p_0 becomes weaker. Ultimately, for $q = 1$, investors always have information about trustees and p_0 has no bearing on their payoffs whatsoever, resulting in neutral drift around the average value $p_0 = 1/2$. If there is a non-zero but small probability of information, $0 < q < 1/b$, the population oscillates between $p_0 = 0$, $r = 0$ and $p_0 = 1$, $r = 1/b + \epsilon$. A formal analysis of both the equilibrium and non-equilibrium cases in a more general scenario is presented in Sections 4.4 and 4.5.

These results offer a straightforward explanation for the trust and trustworthiness observed in experiments. Information about others has generally been available over the course of human evolution as well as in most interactions in modern life (Nowak and Sigmund, 2005). Our model shows that such information can make it adaptive for trustees to return large fractions of what they receive. Trustees who return little are caught sufficiently often that those who are trustworthy outearn the stingy (Kandori, 1992). As a result, trustees have an incentive to be trustworthy and investors have an incentive to be trusting, i.e., to make transfers even when they do not know the trustee's r. As the value created by trust (the multiplier b) increases, less monitoring, just $q \geq 1/b$, is required to enforce trustworthiness. In a world of highly beneficial interactions, $b \gg 1$, even a small chance of having information about trustees is enough to generate trusting and trustworthy behavior.

Fig. 1(b) shows the payoffs corresponding to the scenarios in Fig. 1(a). As q increases, the average investor payoff increases only slightly. The average trustee payoff, on the other hand, increases dramatically. Trustees capture almost all the benefit created by the exchange. Trustees need only return a small fraction of their profits to ensure a transfer from a rational self-interested investor. Thus, trustee accountability actually benefits trustees more than it benefits investors.

3.2. Stochastic information diffusion

We have heretofore abstracted away the details of how information about trustees spreads among investors. In particular, we posited that investors have a probability q of knowing the trustee's r before the transaction begins, where q is fixed, and that information spreads much faster than the players change their strategies. The latter assumption means that when a trustee updates his or her strategy at the end of a round, investors still have the same probability q of knowing that trustee's new r at the beginning of the next round.

While these assumptions simplify the technical analysis, they are not essential for our results. In this section, we describe a particular concrete mechanism for the spread of information and show that the results are qualitatively the same as those for our simplified model.

As in Section 2, we begin with a population evenly split between investors and trustees. Each round consists of a fixed number of games, and each player has a strategy that is fixed for the entirety of the round. At the beginning of the round, investors do not know anything about trustees. An investor and a trustee are chosen uniformly at random to play the game. Since the investor does not know the trustee's r, she makes the transfer with probability p_0. If the transfer is made, the investor informs a fraction q' of investors about the r of the trustee with whom she just interacted. If one of these investors plays the trustee later in the generation, then she will be able to condition how much she transfers on the trustee's r, just as in Section 2. Thus, after each game in which the investor does make the transfer (and regardless of whether or not she had information about the trustee prior to the interaction), a fraction q' of investors are informed of the r of the trustee involved in the game. We call q' the "stochastic information." When an investor and a trustee meet, the investor's strategy depends on whether she has received information about the trustee in the past. If she has, she makes the transfer according to her conditional strategy. If not, she makes the transfer with probability p_0. As more and more games are played, the probability that an investor knows the r of any given trustee increases. The rate at which this probability increases is controlled by q'. We assume that information does not persist between generations, so this process of information-gathering begins anew each round.

Fig. 2 is the analogue of Fig. 1 but with stochastic information. We see that our concrete mechanism produces results very similar to those of the simplified system analyzed in the paper. For a more detailed study of how stochasticity affects the spread of information—in particular, how finite investor memories and conflicting information about trustees can affect outcomes—see Manapat and Rand (2012).

4. Irrational investors

4.1. Behavioral experiments

Thus far we have assumed that investors use information in a payoff-maximizing manner. When they have knowledge of the fraction r the trustee will return, they act to maximize their payoff and transfer with probability 1 if $r > 1/b$ and with probability 0 if $r < 1/b$. But how do investors, when probabilistically given information about trustees, behave in practice?

To gain insight into this question, we performed an experiment in which 175 subjects played a modified trust game with information. We allowed investors to condition their decision on the amount returned by the trustee. Specifically, investors indicated the minimum amount a trustee must return for the investor to make the transfer. If the trustee chose to return less than this amount, no transaction occurred and the investor kept his or her initial stake. Trustees knew that investors could decide whether to make the transfer based on how much the trustee returned but did not know the specific amount demanded (see Section 4.1.1 below for details).

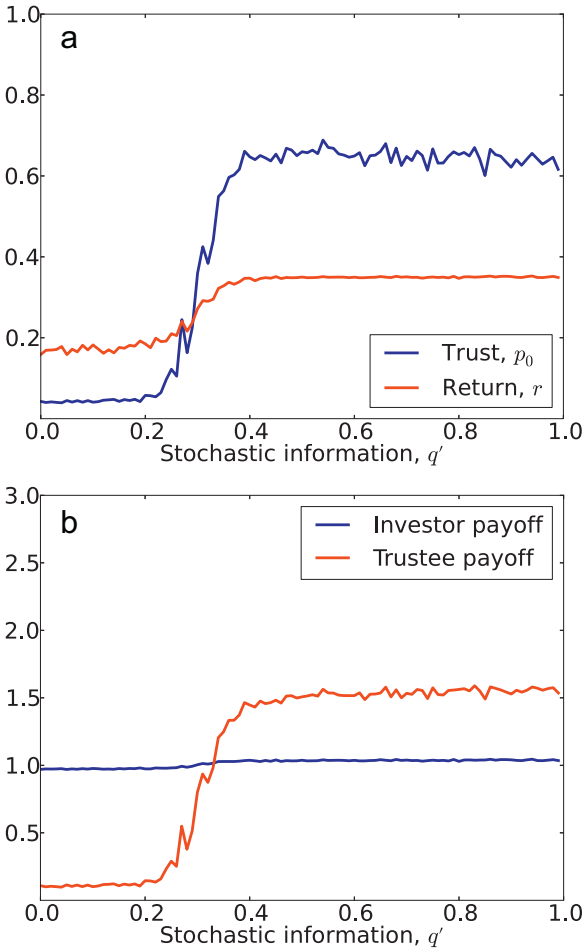

Fig. 2. Stochastic information spread. The analogues of Fig. 1(a) and (b) when information about trustees spreads "stochastically." After each game in which the investor makes the transfer, a fraction q' of the investor population is informed of the r of the trustee involved in the interaction. When an investor and a trustee meet, either the investor has at some point in the past been informed of the trustee's r, in which case she makes the transfer according to her conditional strategy, or she has not been, in which case she makes the transfer with probability p_0. The results are qualitatively similar to those in Fig. 1. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, the selection intensity $\beta = 20$, and the multiplier $b = 3$. Results are averaged over 50 simulation runs, each run consisting of 50, 000 rounds and each round consisting of 1000 random games.

The distribution of the minimum return fractions acceptable to investors is given in Fig. 3(a) and the corresponding cumulative distribution function (CDF) in Fig. 3(b). The CDF gives the fraction of investors that make the transfer for a given return fraction r. Thus, it represents the investor decision rule in the presence of information. For example, if all investors were perfectly rational and self-interested, this CDF would be a step function with a transition from 0 to 1 at $r = 1/b$.

The investor decision rule in the experiment deviates from this step function in two ways. First, the transition from low to high transfer probability occurs gradually as r increases. This means that investors' sensitivity to changes in the trustees' r's is dulled. Because of their "fuzzy minds," investors sometimes make the transfer even when the trustee's r is below what the investor wants and sometimes do not make the transfer even when the trustee's r is above the desired level. Second, the point at which the transfer probability is increasing fastest occurs at a value of r larger than $1/b$. This means that investors demand more than just a tiny profit from trustees. The CDF in Fig. 3(b) is fit well by a sigmoid of the form $1/(1 + e^{-(r-t)/f})$, where f represents the "fuzziness" of investors' minds and t represents the return fraction that investors "demand" from trustees. From our data, we estimated values of $f = 0.11$ and $t = 0.39$ for $b = 3$.

4.1.1. Experimental design

We measured how people behave as investors and trustees with a behavioral experiment in December 2010. Subjects were recruited using the online labor market Amazon Mechanical Turk ("Mturk," https://www.mturk.com/mturk/welcome) (Horton et al., 2011; Rand, 2012). Subjects received a $0.40 show-up fee for participating and were able to earn bonuses of up to $1.20 depending on how they played, consistent with the usual compensation levels in this market. The relatively low stakes on Mturk have been shown in general not to affect the play of individuals when compared to how they act in higher-stakes laboratory experiments. Indeed, a number of standard results in behavioral economics have been replicated on Mturk

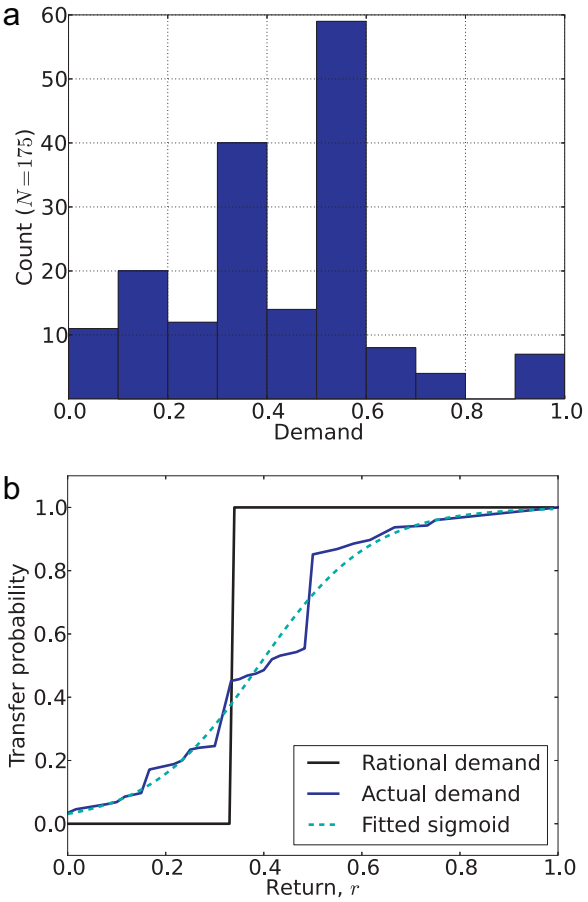

Fig. 3. Behavior in a trust game experiment with information. (a) We conducted a behavioral experiment (with $b = 3$) to determine how much investors demand from trustees when investors have information about the trustees' return fractions. The distribution of demands ($N = 175$) is shown in (a) and the corresponding cumulative distribution function (CDF) in (b). The CDF represents the average probability that an investor makes the transfer for a given amount returned by the trustee. The empirical data is fit well by a sigmoid (dotted line) that deviates from the payoff-maximizing step function in two ways: it is smooth rather than sharp (investors have "fuzzy minds"), and it is shifted to the right (investors demand more than just an ϵ-profit).

with these lower stakes (Horton et al., 2011; Paolacci et al., 2010). For instance, there has been a recent demonstration of quantitative agreement between higher-stakes games in traditional laboratories and low-stakes games run on Mturk using the trust game, public goods game, ultimatum game, dictator game, and prisoner's dilemma (Suri and Watts, 2011; Horton et al., 2011; Amir et al., 2012). This relative insensitivity to stake size in social dilemmas is consistent with previous research in the laboratory (see Camerer and Hogarth, 1999 for a review).

We sought to determine how investors would use information about trustees' return fractions when making the decision to transfer or not. We recruited 175 individuals, each of whom first acted as the "investor" and was asked to indicate an amount such that

- If Player 2 choses to return less than the amount you indicate, you keep the 20 cents, and
- If Player 2 choses to return the amount you indicate or more, then you will make the transfer and earn whatever amount Player 2 returns.

Participants could "demand" any integral amount between 0 and 60 cents (inclusive). The most common demand was 30 cents, or a trustee return fraction of 1/2. The second most common demand was 20 cents, or a trustee return fraction of 1/3, which guarantees that the investor does not incur a loss from making the transfer. The full distribution of investor responses to this question (after normalization) and the corresponding cumulative distribution function are shown in Fig. 3(a) and (b).

Next, the participants were asked to act as trustees and told, "Player 1's decision of whether to transfer is based on the amount you choose to return. If you return more than Player 1's specified minimum amount, the transfer happens; if not, you receive nothing." The 175 individuals returned on average 29.97 cents, just under 50 percent of what they received. This is slightly higher than in the situation in which investors do not get to condition their transfer on the trustee return fraction (Johnson and Mislin, 2011), which is reasonable as trustees might raise their return fractions when they know investors can

have information about them. Note that subjects were not aware that they would subsequently act as Player 2 while making their decisions as Player 1.

The 175 investor demands were randomly matched with the 175 trustee responses to determine the actual payoffs for the participants, each of whom could earn up to 60 cents from his or her play as the investor and up to 60 cents from his or her play as the trustee. The instructions provided in the experiment are available from the authors.

When a perfectly rational investor knows a trustee's r, the probability that the investor makes the transfer is given by a step function of r, the probability being 1 if $r > 1/b$ and 0 otherwise. But we have seen that investors are not perfectly rational and that their irrationality can manifest itself in two ways. First, they can have "fuzzy minds." This means that investors might make errors when interpreting information about trustees. As a result, they sometimes make the transfer even if $r < 1/b$, and they sometimes do not make the transfer even if $r > 1/b$. Second, investors can have unreasonable demands—they make the transfer if and only if $r > t$, where $t > 1/b$. As we saw in the previous section, we can capture both types of "irrationality" by letting the investor's transfer probability be given by a sigmoidal function, $1/(1 + e^{-(r-t)/f})$. The parameter f is the "fuzziness" of the investor's mind. As $f \to 0$, the sigmoid approaches a step function. The parameter t is the demand. For the perfectly rational investor, $f = 0$ and $t = 1/b$. We now explore the consequences of each of these two forms of irrationality.

4.2. Fuzzy minds

First, suppose investors demand the rational amount, $t = 1/b$, but have fuzzy minds, $f > 0$. We begin by assigning all investors the same fixed f. Fig. 4(a) shows the average investor payoff as a function of f. Initially, the payoff for investors increases as their minds become collectively clouded. Confusion benefits the investors: trustees must return larger amounts to ensure that fuzzy-minded investors make transfers. But it is not good for investors if they become too confused, for if they are they do not use information sufficiently wisely and thus cannot exert pressure on trustees. There is an optimal value of f that is payoff-maximizing for investors (see Section 4.6).

Fig. 4, which shows the average investor payoff as a function of the collective f (and t), has the interesting property that there are reversals in the payoff orderings. When $f = 0$, for example, investor payoffs are highest when $q = 1$. When f is large, on the other hand, investor payoffs are highest when $q = 0.5$. What causes this reversal? When $f = 0$, investors are perfectly rational in their use of information. They act to capture any profit, however small. The more information they have, the better it is for them as they are acting as intelligently as possible. When $f \to \infty$, on the other hand, investors make the transfer at random when they have information. If $q = 1$, this means that investors always make the transfer at random (as they always have information). Hence, there is no incentive for trustees to return anything and the investor payoff goes to zero. When $q < 1$, on the other hand, investors sometimes have no information about trustees and make the transfer with probability p_0. But $p_0 \approx 0$ on average: investor behavior when investors have information does not promote trustworthiness because the behavior is random. Thus, with probability $1 - q$ investors withhold the endowment from the exploitative trustees, increasing the average investor payoff. The larger $1 - q$ is (i.e., the smaller q is), the greater the retained payoff.

Collective investor irrationality benefits investors more than it benefits trustees. If all investors make the transfer according to a sigmoid with $f > 0$, then the average investor payoff can be higher than it is when all investors are perfectly rational. For the parameter values we have chosen, our model predicts an optimal f between 0.06 and 0.1, close to the value of $f = 0.11$ observed experimentally.

4.3. Irrational demands

Now suppose that investors are perfectly sharp in their decisions, $f = 0$, but demand a return $t > 1/b$. Again we initially assign all investors the same fixed demand, t. Fig. 4(b) shows the average investor payoff as a function of t. When investors always have information, $q = 1$, trustees have no choice but to satisfy investors' demands. Thus, investors' payoffs increase linearly with t. When information is not always available, $q < 1$, there is an optimal demand. If $t < q$, trustees do best by complying with investor demands, so investors should uniformly increase t. If $t > q$, there are oscillations between two states: one with neither investor trust nor trustee returns ($p_0 = 0$, $r = 0$) and the other with full trust and the demanded return ($p_0 = 1$, $r = t + \epsilon$). The larger the difference $t - q$, the more time the population spends in the former state. Investors thus maximize their payoffs by demanding $t = q$.

Fig. 5 shows the oscillations in trust, p_0, and return, r, when the collective investor demand t exceeds the information level q (in this case, $q = 1/3$ whereas $t = 2/3$). It also provides insight into the question of which arises first, trust or trustworthiness. When both the average p_0 and the average r are close to zero, r increases first and then p_0 follows: trustworthiness leads to trusting behavior. When trustees are not trustworthy, an investor who experiments by trying a larger p_0 will be exploited. Selection thus keeps the average p_0 close to 0. However, a mutant trustee with $r > 1/b$ has an advantage: when the investor knows the trustee's r, the mutant trustee will get the transfer. Therefore mutation and selection together lead to increasing trustee return fractions. Selection then pushes p_0 upwards in response.

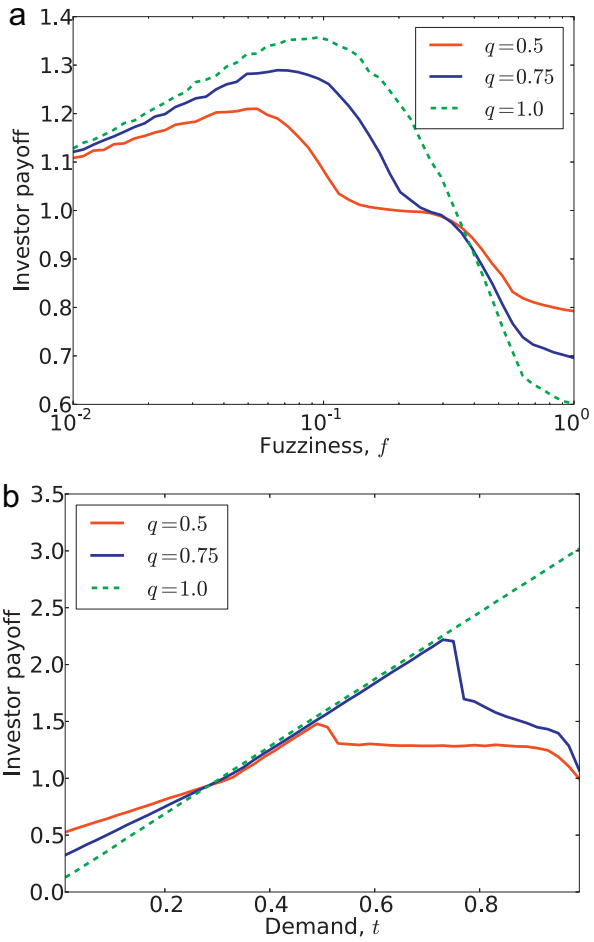

Fig. 4. Fuzzy minds and unreasonable demands benefit investors. (a) The average investor payoff when all investors have the same (exogenously fixed) "fuzziness" of mind. There is an optimal value for f. If investors are too "sharp," then they are satisfied by small profits, giving trustees little incentive to return substantially more than just $1/b$. If they are too "fuzzy," they fail to use information about trustees adequately. (b) The average investor payoff when all investors have the same (exogenously fixed) demand t for the trustees' return fraction. When $q = 1$, the investor payoff increases linearly with t: trustees have no choice but to comply with investor demands. But when $q < 1$, there is an optimal, payoff-maximizing value of $t\,(= q)$ for investors. We see that "fuzzy minds" (a) and "unreasonable demands" (b) benefit investors more than trustees. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, the selection intensity $\beta = 20$, and the multiplier $b = 3$. Results are averaged over 50 simulation runs, each run consisting of 50, 000 rounds.

4.4. Oscillations when $q \le t$, $t \ge 1/b$

In this section, we study in detail the oscillations in strategies that occur when $q \le t$ and $t \ge 1/b$. In Fig. 1(a), the values of p_0 and r when $0 < q \le 1/b$ are in fact the results of averaging over cycles in which p_0 and r are oscillating between the $p_0 = 0$, $r = 0$ state and the $p_0 = 1$, $r = 1/b + \epsilon$ state. There are no equilibrium strategies here because $t = 1/b$ but $q \le t$. Similarly, in Fig. 4(b), r increases linearly with t for t between 0 and q since $r = t + \epsilon$ is the equilibrium strategy for the trustee when $t < q$ (see Section 4.5 below). However, once $t \ge q$, the system degenerates into cycles and the decreasing q is the result of averaging over those cycles.

Let us obtain a more precise understanding of this phenomenon. Consider a game between a single investor and a single trustee in which the investor's demand $t \ge 1/b$ is larger than the information q but the investor's mind is not at all "fuzzy" ($f = 0$). If the trustee's r is less than $1/b$, $p_0 = 0$ is optimal for the investor. If the trustee's r is greater than $1/b$, $p_0 = 1$ is optimal for the investor. Generically, only $p_0 = 0$ and $p_0 = 1$ are rational investor strategies.

Suppose first that $p_0 = 1$. Among all the trustee return fractions that satisfy the investor's demand, clearly the optimal one for the trustee is $r = t + \epsilon$. But when $t \ge q$, the trustee is indifferent between $r = t + \epsilon$ and

$$r = \frac{t - q}{1 - q} + \epsilon', \tag{7}$$

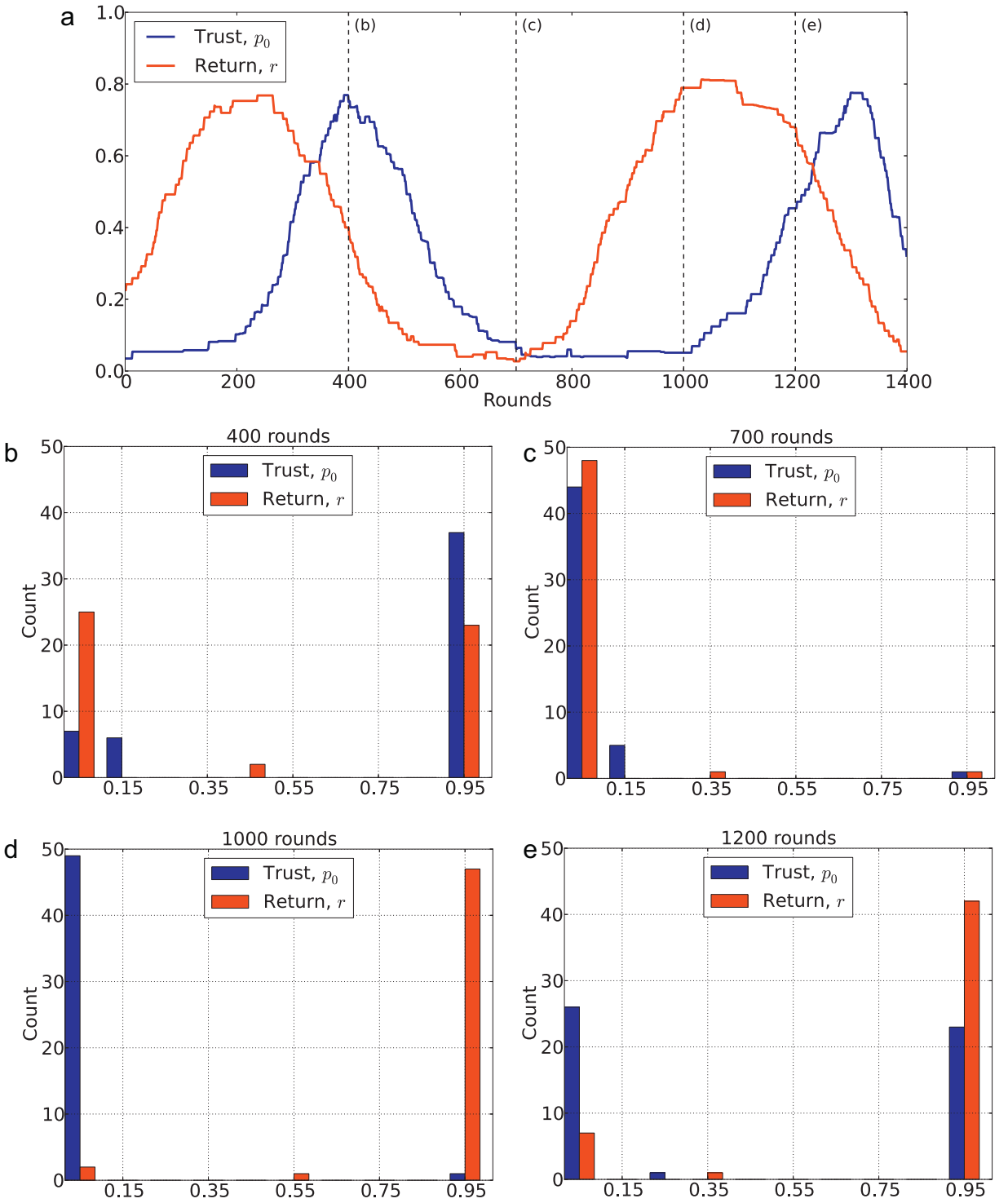

Fig. 5. Oscillations in trust and trustworthiness. (a) An example of how the average investor trust p_0 and the average trustee return r evolve over time when $t = 2/3 > 1/3 = q$. When the collective investor demand t is larger than the information q, there are no equilibria and both investors and trustees oscillate between strategies. (b) The average p_0 is peaking while the average r is decreasing. Investors eventually adapt to lower trustee r's by reducing their p_0's. **(c)** Both the average p_0 and the average r are close to zero. Trustees who return more do better when investors have information, so the average r begins to increase. Trustworthiness arises first, then trust follows. **(d)** The average p_0 is low while the average r is high. Since trustees are returning a large fraction of what they receive, investors evolve to make the transfer even when they have no information. **(e)** Both the average p_0 and the average r are high. Trustees with lower r's profit more by "cheating" the mostly trusting investors, so the average r decreases. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, the selection intensity $\beta = 20$, and the multiplier $b = 3$.

where $\epsilon' = \epsilon/(1-q)$. This follows immediately from the trustee's payoff function:

$$\pi_T = \begin{cases} b(1-r)(p_0 - p_0 q), & \text{if } r \le t, \\ b(1-r)(p_0 - p_0 q + q), & \text{if } r > t. \end{cases} \tag{8}$$

Thus, when the investor demands more than the information q, the trustee has a strategy that does not satisfy the investor's demand but has a payoff as high as the optimal demand-satisfying strategy. In fact, the trustee's payoff is a decreasing function of r between $r = 0$ and

$$r = \frac{t-q}{1-q} + \epsilon', \tag{9}$$

so the trustee maximizes his payoff by switching to $r = 0$.

Once $r = 0$, the investor should switch to $p_0 = 0$. But once $p_0 = 0$, the trustee increases his payoff by switching to $r = t + \epsilon$. The investor follows by switching to $p_0 = 1$ (since $r = t + \epsilon > 1/b$), and the cycle repeats. Fig. 5 shows the oscillations in the average p_0 and r over time. The waves are phase-shifted as expected.

The oscillations in trust and trustworthiness are inconsistent with the notion that a socially optimal general equilibrium always exists (Gowdy et al., 2012). Knowing that a particular socially optimal situation is not stable, however, is important for public policy reasons. Just as an understanding of the business cycle and the role of proper fiscal and monetary policy allows governments and central banks to take countercyclical action—employing loose fiscal and monetary policy in times of economic distress—so too can an understanding of the oscillations in social behavior allow institutions to apply policies that reduce the frequency (and amplitude) of the oscillations in an attempt to keep society closer to the social optimum.

4.5. Nash equilibrium analysis

Here we determine explicitly which (pure) strategies are equilibria when investors do not have fuzzy minds ($f = 0$) but may have irrational demands ($t \neq 1/b$). As in the previous section, we consider a game between a single investor and a single trustee. When the investor knows the trustee's r, she makes the transfer if and only if $r > t$. The demand t is exogenously fixed, and strategies for the investor and trustee consist of a choice of p_0 and r (respectively) given this fixed t. We assume that all individuals are risk-neutral and seek to maximize their expected payoffs.

Theorem 1. *The following are all the pure Nash equilibria of the trust game with information:*

- $p_0 = 0$ *and* $r \le 1/b$ *when* $q = 0$,
- $p_0 \in [0, 1]$ *and* $r = t + \epsilon$ *when* $q = 1$,
- $p_0 = 1$ *and* $r = t + \epsilon$ *when* $1/b \le t < q < 1$.
- $p_0 = 0$ *and* $r = t + \epsilon$ *when* $t < 1/b$ *and* $0 < q < 1$.

Proof. The investor's payoff is given by

$$\pi_I = \begin{cases} 1 - (1 - br)(p_0 - p_0 q), & \text{if } r \le t, \\ 1 - (1 - br)(p_0 - p_0 q + q), & \text{if } r > t, \end{cases} \tag{10}$$

and the trustee's payoff by

$$\pi_T = \begin{cases} b(1-r)(p_0 - p_0 q), & \text{if } r \le t, \\ b(1-r)(p_0 - p_0 q + q), & \text{if } r > t. \end{cases} \tag{11}$$

If $r > 1/b$, the investor maximizes her payoff by choosing $p_0 = 1$. If $r < 1/b$, the investor maximizes her payoff by choosing $p_0 = 0$. Generically, only $p_0 = 0$ and $p_0 = 1$ can be equilibrium strategies for the investor.

Suppose first that $q = 0$, so

$$\pi_I = 1 - (1 - br)p_0, \tag{12}$$

$$\pi_T = b(1-r)p_0. \tag{13}$$

If $p_0 = 1$, then clearly the trustee should choose $r = 0$. But if $r = 0$, the investor should choose $p_0 = 0$. Thus, there are no equilibria with $p_0 = 1$. If $p_0 = 0$ and $r > 1/b$, the investor increases her payoff by switching to $p_0 = 1$. Thus, the only Nash equilibria when $q = 0$ are $p_0 = 0$, $r \le 1/b$.

Now suppose that $q = 1$, so

$$\pi_I = \begin{cases} 1, & \text{if } r \le t, \\ br, & \text{if } r > t \end{cases} \tag{14}$$

and

$$\pi_T = \begin{cases} 0, & \text{if } r \le t, \\ b(1-r), & \text{if } r > t. \end{cases} \tag{15}$$

Then clearly the trustee always has an incentive to switch to $r = t + \epsilon$. An investor's p_0 does not affect her payoff, so any $p_0 \in [0, 1]$, with $r = t + \epsilon$, is a Nash equilibrium. (As we saw in Fig. 1(a), when $q \approx 1$, there is neutral drift around $p_0 = 1$.)

Next, suppose that $1/b \le t \le q < 1$. To see that $p_0 = 1$ and $r = t + \epsilon$ is a Nash equilibrium, we argue as follows. First, the investor has no incentive to lower her p_0 since $r = t + \epsilon > 1/b$, and an investor maximizes her payoff by choosing $p_0 = 1$ when a trustee returns $r > 1/b$. Similarly, the trustee has no incentive to increase his r since doing so would just reduce the fraction of the transfer he retains for himself.

For the trustee to have no incentive to choose a lower r, the following inequality must hold:

$$b(1 - t - \epsilon)(p_0 - p_0 q + q) \ge b(1 - r)(p_0 - p_0 q). \tag{16}$$

The left side is the trustee's payoff when he returns a fraction $t + \epsilon$ of what he receives and the right side is his payoff when he returns a fraction $r \le t$. Since $p_0 = 1$, we obtain the inequality

$$q \ge \frac{t + \epsilon - r}{1 - r}, \tag{17}$$

which in the limit $\epsilon \to 0$ we can write as

$$q > \frac{t - r}{1 - r}. \tag{18}$$

We want this inequality to hold for all r such that $0 \le r \le t$.

Now

$$\frac{d}{dr}\left(\frac{t - r}{1 - r}\right) = \frac{t - 1}{(1 - r)^2} < 0, \tag{19}$$

so $(t - r)/(1 - r)$ is maximized at $r = 0$, where it equals t. Thus, $q > t$ ensures that $p_0 = 1$, $r = t + \epsilon$ is an equilibrium. In Section 4.4, we saw that there are oscillations when $t \ge 1/b$ and $0 < q \le t$.

Finally, suppose $t < 1/b$ and $0 < q < 1$. We claim $p_0 = 0$ and $r = t + \epsilon$ is an equilibrium. Certainly the investor has no incentive to increase her p_0 since $r < 1/b$. The trustee is getting the maximum possible transfer when the investor has information and would not get more by increasing his r. Conversely, the trustee would lose the transfer (in the presence of information) by decreasing his r. Since neither the investor nor the trustee has an incentive to change strategies, $p_0 = 0$, $r = t + \epsilon$ is an equilibrium. It is straightforward to see that there are no other equilibria when $t < 1/b$, so we omit the argument. □

4.6. Optimal trustee strategies when investors have fuzzy minds

In this section, we determine the optimal collective fuzziness f for investors. Recall that when the investor knows the trustee's r, she makes the transfer with probability $1/(1 + e^{-s(r-t)})$, where we have set $s = 1/f$ for notational convenience (s can be thought of as the "sharpness" of the investor). When r is small, the investor transfers essentially nothing in expectation. When r is large, the investor transfers essentially 1 in expectation. There is a steep transition in the expected amount transferred at $r = t$. The parameter s controls how steep this transition is. The larger s is (the smaller f is), the steeper the transition. The constant response of an individual who ignores the available information is obtained by taking $s = 0$ ($f \to \infty$). The "binary" response of the perfectly rational individual is obtained by taking the limit $s \to \infty$ ($f = 0$). If the investor does not know the trustee's r, then she makes the transfer with probability p_0 (i.e., she transfers the amount p_0 average). Hence, the expected amount transferred by the investor is

$$q\frac{1}{1 + e^{-s(r-t)}} + (1 - q)p_0. \tag{20}$$

We have considered "deviations from rationality" in which s varied but the return fraction t "demanded" by investors was the rational $1/b$. Here we will consider the general case in which t might not be $1/b$.

To see why the payoff to investors is maximized at an intermediate value of f, we argue as follows. When investors are perfectly sharp ($f = 0$, or $s \to \infty$), there is no incentive for trustees to be more than just marginally trustworthy. An informed investor with $f = 0$ will make the transfer as long as a trustee's r is greater than $1/b$. When $f > 0$ (s is finite), however, an increase in r always leads to an increase in the transfer probability. Thus, trustees have an incentive to increase their r's above $1/b$. At the same time, a larger r means more of the transfer is sent back to the investor. These two forces balance each other out at some intermediate value of $r > 1/b$. We now formalize this argument.

Suppose the investor transfers with probability p and the trustee returns a fraction r. The (expected) payoff to the trustee is $pb(1 - r)$. Now let r change to $r + \Delta r$ and p change to $p + \Delta p$. Then the new payoff to the trustee is

$$b(p + \Delta p)(1 - r - \Delta r). \tag{21}$$

After some arithmetic, we find that the new payoff will be larger than the old payoff when

$$(1-r)\frac{\Delta p}{\Delta r} > p + \Delta p. \tag{22}$$

Letting $\Delta r \to 0$, we obtain

$$\frac{dp}{dr} > \frac{p}{1-r}. \tag{23}$$

When the inequality (23) holds, trustees should increase the fraction r that they return.

We take p to be the function of r given by

$$p(r) = \frac{1}{1+e^{-s(r-t)}}. \tag{24}$$

Now

$$\frac{dp}{dr} = \frac{se^{-s(r-t)}}{(1+e^{-s(r-t)})^2} \tag{25}$$

and so the condition $dp/dr > p/(1-r)$ becomes

$$(s(1-r)-1)e^{-s(r-t)} > 1. \tag{26}$$

Taking the logarithm of both sides and assuming $s \gg 0$, we obtain

$$r < \frac{\log s + st}{s+1}. \tag{27}$$

Thus, as long as r satisfies (27), rational trustees will have an incentive to increase their r's. The estimate for r given by (27) is a one-humped function of S, consistent with simulation results (and explaining why the average investor payoff is a one-humped function of f). In the limit of perfect rationality, $S \to \infty$, (27) reduces to $r < t$. We thus obtain in another way our earlier result that r evolves to approximately t. Note that the foregoing analysis only applies when $q \approx 1$.

For all but the smallest values of s, we have

$$\frac{\log s + st}{s+1} > t. \tag{28}$$

Hence, when investors are not perfectly rational, the average r is larger than t. This effect is dampened by the fact that q is generally less than 1, but it provides an explanation for why there is a general increase in how much trustees return when investors have fuzzy minds.

A straightforward calculation (after more approximations) shows that the s that maximizes r is given approximately by

$$s_{\text{opt}} \approx e^{1+t} + 1. \tag{29}$$

Collectively "fuzzy" investor minds are advantageous to investors, but it is not the case that investor payoffs always increase as $f = 1/s$ increases: there is an optimal level of fuzziness given by

$$f_{\text{opt}} = \frac{1}{e^{1+t}+1}. \tag{30}$$

This approximate formula is consistent with our simulation results up to an order of magnitude.

4.7. Evolution of the investor decision rule

Thus far, we have been assigning all investors the same fuzziness f and demand t. What happens if the fuzziness and demand parameters are subject to evolution? We find that evolution favors rational self-interest, $f \to 0$, and demands that ensure just an $\epsilon > 0$ profit, $t \to 1/b$. Hence, the investor trust p_0 stays high, but the trustee return r evolves back to just slightly more than $1/b$. Collective irrationality can be seen as a public good for investors. It is best for all investors to coordinate and demand a large return fraction t. Yet a mutant investor who demands $t - \epsilon$ earns more: she engages in more transactions with trustees, who are returning large fractions because of the demands of other investors. The same logic applies to f. A collectively confused investor population results in higher investor payoffs. But a mutant with a sharper mind (smaller f) makes more deals and performs better. Evolution thus leads to a "tragedy of the commons" in which investors become more and more perceptive as their decision rules converge to the perfectly rational step function, resulting in ever lower investor payoffs. While there is an optimal level of irrationality for the investor population as a whole, selection works against collusion.

To gain further intuition regarding these results, we again consider our game as a hybrid of trust and bargaining. We saw that the average payoff of a rational, self-interested investor increases only slightly as q increases. This makes sense because, in the bargaining situation, the responder (the investor) should accept any positive offer, however small. This is the unique subgame perfect Nash equilibrium in the ultimatum game. Second, we saw that fuzzy minds and unreasonable demands

lead to higher investor payoffs. This makes sense because responders in the ultimatum game can improve their payoffs by making "threats" (i.e., by deviating from the equilibrium).

Given that both forms of irrationality are not evolutionarily stable, the presence of these effects in our behavioral experiment is surprising. Explaining the maintenance of these behavioral characteristics is an important direction for future study.

5. Partner choice

So far we have been considering interactions involving a single investor and a single trustee. What happens if investors can choose among multiple trustees? Partner choice, and the related phenomena of dynamic interaction networks, are mechanisms for promoting prosocial behavior that have received considerable attention in both experimental (Barclay, 2004; Barclay and Willer, 2007; Rand et al., 2011; Wang et al., 2012) and theoretical (Santos et al., 2006; Pacheco et al., 2006; Fu et al., 2008; Skyrms and Pemantle, 2000) studies in recent years. Here we study the effects that partner choice has on the evolution of trust.

Suppose that when an investor wants to make an "investment," she begins by selecting k trustees uniformly at random. The investor knows a given trustee's return fraction with probability q, and her knowledge of any one trustee's return fraction is independent of her knowledge of those of other trustees. We then assume that the probability that a particular trustee (out of the k) gets the transfer is proportional to his return fraction if it is known to the investor and to an imputed return fraction if it is not known (see below for details). As Fig. 6 shows, if $k > 1$, then the average investor payoff increases significantly with the information q. Furthermore, the critical information level q^* to achieve maximum trust is a decreasing function of k. We will see that the investor must have a probability $1/b$ of knowing the actual r of at least one trustee in the comparison set for selection to favor trust. Thus, $q_k* = 1 - \left(1 - 1/b\right)^{1/k}$.

We now make this analysis precise. Earlier, we captured the notion of "irrationality" by allowing the investor's transfer probability to be a sigmoidal function of r. More precisely, in the fraction q of cases in which the investor knows the trustee's r, she makes the transfer with probability

$$\frac{1}{1 + e^{-(r-1/b)/f}}. \tag{31}$$

We called f the "fuzziness" of the investor's mind. When $f \to 0$, the investor always makes the transfer if $r > 1/b$ and never makes it if $r < 1/b$. When $f \to \infty$, or when $r = 1/b$ and f is arbitrary, the investor makes the transfer with probability 1/2.

We will reformulate this decision rule slightly (without changing its essence) so that we can extend it to the case of multiple trustees. An investor of "fuzziness" f makes the transfer to an investor with known return fraction r with probability

$$\frac{r^{1/f}}{r^{1/f} + (1/b)^{1/f}}. \tag{32}$$

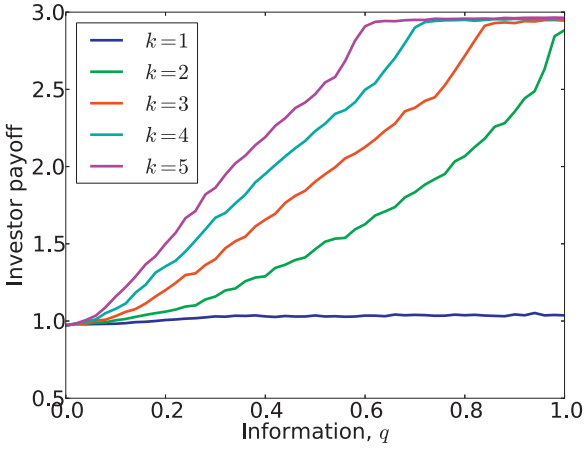

Fig. 6. Investor payoffs and partner choice. The average investor payoff as the information, q, and the number of trustees compared, k, vary. When $k > 1$, investors do better when more information is available. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, the selection intensity $\beta = 20$, and the multiplier $b = 3$. Results are averaged over 50 simulation runs, each run consisting of 50, 000 rounds and each round consisting of 500 games between randomly chosen investors and trustees.

In general, (32) is a sigmoidal function of r. The sigmoid is steepest when $r = 1/b$. The smaller f is, the sharper the transition (i.e., the less fuzzy the investor's mind). We note that the f of (32) is not precisely the same as the f of (31), though they are functionally equivalent. As $f \to 0$,

$$\frac{r^{1/f}}{r^{1/f} + (1/b)^{1/f}} \to \begin{cases} 0, & \text{if } r < 1/b, \\ \frac{1}{2}, & \text{if } r = 1/b, \\ 1, & \text{if } r > 1/b. \end{cases} \tag{33}$$

Thus, the limit $f \to 0$ yields the behavior of the perfectly rational investor.

Suppose now that when an investor wants to make an "investment," she begins by selecting k trustees uniformly at random. The trustees return fractions $r_1, r_2, \ldots, r_k$ of what they receive. The investor knows trustee i's return fraction r_i with probability q, and her knowledge of any one trustee's return fraction is independent of her knowledge of other trustees' fractions.

Without loss of generality, suppose the investor knows $r_1, \ldots, r_j$ but does not know $r_{j+1}, \ldots, r_k$. Her decision rule is as follows. She chooses to make the transfer to trustee i, $1 \le i \le j$, with probability

$$\frac{r_i^{1/f}}{r_1^{1/f} + \cdots + r_j^{1/f} + k(1/b)^{1/f}}. \tag{34}$$

She chooses to make the transfer to trustee i, $j < i \le k$, with probability

$$\frac{p_0(1/b)^{1/f}}{r_1^{1/f} + \cdots + r_j^{1/f} + k(1/b)^{1/f}}. \tag{35}$$

And she chooses to withhold the transfer with probability

$$\frac{(k - p_0 k + p_0 j)(1/b)^{1/f}}{r_1^{1/f} + \cdots + r_j^{1/f} + k(1/b)^{1/f}}. \tag{36}$$

Suppose $k = 1$. If the investor does not know the trustee's return fraction, she makes the transfer with probability

$$\frac{p_0(1/b)^{1/f}}{(1/b)^{1/f}} = p_0 \tag{37}$$

and does not make the transfer with probability

$$\frac{(1 - p_0 + p_0 \cdot 0)(1/b)^{1/f}}{(1/b)^{1/f}} = 1 - p_0. \tag{38}$$

If she does know the trustee's return fraction r_1, she makes the transfer with probability

$$\frac{r_1^{1/f}}{r_1^{1/f} + (1/b)^{1/f}}, \tag{39}$$

which is precisely (32), and does not make the transfer with probability

$$\frac{(1/b)^{1/f}}{r_1^{1/f} + (1/b)^{1/f}}. \tag{40}$$

Thus, our generalized rule (34) agrees with the original one when $k = 1$.

To motivate the form of the rule when $k > 1$, suppose that a risk-neutral investor is generally very trusting, $p_0 \gg 0$, and suppose that she is comparing a trustee whose r she knows to a trustee whose r she does not know. The investor would prefer a trustee who is known to return $r > 1/b$ to one about whom she has no information, and she would prefer a trustee about whom she has no information to one who is known to return $r < 1/b$. Thus, a trustee of unknown r is effectively the same as one whose r is known to be $1/b$, at least as far as the investor's decision process is concerned. However, the less trusting an investor is, the less weight she will assign to a trustee of unknown r. Thus, we scale the imputed return of $1/b$ by the level of trust p_0, leading to (34).

Fig. 7(a) shows the average investor trust, p_0, as the number of trustees compared, k, and the information, q, vary. We assume all investors are perfectly rational, $f \to 0$, and that the multiplier b is 3. When $k = 1$, p_0 at first increases with q, reaching its maximum value when $q = 1/b$. The decrease in p_0 for $q > 1/b$ is a result of the decreasing selection pressure on p_0. When investors often know the return fractions of trustees, the investors' p_0's have little bearing on their payoffs. Ultimately, when $q = 1$, investors always have information about trustees and never use their p_0's, resulting in neutral drift around the average $p_0 = 1/2$ as in our analysis without partner choice (see Section 3.1 and Fig. 1).

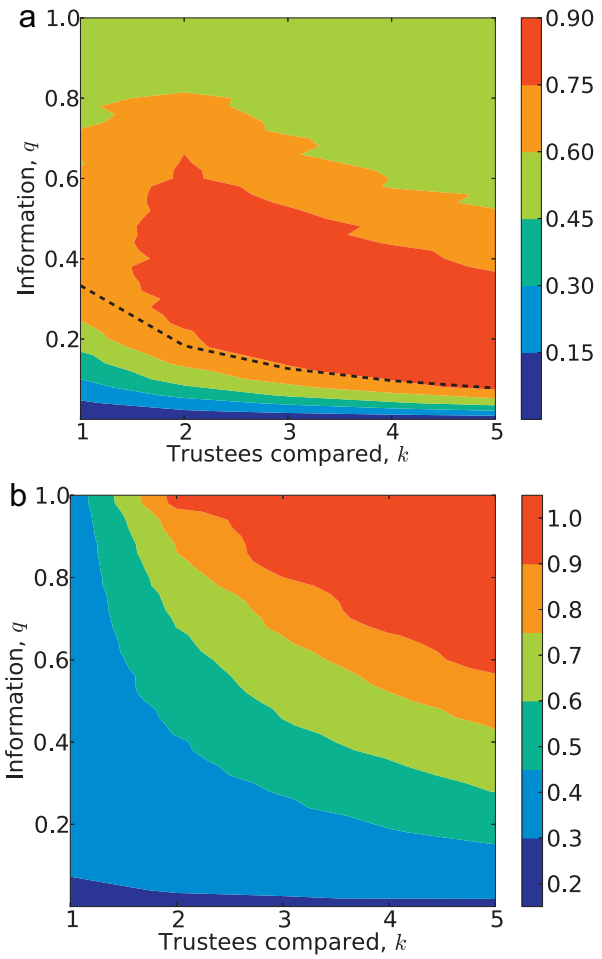

Fig. 7. Partner choice (a) The average trust, p_0, as the information, q, and the number of trustees compared, k, vary. Investors are perfectly rational. As k increases, the level of information, q, needed to achieve trusting behavior decreases. When the investor compares k trustees, the probability that she knows the return fraction of at least one of them must be at least $1/b$ for trusting behavior to be selected (dashed line). The decrease in p_0 for fixed k as $q \to 1$ is a result of decreasing selection pressure on p_0. (b) The analogous plot for the average return, r. When $k = 1$, the average r plateaus at a value of $1/b + \epsilon$ once $q \geq 1/b$. When $k > 1$, on the other hand, competition between trustees results in the selection of higher return fractions. We use the following parameters: the total population size $N = 100$, the mutation rate $\mu = 0.01$, the selection intensity $\beta = 20$, and the multiplier $b = 3$. Results are averaged over 50 simulation runs, each run consisting of 50, 000 rounds and each round consisting of 500 games between randomly chosen investors and trustees.

When $k > 1$, p_0 reaches its maximum value at some $q_k{}^* < 1/b$. The larger k is, the smaller $q_k{}^*$ is. From Theorem 1, we know that the probability q that an investor knows a trustee's r must satisfy $q \geq 1/b$ for $p_0 = 1$ to be an equilibrium. Hence, $q_1{}^* = 1/b$. When partner choice is possible, the generalization of this condition is the following: the probability that the investor knows the return fraction of at least one of the k trustees must be at least $1/b$. We can write this as

$$1 - (1 - q)^k \geq \frac{1}{b}, \tag{41}$$

and so

$$q_{k*} = 1 - \left(1 - \frac{1}{b}\right)^{1/k}. \tag{42}$$

The dashed line in Fig. 7(a) is the curve (42). Thus, the more potential partners an investor considers before engaging in a transaction, the less information she needs (about any one of them) for trusting behavior to be selected.

Fig. 7(b) is the analogous plot for the average return, r. When $k = 1$, the average r attains its maximum value of $1/b + \epsilon$ at all $q \geq 1/b$. Trustees are capturing essentially all the profits, leaving investors with just a minimal positive return. When $k > 1$, on the other hand, the average r has a maximum value (achieved at some $q > 1/b$) significantly greater than $1/b + \epsilon$.

We can understand this intuitively as follows. When $k = 1$, all a trustee must do to satisfy a perfectly rational investor is return a positive amount, however small. When $k > 1$, on the other hand, a trustee who increases his return fraction r beyond $1/b + \epsilon$ increases his probability of being chosen by the investor (see (34)). At the same time, a larger r means that

the trustee is returning more of what he receives to the investor when he does receive the transfer. A trustee should keep increasing his r until these two forces balance out, i.e., until his marginal payoff is zero. This optimizing behavior results in return fractions that are significantly greater than $1/b$. And the larger k is, the more intense the competition and the larger the average r. Partner choice, even when the choice is just between $k = 2$ possible trustees, therefore results in the selection of return fractions significantly greater than $1/b + \epsilon$.

If humans evolved in situations in which individuals had a choice as to whom deserved their trust, the levels of trustworthiness observed in laboratory experiments would then be favored by natural selection. Indeed, small, face-to-face groups—which are able to enforce norms at low cost through mechanisms such as gossip—were "the human social environment for many thousands of generations, prior to the advent of agriculture only about 13,000 years ago," and it has been argued that "economic assumptions about human social preferences should be based upon the psychological traits that evolved to enable human groups to function adaptively at this scale" (Gowdy et al., 2012). Partner choice in such small groups is thus a plausible explanation for the evolution of more than just marginal trustworthiness. Furthermore, this logic need not rest on assumptions about how the environment of ancestral humans shaped genetic evolution. Partner choice is also a powerful factor in many modern day social interactions, such as the choice of friends or business partners. Thus, the development of strategies through social learning over the course of a single lifetime could also lead to the same trusting and trustworthy outcomes.

6. Discussion

We have studied how information fundamentally changes the trust game and transforms it into something akin to an ultimatum game. The resulting amalgam of trust and bargaining leads to the evolution of prosocial behavior. In a related study, McNamara et al. added information in a different way to a discrete version of the trust game (McNamara et al., 2009). In their model, trustees who received the transfer faced a binary choice—they could keep everything or return a fair amount to the investor. Investors facing a particular trustee could pay a cost to learn how that trustee acted in n random previous interactions, where n is an exogenously fixed parameter of the world in which that game occurs. Investors were characterized by a strategy parameter l indicating the number of times (out of the n) that the trustee must have chosen to make the fair return in order for the investor to be willing to make the transfer. McNamara et al. found that when $n > 1$ (or when $n = 1$ and the mutation rate is sufficiently high to maintain a significant amount of diversity among trustees), trust and trustworthiness can arise. The most salient difference between their setup and ours is that in their model information about trustees is always available, such that the decision to trust involves choosing not to pay to access that information; whereas in our model, information is sometimes available (at not cost) and other times unavailable, such that the decision to trust involves transferring "blind" in anonymous games where nothing is known about the trustee. Thus our model helps to explain behavior in laboratory experiments, where investors have no option of buying information about their trustees. We have shown that even a relatively small probability of knowing a trustee's return fraction is sufficient to lead to trust and trustworthiness, a condition which is likely fulfilled in a wide range of real-world scenarios. We also make the interesting observation that if the level of trustworthiness demanded by investors—our t, comparable to the fraction l/n in McNamara et al.'s model—exceeds the information q, then investors and trustees are locked into cycles in which trust and trustworthiness emerge and then collapse in never-ending oscillations.

7. Conclusion

Information and the investor response to information have surprising implications for the payoffs of investors and trustees. Counter-intuitively, giving investors information about trustees benefits trustees more than investors. This is a general phenomenon that is true across a wide range of situations: when one player knows the other's decision, the first player is constrained (if rational) to play a best response, and this can improve payoffs for the second player (Maynard Smith, 1982). For example, Pen and Taylor (2005) show that giving workers information about the queen's decision in a sex allocation game can benefit the queen more than the workers. But when the player with the information has a fuzzy mind, this effect is dampened and the payoff for the "confused player" can increase. This paradoxical situation is in contrast with results showing that errors generally decrease payoffs.

While investors do better if they are collective irrational, such coordination is not evolutionarily stable. Using (limited) information to choose between several trustees provides a way for investors to improve their payoffs in a stable manner. Thus, information (or "reputation") effects together with "partner choice" explain, both qualitatively and quantitatively, the high degrees of trust and trustworthiness exhibited by humans in one-shot anonymous interactions. Our results elucidate the mechanism that lead to the evolution of trust and are more than just a reformulation of the proximate causes of this altruistic behavior (Wilson and Gowdy, 2012). In related work, we have similarly shown have an evolutionary approach can explain experimentally observed behavior in the context of the centipede game (Rand and Nowak, 2012), the traveler's dilemma (Manapat et al., 2012), and anti-social punishment of cooperators (Rand et al., 2010; Rand and Nowak, 2011). Furthermore, in the present study, an emphasis on the evolutionary dynamics and not just on stable equlibria—which often do not exist, as we have seen—provides a more complete understanding of the forces shaping trust and trustworthiness. An

awareness of these dynamical issues makes the formulation of pro-social public policy more plausible, an advantage that the evolutionary approach has over standard (and static) economic analysis.

Acknowledgements

We thank Max Bazerman, Carl Bergstrom, Anna Dreber, Tore Ellingsen, Deepak Malhotra, Hisashi Ohtsuki, and Daniel Rosenbloom for helpful comments. Support from the John Templeton Foundation and the NSF/NIH joint program in mathematical biology (NIH grant R01GM078986) is gratefully acknowledged.

References

Amir, O., Rand, D.G., Gal, Y.K., 2012. Economic games on the internet: the effect of $1 stakes. PLoS One 7, e31461.
Barclay, P., 2004. Trustworthiness and competitive altrusim can also solve the "tragedy of the commons". Evolution and Human Behavior 25, 209–220.
Barclay, P., Willer, R., 2007. Partner choice creates competitive altruism in humans. Proceedings of the Royal Society B 274, 749–753.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games and Economic Behavior 10, 122–142.
Bohnet, I., Huck, S., 2004. Reputation and repetition: implications for trust and trustworthiness when institutions change. American Economic Review 94, 362–366.
Bohnet, I., Zeckhauser, R.J., 2004. Trust, risk and betrayal. Journal of Economic Behavior and Organization 55, 467–484.
Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90, 166–193.
Brandt, H., Sigmund, K., 2005. Indirect reciprocity, image scoring, and moral hazard. Proceedings of the National Academy of Sciences of the United States of America 102, 2666–2670.
Camerer, C.F., Hogarth, R.M., 1999. The effects of financial incentives in experiments: a review and capital-labor-production framework. Journal of Risk and Uncertainty 19, 7–42.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817–869.
Cox, J.C., 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260–281.
Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Review of Economic Studies 74, 685–704.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268–298.
Falk, A., Fishbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.
Fehr, E., 2009. On the economics and biology of trust. Journal of the European Economic Association 7, 235–266.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114, 817–868.
Frank, R.H., 1987. If Homo Economicus could choose his own utility function, would he want one with a conscience? American Economic Review 77, 593–604.
Fu, F., Hauert, C., Nowak, M.A., Wang, L., 2008. Reputation-based partner choice promotes cooperation in social networks. Physical Review E 78, 026117.
Fudenberg, D., Imhof, L., 2006. Imitation processes with small mutations. Journal of Economic Theory 131, 251–262.
Glaeser, E.L., Laibson, D.I., Scheinkman, J.A., Soutter, C.L., 2000. Measuring trust. The Quarterly Journal of Economics 115, 811–846.
Gowdy, J.M., Dollimore, D., Witt, U., Wilson, D.S., 2012. Economic cosmology and the evolutionary challenge. Journal of Economic Behavior and Organization.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.
Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. Experimental Economics 14, 399–425.
Johnson, N.D., Mislin, A.A., 2011. Trust games: a meta-analysis. Journal of Economic Psychology 32, 865–889.
Kandori, M., 1992. Social norms and community enforcement. Review of Economic Studies 59, 63–80.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R., 2005. Getting to know you: reputation and trust in a two-person economic exchange. Science 308, 78–83.
Kosfeld, M., Heinrichs, M., Zak, P.J., Fishbacher, U., Fehr, E., 2005. Oxytocin increases trust in humans. Nature 435, 673–676.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593–622.
Malhotra, D., 2004. Trust and reciprocity decisions: the differing perspectives of trustors and trusted parties. Organizational Behavior and Human Decision Processes 94, 61–73.
Manapat, M.L., Rand, D.G., 2012. Delayed and inconsistent information and the evolution of trust. Dynamic Games and Applications., http://dx.doi.org/10.1007/s13235-012-0055-6.
Manapat, M.L., Rand, D.G., Pawlowitsch, C., Nowak, M.A., 2012. Stochastic evolutionary dynamics resolve the Traveler's Dilemma. Journal of Theoretical Biology 303, 119–127.
Maynard Smith, J., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.
McNamara, J.M., Houston, A.I., 2002. Credible threats and promises. Philosophical Transactions of the Royal Society B 357, 1607–1616.
McNamara, J.M., Stephens, P.A., Dall, S.R.X., Houston, A.I., 2009. Evolution of trust and trustworthiness: social awareness favours personality differences. Proceedings of the Royal Society B 276, 605–613.
Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.-J., 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proceedings of the Royal Society B 268, 2495–2501.
Milinski, M., Semmann, D., Krambeck, H.-J., 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415, 424–426.
Nowak, M.A., Page, K.M., Sigmund, K., 2000. Fairness versus reason in the ultimatum game. Science 289, 1773–1775.
Nowak, M.A., Sasaki, A., Taylor, C., Fudenberg, D., 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428, 646–650.
Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.
Nowak, M.A., Sigmund, K., 2004. Evolutionary dynamics of biological games. Science 303, 793–799.
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 ca maintain cooperation by indirect reciprocity. Journal of Theoretical Biology 239, 435–444.
Ohtsuki, H., Iwasa, Y., Nowak, M.A., 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457, 79–82.
Pacheco, J.M., Traulsen, A., Nowak, M.A., 2006. Active linking in evolutionary games. Journal of Theoretical Biology 243, 437–443.
Panchanathan, K., Boyd, R., 2004. Indirect reciprocity can stabilize cooperation without the second-order free-rider problem. Nature 432, 499–502.
Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on amazon mechanical turk. Judgment and Decision Making 5, 411–419.
Pen, I., Taylor, P.D., 2005. Modelling information exchange in worker-queen conflict over sex allocation. Proceedings of the Royal Society B 272, 2403–2408.
Pfeiffer, T., Tran, L., Krumme, C., Rand, D.G., 2012. The value of reputation. Journal of the Royal Society Interface, doi:10.1098/rsif.2012.0332.
Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.
Rand, D.G., 2012. The promise of mechanical turk: how online labor markets can help theorists run behavioral experiments. Journal of Theoretical Biology 299, 172–179.
Rand, D.G., Arbesman, S., Christakis, N.A., 2011. Dynamic networks promote cooperation in experiments with humans. Proceedings of the National Academy of Sciences of the United States of America 108, 19193–19198.

Rand, D.G., Armao, 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., Greene, J.D., Nowak, M.A., 2012. Spontaneous giving and calculated greed. Nature 489, 427–430.
Rand, D.G., Nowak, M.A., 2011. The evolution of antisocial punishment in optional public goods games. Nature Communications 2, 434.
Rand, D.G., Nowak, M.A., 2012. Evolutionary dynamics in finite populations can explain the full range of cooperative behaviors observed in the centipede game. Journal of Theoretical Biology 300, 212–221.
Santos, F.C., Pacheco, J.M., Lenaerts, T., 2006. Cooperation prevails when individuals adjust their social ties. PLoS Computational Biology 2, e140.
Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. Journal of Economic Theory 97, 273–297.
Skyrms, B., Pemantle, R., 2000. A dynamic model of social network formation. Proceedings of the National Academy of Sciences of the United States of America 97, 9340–9346.
Suri, S., Watts, D.J., 2011. Cooperation and contagion in web-based, networked public goods experiments. PLoS One 6, e16836.
Traulsen, A., Pacheco, J.M., Nowak, M.A., 2007. Pairwise comparison and selection temperature in evolutionary game dynamics. Journal of Theoretical Biology 246, 522–529.
Wang, J., Suri, S., Watts, D.J., 2012. Cooperation and assortativity with dynamic partner updating. Proceedings of the National Academy of Sciences of the United States of America 109, 14363–14368.
Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.
Wilson, D.S., Gowdy, J.M., 2012. Evolution as a general theoretical framwork for economics and public policy. Journal of Economic Behavior and Organization.