Interdependent Preferences and Reciprocity*

Joel Sobel †

October 21, 2004

*I presented a version of this paper at the First World Congress of the Game Theory Society and to my colleagues at the Center for Advanced Study in the Behavioral Sciences. I thank Eli Berman, Antonio Cabrales, Miguel Costa-Gomes, Vincent Crawford, David Kreps, Herbert Gintis, Mark Machina, Efe Ok, Luís Pinto, Matthew Rabin, Paul Romer, Klaus Schmidt, Uzi Segal, Joel Watson, and Kang-Oh Yi for discussions, references, and comments. I am especially grateful to two referees who supplied detailed, intelligent, and constructive comments on an earlier version of the manuscript and to John McMillan for his advice and encouragement. I worked on this project while a Fellow at the Center for Advanced Study in the Behavioral Sciences. I thank my classmates at the Center for conversations and encouragement and the Center for financial and clerical support. NSF funding is also gratefully acknowledged.

†University of California, San Diego. Email: jsobel@ucsd.edu.

Abstract

Experiments, ethnography, and introspection provide evidence economic agents do not act to maximize their narrowly defined self interest. Expanding the domain of preferences to include the utility of others provides a coherent way to extend rational choice theory.

There are two approaches for including extended or social preferences in strategic models. One posits that agents have extended preferences, but maintains the conventional assumption that these preferences are stable. Prominent examples of this approach permit agents to exhibit concern for status, inequality, and social welfare. The other approach permits the strategic context to determine the nature of individual preferences. Context-dependent preferences can capture the possibility that agents are motivated in part by reciprocity. They may sacrifice personal consumption in order to lower the utility of unkind agents or to raise the utility of kind agents.

This paper surveys the evidence in favor of social preferences and describes the implications of the leading theoretical models of extended preferences. It presents behavioral assumptions that characterize different types of social preferences. It investigates the extent to which social preferences may arise as the limit of evolutionary processes. It discusses the relationship between norms of reciprocity and social preferences in repeated interactions. *Journal of Economic Literature* Classification Numbers: C7, D9; Keywords: reciprocity, interdependent preferences, altruism, spite, evolution of preferences.

Contents

1 Introduction

Much of economic analysis stems from the joint assumptions of rationality and individual greed. Common sense and experimental and field evidence point to the limits of this approach. Not everything of interest to economists can be well understood using these tools. This paper reviews evidence that narrow conceptions of greed and rationality perform badly. The evidence is consistent with the view that economic incentives influence decision making. Hence there is a role for optimizing models that relax the assumption of individual greed. I discuss different ways in which one can expand the notion of preferences.

I pay particular attention to how reciprocity influences decision making. Reciprocity refers to a tendency to respond to perceived kindness with kindness and perceived meanness with meanness and to expect this behavior from others. I introduce models of **intrinsic reciprocity** in Section 3.4. Intrinsic reciprocity is a property of preferences. The theory permits individual preferences to depend on the consumption of others. Moreover, the rate at which a person values the consumption of others depends on the past and anticipated actions of others. An individual whose preferences reflect intrinsic reciprocity will be willing to sacrifice his own material consumption to increase the material consumption of others in response to kind behavior while, at the same time, be willing to sacrifice material consumption to decrease someone else's material consumption in response to unkind behavior.

It is more traditional to view reciprocity as the result of optimizing actions of selfish agents. Responding to kindness with kindness in order to sustain a profitable long-term relationship or to obtain a (profitable) reputation for being a reliable associate are examples of **instrumental reciprocity**. Economics typically describes instrumental reciprocity using models of reputation and repeated interaction. This approach is quite powerful as essentially all exchanges in natural settings can be viewed as part of some long-term interaction. Consequently one could argue that the models of Section 3.4 are unnecessary. This essay presents the counter argument that models of intrinsic reciprocity can provide clearer and more intuitive explanations of interesting economic phenomena. An openness to the possibility of intrinsic reciprocity leads to a new and useful perspective on important problems.

The next section contains a stylized example that illustrates the limitations of standard models. I use the example to provide an informal introduction to alternative theoretical approaches and motivate the paper. Section 3 introduces these models formally. Section 4 describes some economic settings in which the modeling approaches of Section 3 may be particularly useful. Section 5 reviews literature on the evolution of preferences. Section 6 responds to stylized arguments against the

approach and Section 7 is a conclusion.

2 An Informal Guide to the Concepts

We regularly read accounts of dissatisfied or recently fired employees destroying property, sabotaging computer files, or even "going postal" and killing people at their workplace. The sense of outrage at an apparent injustice is real. Many people are willing to take destructive actions as part of the outrage. This kind of destructive behavior is unlikely to be in the material interest of a fired employee: it takes time, it is not compensated, and it carries the risk of criminal penalties. How should we think about it? For concreteness, imagine that Paul was a high-ranking executive who had worked for a company for more than ten years. He lost his job when business turned bad. On his last day on the job, Paul destroyed vital company documents and continued to sabotage computer files until he was caught six months later. In this section I will use Paul's story to introduce and motivate the ideas I review in the manuscript.

2.1 Descriptions

There are several ways to react to Paul's destructive activity. One response is to treat it as an emotional response not subject to economic analysis. We should not give up so easily. Paul may be crazy, but his former boss, Marsha, probably is not. Unless Paul's actions are completely unrelated to the environment, Marsha will want to understand how to reduce the chance of adverse behavior. Marsha may need to hire lawyers or psychiatric consultants (instead of economists) to tell her how to deal with Paul or reduce the risk of costly outbursts by employees, but she should evaluate her options using economic models.

I will concentrate on descriptions that are consistent with the hypotheses of optimization and equilibrium. Once we allow that Paul maximizes something more that his own monetary reward, there are many stories like this available. This section introduces some potential descriptions informally. Section 3 provides a more systematic treatment.

2.1.1 Static Income Maximization

Hypothesis. The narrowest version of economic theory assumes that Paul seeks to maximize his utility and that his utility depends on the quantities of the private material goods he consumes. In static income maximization, Paul balances the

immediate cost and benefits of actions rather than the long-term implications of these decisions. In simple settings, this hypothesis reduces to the assumption that Paul maximizes his monetary income.

Analysis. Paul would carry out his destructive action only if he imagined that it would lead to direct material gain. It is hard to rationalize Paul's behavior under these assumptions. His actions may advance his immediate material interests if Marsha gives him back his job or if he receives a payment to stop sabotaging the company, but a more elaborate description seems necessary.

2.1.2 Interdependent Preferences

Hypothesis. Paul maximizes a utility function that depends on Marsha's consumption of material goods in addition to his own.

Analysis. If Paul's utility is decreasing in the material wealth of Marsha, then Paul will be willing to sacrifice his own material well being to punish Marsha. These preferences explain why Paul would wish to harm Marsha, but do not explain why he waits until after he is fired to do so. There are two possibilities. In the midst of an ongoing employment relationship Paul does not harm Marsha because he fears that Marsha will fire him, which would be a sufficiently great punishment to deter him from hurting Marsha.

Alternatively, the marginal rate of substitution between Paul and Marsha's material income in Paul's preferences may change as a result of Paul's termination. The impoverished Paul is willing to sacrifice to make Marsha worse off. This explanation only works if Paul's income after being fired is lower than after a voluntary separation (otherwise any separation would trigger Paul's disruptive behavior).

For the example, it makes sense to assume that Paul's utility is decreasing in Martha's income. The interdependent preference approach permits Paul to be willing to sacrifice material welfare to decrease the income of others.

2.1.3 Preferences over General Consumption Goods

Hypothesis. Paul maximizes a utility function that is a function of "consumption goods" that are derived from marketed goods through a personal production process.

Analysis. This approach generates several possible stories. One possibility is that Paul's behavior demonstrates that he has a marketable characteristic – that is, he is

not the type of person who can be pushed around, he is not afraid to stand up for injustice, or he is capable of identifying weaknesses in a firm's security. By hurting Marsha, Paul gains because he positions himself to get another job (which he may be less likely to lose) or sell a book about his experiences. Under these circumstances, Paul may have preferences defined over both his monetary wealth and his "sense of honor." If the preferences are increasing in both arguments, then he will be willing to make material sacrifices in order to increase in honor. If Paul only cares about honor because it enables him to increase his monetary payoff, then this explanation reduces to income maximization.

Another possibility is that Paul takes pleasure directly from the act of sabotage. That is, his preferences contain an additional argument ("sabotage"). Paul will not maximize his material payoff, but he is selfish and goal oriented. This explanation does not explain why Paul turns to sabotage only after he was fired. Perhaps the advantage of maintaining the employment relationship deterred his urge to destroy files until he was fired, but this explanation suggests that the sabotage levels would be the same whether the employee was fired or separated voluntarily.

2.1.4 Intrinsic Reciprocity

Hypothesis. Paul's utility depends on the material wealth of Marsha. Moreover, Paul's perception of Marsha's behavior determines the direction of Paul's preferences. The marginal utility of Paul with respect to an increase in Marsha's income increases when Marsha is kind to Paul and decreases when Marsha is unkind.

Analysis. Paul believes that Marsha acted unfairly when she fired him. Consequently, his preferences changed and he becomes (more) willing to sacrifice his own income to return the insult he received.

2.1.5 Commitment

Hypothesis. Paul seeks to maximize a utility function that depends only on his material consumption. He can commit himself to taking future actions that are in his best interest when he makes his plans, but may not be in his best interest when he enacts his plans.

Analysis. If fired workers retaliate, Marsha might be reluctant to fire people or she may offer attractive separation packages. These actions benefit workers, so if workers could commit to destructive actions after being fired, then it might be in

their interest to do so.[1] Marsha fired Paul because she concluded that he would do more damage as an employee than not, but there is no explanation for why Paul carried out his threat after he was fired. Standard equilibrium concepts rule out this kind of commitment. For this reason, I do not treat commitment as a description consistent with equilibrium and optimization. I discuss evolutionary arguments for why individuals may maintain commitment ability in Section 5.

2.1.6 Repeated Games

Hypothesis. Paul seeks to maximizes his material consumption, but he views the relationship as ongoing. More precisely, he is engaged in a repeated game with Marsha, and he seeks to maximize a discounted sum of single-period payoffs.

Analysis. Actions have implications for future payoffs in repeated games, so equilibrium behavior does not require short-term maximization. It is natural to assume that Paul's destructive action imposes a short-term cost on Paul but an even greater cost on Marsha, punishing her for firing him. Equilibrium strategies in repeated games often specify that one player punish another player following a deviation from equilibrium behavior. In the simplest cases, potential punishments deter the behavior that would trigger them. So one would never see punishments. When Marsha is uncertain about Paul's willingness or ability to sabotage, it might be in her best interest to take actions that lead to Paul's destructive behavior.

This hypothesis works like the commitment hypothesis in this example, and it has the same problem: Paul lacks incentives to carry out his threat after he loses his job. For the repeated-game hypothesis to apply, Paul must expect to receive benefits after he punishes Marsha that he would not receive otherwise. These benefits may come because Paul receives rewards from third parties (friends or future employers) after he punishes Marsha.

2.2 Sorting out the Explanations

In this section I have presented several explanatory models in an informal context. At a conceptual level, it is sometimes difficult to differentiate the models. I make an attempt to do so in the next section, where I discuss the explanations more precisely in an explicit game-theoretic context. Nevertheless, the distinction is often linguistic:

[1] Alternatively, Marsha might institute tighter security or announce that she will seek severe punishments for destructive activities.

The different approaches sometimes just use different terminology to describe the same thing.

At an empirical level, the different descriptions identify different reasons for Paul's behavior. Fully specified models will lead to different ways to organize the work environment in response to the threat of dissatisfied workers. On the other hand, given a particular observation, it is usually possible to construct a "just-so" story from the perspective of one's favorite class of descriptions that is consistent with the observation. My taxonomy does not generate a fully specified testable model in each category, but rather a family of models with the same underlying mechanism. Rejecting a model is easy, rejecting an approach is nearly impossible.

Natural formulations of the different hypotheses do have different implications in some situations, however, in part because they have different conceptions of what the benefit of Paul's actions are. Factors that may distinguish the descriptions are how long Paul has worked for Marsha and the nature of their relationship, how widely known his destructive actions become, how much discretion Marsha has in her decision, and how old Paul is when he is fired. I close the section with a few examples.

In the repeated-game explanation, Paul punishes Marsha to influence his future returns. If his future returns decrease with age and the cost of sabotage does not change, then the older Paul is, the less likely he is to sabotage Marsha. It is not necessary in the repeated-game story for Paul's behavior to depend on the amount of severance pay he received or the number of other employees laid off. Simple stories based on intrinsic reciprocity would not predict Paul's behavior to depend on his age. Stories based on interdependent preferences would predict that the wealth of Marsha and other employees would influence Paul's behavior.

On the other hand, considerations based on reciprocity are largely retrospective. Paul would be less inclined to punish Marsha if he thought that she had no choice but to fire him or if steps were taken to inform him and reward him for service prior to separation. One would expect destructive behavior to decrease with the amount of goodwill Marsha has accumulated during their relationship. The consequences of Paul's behavior after he is fired are less important under the hypothesis of reciprocity than in the repeated-game story or general consumption good story.

The simplest explanation of Paul's behavior based on reciprocity involves only the relationship between Paul and Marsha. Paul destroys data to hurt Marsha. He may want it known that something bad has happened to the company because this revelation may hurt Marsha, but he gains nothing from advertising his own connection to the crime. For at least some of the explanations based on generalized consumption goods or repeated games, it is important Paul's association with the

sabotage to become public.

The interdependent-preference hypothesis predicts less punishment if it is clear that Marsha is suffering material losses when she fires Paul. It would be to Marsha's advantage to coordinate firings with reductions in pay of employees under these circumstances.

The different implications of the hypotheses imply that more precise versions of the stories can be supported or rejected by data. The different hypotheses also suggest different ways for Marsha to modify the environment to improve outcomes. If Marsha thought that the repeated-game or generalized consumption good hypotheses were the best explanations of Paul's destructive behavior, she would try to reduce sabotage by manipulating Paul's incentives after he leaves the firm. If Marsha thought that Paul was motivated by intrinsic reciprocity, then she would focus on changing behavior during the employment.

3 Models

> Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.
>
> Karl Popper

Several different models have been developed to describe and organize the evidence of non-selfish behavior. No model provides a complete description of the observational findings. A sensible approach will take ideas from different models. This section parallels Section 2. It introduces formal models for the different descriptions of Paul and Marsha's conflict and the theoretical questions raised by the use of these models. I maintain the assumption that individuals have well defined preferences and they behave to maximize their preferences subject to resource constraints. For this reason, there is a clear sense in which all of the behavior I discuss is selfish. I will use the term selfish preferences in a more limited way to mean preferences that do not directly depend on the consumption of others.[2] Preferences are altruistic if they are increasing in the material consumption of others.

The narrowest formulation of rationality demands that an economic agent maximize a utility function that depends only on his current consumption of material goods. This formulation is easily refuted, but it is an unnecessarily restrictive view of

[2]The discussion of evolutionary models in Section 5 permits a less arbitrary identification of selfishness with fitness maximization.

rationality. General models of rational behavior permit a wider range of arguments to enter utility functions. Somewhat arbitrarily, I classify the models according to the way they extend preferences. Section 3.2 looks at models in which an individual cares about people other than himself. Section 3.3 discusses models in which preferences depend on more things than marketed goods.

Paul's behavior towards Marsha illustrates an apparent willingness to risk material well being in order to damage someone else. Spite, outrage, moralistic aggression, and the desire for revenge are behaviors that are as familiar to social scientists as they are inconvenient to the economists' narrow notion of self interest. But while people will go out of their way to harm enemies, they will make sacrifices to help their friends. I will call repaying unkindness with unkindness **destructive reciprocity** and repaying kindness with kindness **constructive reciprocity**.[3] Individuals with an intrinsic preference for reciprocity will have different preferences over allocations depending on the context. Unkind behavior of their neighbors induces destructive reciprocity while kindness induces constructive reciprocity. Section 3.4 examines models in which the process that creates outcomes influences preferences. These models are formal representations of intrinsic reciprocity.

Section 3.5 briefly discusses how commitment power influences predictions. Section 3.6 discusses how the theory of repeated interactions may lead selfish individuals to behave as if they cared about the welfare of others. This provides a foundation for theories of instrumental reciprocity.[4]

Throughout this section I illustrate models using the ultimatum game. The ultimatum game provides a powerful challenge to the hypothesis that income max-

[3]Reciprocity has many definitions, so it is not surprising that adjectives modifying reciprocity take on different meanings depending on the author. Anthropologists Sahlins (1968, page 82) and Service (Service 1966, pages 14 and 15) generously credit each other for definitions of generalized, balanced, and negative reciprocity. While economists do not use the terms generalized and balanced reciprocity, they use negative reciprocity to describe the tendency to punish people who treat you badly. On the other hand, for Sahlins and Service negative reciprocity describes many standard economic "transactions opened and conducted towards net utilitarian advantage" (Sahlins (1968, page 83)). To avoid unnecessary interdisciplinary confusion, I propose the terms destructive and constructive reciprocity as alternatives to negative and positive reciprocity. This is not the only possible source of confusion. Alexander (1987) uses the term direct reciprocity to describe bilateral exchanges of favors: the person receiving a benefit compensates the person who generated the benefit and indirect reciprocity when the return favor comes from a third party. Trivers (1971) uses generalized reciprocity in the sense that Alexander uses indirect reciprocity. Gintis (2000) uses weak reciprocity to describe reciprocal interactions that are instrumental and strong reciprocity in the way that I use intrinsic reciprocity.

[4]Fehr and Gächter (2000) and Fehr and Schmidt (forthcoming) also review of the material in this section.

imization and equilibrium describe economic interactions. In the ultimatum game, two players bargain over the distribution of a surplus of fixed size 1. The first player (proposer) can choose any distribution of the surplus $s \in [0, 1]$. The second player (responder) then either accepts or rejects the proposal. If the responder accepts the proposal s, then the proposer's monetary payoff is $1-s$ and the responder's monetary payoff is s. Otherwise, both receive nothing.

Game theory, assuming that players seek to maximize their monetary payoff, makes two predictions. First, in Nash Equilibrium, all positive offers must be accepted. Second, in subgame-perfect Nash Equilibrium, the proposer must offer $s = 0$ (or, if the set of feasible proposals is discrete, either $s = 0$ or the minimum positive proposal).[5]

The ultimatum game is a beautiful subject for experimental study. It is simple. The combination of payoff maximization and rationality lead to sharp predictions. Experimental subjects repeatedly violate the theoretical predictions.[6] The violations are systematic: low ($s < .2$) proposals are rare; proposals are rejected; proposals rarely give the second player more than half of the surplus (so that $s > .5$ is rarely observed); and equal or nearly equal splits ($.4 < s \leq .5$) are common, occurring more than half the time in typical experiments. In addition, rejections from the responder decrease as s increases. This finding is consistent with, although much weaker than, the stark theoretical prediction since subjects are more likely to choose actions that maximize their material payoffs the larger the (material) benefit associated with doing so. These results continue to hold under a range of conditions.

Before describing the models, I introduce basic notation. Limit attention to a strategic environment with I agents. The strategy set of agent i is S_i. Any strategy profile $s = (s_1, \ldots, s_I)$ (where $s_i \in S_i$ for all $i = 1, \ldots, I$) determines an outcome $O(s)$. Conventional game theory adds to this formulation the assumption that agents have well defined preferences over outcomes. Assume that agent i's preferences can be represented by a utility function $u_i(\cdot)$ defined over outcomes. Without additional assumptions, selfish behavior is not defined. The abstract definition of outcome does not identify a consumption bundle for each agent.

[5]It is a Nash equilibrium for the proposer to offer $s > 0$ and for the responder to accept any offer greater than or equal to s. This equilibrium fails to be subgame perfect because it relies on the responder's threat to reject positive offers less than s.

[6]See Güth, Schmittberger, and Schwarze (1982) for early experiments and Güth (1995b) and Roth (1995) for reviews.

3.1 Income Maximization

In this setting, $O(s) = (x_1, \ldots, x_I)$, where x_i is an allocation to player i and $u_i(O(s))$ is an increasing function of x_i and independent of x_j for $j \neq i$. That is, the outcome consists of private goods allocated to each player and each player cares only about his or her own consumption. In many applications the private goods are one-dimensional monetary payoffs. For simplicity, x_i will refer to a monetary payoff in this section. This model is simple and leads to clear predictions. The predictions are systematically wrong in many interesting situations. It is this model, combined with the assumption that players use (subgame perfect) equilibrium strategies, that leads to the prediction that the first player demands essentially everything in the ultimatum game.

3.2 Interdependent Preferences

This subsection describes models that assume individuals seek to maximize well-defined preferences, and that base predictions on equilibrium behavior, but permit preferences to depend on the consumption of others.

As in the case of income maximization, let $O(s) = (x_1, \ldots, x_I)$ denote the outcome, where x_i is the material allocation of player i. With interdependent preferences, agents care about the distribution of material goods and not simply their own allocation. That is, $u_i(O(s))$ may now depend non-trivially on x_j, for $j \neq i$.

Several authors have proposed specific functional forms for interdependent preferences. For these authors, material allocations are one dimensional – conformable to monetary payoffs in an experiment. To review these models, consider the utility function

$$u_i(x) = x_i + \sum_{j \neq i} \lambda_{ij}(x_i - x_j)x_j. \tag{1}$$

The simplest form of interdependent preferences consistent with (1) arises when $\lambda_{ij}(\cdot)$ is constant. A positive $\lambda_{ij}(\cdot)$ reflects altruism (in the sense that an agent is willing to decrease his own consumption in order to increase the consumption of another agent); a negative $\lambda_{ij}(\cdot)$ reflects spite.

Charness and Rabin (2002) and Fehr and Schmidt (1999) offer specifications that are special cases of (1). These papers impose the further restriction that $\lambda_{ij}(\cdot)$ is independent of i and j and depends on only the sign of $x_i - x_j$.

Charness and Rabin (2002) opt for an average of functional forms that place positive weight on the selfish monetary payoff, the monetary payoff received by the

least well off agent, and the total payoff. With this specification, $\lambda_{ij}(\cdot) > 0$, but is greater when $x_j > x_i$. That is, individual i always places positive weight on the consumption of others and places more weight on the consumption of individuals poorer than he is than on richer individuals.

For the inequity aversion approach suggested by Fehr and Schmidt, $\lambda_{ij}(\cdot)$ is positive if $x_i > x_j$ and negative if $x_i < x_j$. Under this specification, an agent cares about his own monetary payoff and, in addition, would like to reduce the inequality in payoffs between the two players. Bolton and Ockenfels's (2000) ERC (for "Equity, Reciprocity, and Competition") model has a similar motivation, but proposes a utility function that is not in form (1). Instead, Bolton and Ockenfels assume agent i's preferences are an increasing (possibly non-linear) function of x_i and agent i's relative income ($\frac{x_i}{\sum_{j=1}^{N} x_j}$).

The simplest model of interdependent preferences provides the flexibility to some of observed violations of income maximization in the ultimatum game. For example, if the responder is spiteful, so that $\lambda_{ij} = -\gamma$ is constant and negative, then when she is offered the share s she will reject it if $s - \gamma(1 - s) < 0$. If both players have $\gamma \in (0, 1)$, then the unique equilibrium of the ultimatum game would be for the first player to offer the second player the share $\frac{\gamma}{1+\gamma}$, which is positive but less than one half.

This resolution is unsatisfactory. Intuition suggests that at least some of the behavior in the ultimatum game comes from generosity and not fear of rejection. One would not expect a spiteful individual to make charitable contributions or give a positive share to his opponent in a dictator-game version of the ultimatum game (in which the second player is required to accept any feasible proposal). Allowing the sign of λ to change depending on the income distribution is consistent with this behavior.[7]

Levine (1998) assumes that the extent to which agents care about another player's material utility is a weighted average of a pure altruism parameter and the altruism parameter of the other player. Levine assumes that people differ in the degree to which they care about others and that people care more about the material payoffs of nice people. Formally, he assumes that i maximizes

$$x_i + \sum_{j \neq i} \frac{\alpha_i + \beta_i \alpha_j}{1 + \beta_i} x_j.$$

[7]Models of interdependent preferences predict all positive offers are accepted when λ_{ij} is non-negative. Hence Charness and Rabin (2002) also combine their functional form with preferences that depend on context in order to explain some of the observed responder behavior in ultimatum games.

Here α_i is the altruism parameter of player i and β_i is the weight player i places on j's preferences. If $\beta_i = 0$, then the weight player i places on j's material payoff is independent of j's degree of altruism; otherwise, the weight is an increasing function of j's altruism parameter. This specification is a special case of (1). In contrast to inequity aversion the weight placed on the material payoff of another player depends on the identity of that player. In Levine's model an individual wants to be kind to a kind person. Levine uses this model to describe experimental results. In order to do so, he assumes that players are uncertain about their opponent's preferences and solves for the equilibrium of incomplete information games. Agents want to identify altruistic (high α) people so that they can be nice to them. Players draw inferences from the strategies of other people. This permits a form of reciprocity to arise in equilibrium. If agents with higher αs choose nicer strategies, then players place higher weights on the material payoffs of people who play nice strategies, because playing nice strategies signals that you really are nice.

Models of interdependent preferences raise theoretical issues about how to determine the arguments of the utility function. Imagine an agent who is motivated by the desire to maximize a utility function that reflects inequity aversion. Exactly whose utility enters into this function and how should this utility be measured? An experimental subject could try to maximize her monetary payoff in the laboratory and then redistribute her earnings to deserving people later. This behavior would be appropriate if the concern for inequity was 'broadly bracketed' (Read, Loewenstein, and Rabin (1999) and Thaler (1999)) in that concerns outside of the laboratory entered into the decision making in the laboratory. If the second player in the ultimatum game learned the proposer was relatively poor, would she be willing to accept small offers?[8] If experimental subjects had to earn the right to play a role in a game in which the equilibrium prediction (assuming narrowly self-interested behavior) gave unequal payoffs would the effect of inequity aversion be reduced? Would it matter whether attractive positions were allocated by scoring high on a test of general intelligence or by a pseudo-random device, like being born on an odd day of the year? Do winners of lotteries attempt to find out the names of the losers in order to reduce inequality (or increase the wealth of the agent who did poorly on this particular transaction)?

Concerns that always arise in experimental settings are especially salient here. Should the experimenter's payoff enter into the subject's utility function? Subjects

[8]Goeree and Holt (2000) attempt to test for this effect in the laboratory. They vary the lump-sum payment received by subjects in a perfect-information bargaining game. Results from the experiment are consistent with the hypothesis that subjects take into account these payments, which are irrelevant to standard models of the bargaining process.

have some idea that the experimenter is budget constrained (or at least the money from the experiment comes from somewhere). So all allocations are just transfers. Since the total monetary payments are constant, utilitarian objectives are not relevant. Also, the subjects must be aware that they make more than one decision during an experimental session. Even if subjects do not bring their lifetime decision problem into the lab, they may impose notions of fairness or social preferences across the entire experimental session rather than just one decision at a time.

An optimizing agent whose utility function places positive weight on the wealth of others could allocate income outside the lab carefully so that one could not infer the true nature of preferences from lab behavior. This agent would optimize his interdependent preferences before entering the laboratory, so at the margin he would be indifferent between allocating his laboratory winnings to increase his personal income or to decrease inequality. If one accepts the possibility that laboratory behavior takes into account decisions made outside of the laboratory, then this argument suggests that experiments overestimate the amount of (narrowly) selfish behavior, since selfish people must be selfish in all situations and others may appear selfish in the laboratory in order to pursue their non-selfish interests outside the laboratory more effectively.[9]

There are tacit assumptions in the models of interdependent preferences. The modeler makes the assumptions when specifying the identities of the players in the game and their initial wealth levels. In applications, subjects are assumed to care only about the welfare of other active participants in the game and to make relative income comparisons based only on the material payoffs of the game. It is typical (and necessary) to identify a "small world" in which to apply decision-theoretic arguments (particularly in models involving choice under uncertainty). The problem seems especially critical when using interdependent preferences, however. In deciding how to interpret these models, one must understand why the subject cares about the utility of other subjects, but not the utility of the experimenter. In deciding how to apply these models to a contracting problem, one must decide whether preferences are defined over co-workers or just the parties to the contract. In deciding how to apply these models to the labor market, one must decide whether workers care about inequity across labor and management, across all workers, or only across workers in similar jobs. One must also decide whether to invoke an interdependent utility function to determine decisions separately or whether agents make decisions that reduce inequity over longer intervals.

[9]On the contrary, one could also argue that subjects may have more incentive to curb their selfish instincts in the laboratory than in other settings if convincing fellow experimental subjects and experimenters that they are not selfish is the best way to gain future riches.

3.3 Preferences over General Consumption Goods

The "Chicago School" pursues the goal of using the optimizing models of self-interested agents constrained by a market environment – a limited endowment, existing prices, and economic institutions – to explain a broad range of economic phenomena. The approach is advocated powerfully in the work of George Stigler and Gary Becker (for example, Stigler and Becker (1977)). The theory explicitly exploits the possibility that self interest has a broad definition. Preferences are not defined over marketed goods, but general commodities that individuals transform into consumption goods. Models in this tradition therefore do not require that experimental subjects base their decisions solely on their own monetary payoffs.

Stigler and Becker's posit that the decision maker (household) has a utility function

$$U(Z_1, \ldots, Z_m) \tag{2}$$

where for $i = 1, \ldots, m$

$$Z_i = f_i(X_{1i}, \ldots, X_{ni}, t_{1i}, \ldots, t_{li}, S_1, \ldots, S_l, Y_i) \tag{3}$$

Z_i are the generalized consumption goods; $f_i(\cdot)$ is the production function for commodity i, X_{ji} is the quantity of the j^{th} market good, t_{ki} in the time input of individual k, S_k is the human capital of the k^{th} person, and Y_i represents all other inputs. Given wages and prices for the market goods, a household selects X_{ji} and t_{ki} to maximize (2) subject to (3). In any application, the definition of market goods will not be controversial. The quantities X_{ji} and t_{ki} will be observable. In the general specification of the theory, however, the levels of human capital, the functional form of the production functions, and, indeed, the nature of generalized consumption goods can be freely selected by the modeler.

While the approach of Stigler and Becker is firmly grounded in ideas familiar to economists – budget constraints and individual maximization – it also reflects a common view among cultural anthropologists. Sahlins (1968, page 9), who provides a useful taxonomy of reciprocity, observes that "in an uncommon number of tribal transactions material utility is played down, to the extent that the main advantages appear to be social, the gain coming in good relations rather than good things." Sahlins recognizes the need to look beyond immediate material gain to understand the workings of simple economies. While Stigler and Becker do not explicitly substitute "good relations" for "good things," their formulation broadens the notion of consumption good and concedes that economic exchange may be motivated by more than short-term material gain even in developed market economies.

When stripped to its mathematical core, the Stigler-Becker model posits that decision makers have preferences over their choice variables and that these preferences depend on something that is determined in part by choices and in part by other factors that are left unconstrained in the basic formulation.[10]

Everything can be written

$$u_i(O(s); \alpha(s; \theta)) \tag{4}$$

where $s = (s_1, \ldots, s_I)$ is strategy profile (individual i chooses s_i), θ describes personal characteristics, and $\alpha(\cdot)$ is a parameter.[11] In making the transformation, s_i represents the choice variables of the household (quantities of market goods and labor); θ denotes the additional parameters (human capital and "other inputs"); $O(s)$ denotes those generalized consumption goods whose production does not depend on θ and $\alpha(s; \theta)$ all other generalized consumption goods.

The description of the utility function in (2) appears to rule out externalities, since as agent's utility is a function only of his own generalized consumption goods. The reduced-form (4) allows utility to depend on the entire outcome, so it permits interdependent utilities. Characteristics of other individuals in the economy enter through the production function: it includes as arguments the human capital of all agents. Further, since the general model makes no restrictions on the relationship between choices of other agents and their level of human capital on one hand and the definition of generalized consumption goods on the other, nothing prevents agent i's optimization problem from having an arbitrary dependence on agent k's market decisions.

What distinguishes this formulation from the models of income maximization and interdependent preferences is that optimizing decisions can be based on more than the distribution of material goods. There are two aspects to this difference. First, the model of generalized consumption goods does not require that the outcome is a vector of observable private allocations. Models with generalized consumption goods permit some of the goods to be standard private goods, but are flexible enough to include public goods. The arguments could also include quantities that cannot be measured directly, like a warm glow from giving. The second difference is that in models of generalized consumption goods preferences over outcomes may vary with parameters. In terms of the representation in (4), an individual's preferences over

[10] In the Stigler-Becker formulation, the household utility function does not depend on the household, but heterogeneity may enter through the human capital variables.

[11] I use this formulation rather than simply $U_i(s, \theta)$ in order to connect this model to explicitly strategic models that I introduce in Section 3.4.

outcomes $O(\cdot)$ can depend on the parameter α, which is difficult to observe, let alone control.

The reduced-form description of the Stigler-Becker model also connects it to an approach that, at least on the surface, appears quite different.

Akerlof and Kranton (2000)'s formulation of identity posits that the decision maker has a utility function

$$U_i(a_i, a_{-i}, I_i(a_i, a_{-i}; c_i, \epsilon_i, P)) \tag{5}$$

where a_i is the action of individual i, $I_i(\cdot)$ is the individual's identity; c_i is the individual's assigned social categories; ϵ_i is individual i's characteristics; and P are prescriptions that "indicate the behavior appropriate for people in different social categories in different situations" (Akerlof and Kranton (2000, page 719)). Given c_i, ϵ_i, P, and a_{-i} individual k selects a_i to maximize U_i. In this theory, action choices are observable.

Akerlof and Kranton's model also reduces to (4) by denoting the action variable a by s, letting θ contain the variables associated with social and individual characteristics and prescriptions, and defining $\alpha(s;\theta) = I_i(a_i, a_{-i}; c_i, \epsilon_i, P)$. The models of Akerlof-Kranton and Stigler-Becker are thus **mathematically** identical. It is curious that these formally equivalent approaches are associated with schools of thought that often are viewed as opposites.

The theories are identical because they are consistent with precisely the same set of observations. Any observation consistent with the first must be consistent with the second and conversely. The theories have different social scientific implications because they lead one to look for different ways to describe observations. For example, the Chicago school may assume that generous behavior is a consumption good that directly enters the utility function (possibly because the appearance of kindness that will be useful in the future). Akerlof and Kranton might posit that an individual's identity required behaving according to accepted norms of fairness (and therefore the proposer loses utility if he offers unequal divisions or the responder loses utility if she accepts unequal divisions).

Fremling and Posner (1999) provide a more elaborate example of this approach. They sketch a model in which an individual's utility depends on status and non-conspicuous consumption. In this formulation, non-conspicuous spending is the category that summarizes expenditures on standard consumption items. Inserting a status argument in the utility function provides a reduced-form meant to capture the instrumental value of increasing status. Agents have the same underlying preferences, but differ in their given endowment of status.[12] Individuals allocate their

[12]The fixed component of status derives from genetic endowment – inherited wealth, titles, race,

income over non-conspicuous consumption items and expenditures that influence their variable component of status. Because agents will forgo consumption to increase their status, this formulation is sufficient to be consistent with many apparent departures from self interest. Fremling and Posner (1999) suggest that dictators will not take the entire surplus in order to signal that they are altruistic. Being known as a generous person enhances your status, which will put you in a better position to advance your material self interest in the future. The decision to sacrifice non-conspicuous consumption for increased status is a standard economic tradeoff. Proposers will not make low offers in the ultimatum game for the same reason and also to avoid challenging the second player's status.[13] Responders reject low offers in the ultimatum game in order to signal that small amounts of money are not important to them. Voluntary contributions to public goods arise if status is enhanced by contributing to charitable projects.

Fremling and Posner (1999) presumably wish to maintain the Chicago tradition and base their explanations of differences in given status rather than preferences. They argue that heterogeneous behavior arises because different agents have different endowments of status, but their formulation provides no way to measure endowments of status. Further, this position is a bit strained response to heterogeneity of laboratory behavior, since experimenters make great efforts to suppress information about given status.

Fremling and Posner's (1999) model is not fully specified. They do not provide a complete, operational definition of status. The status argument that enters their utility function is not observable. Hence there is no way in which it can be controlled in the laboratory.[14] They do not provide an explanation of why status should enter the utility function, nor do they place substantial restrictions on the form that it enters. Yet their motivation is powerful and consistent with casual intuition. There have been successful attempts to incorporate status concerns in preferences,[15] which

---

and gender – in addition to other attributes obtained in the past. Decisions made in the current period cannot influence the fixed component of status.

[13]This effect should arise in the dictator game as well, so the difference in proposer behavior in the two games must depend on some expectation that low offers will be rejected in the ultimatum game.

[14]Alexander's (1987) discussion of the evolution of morality contains discussions consistent with the view of Fremling and Posner. In his discussion of (apparently) altruistic giving on pages 159 and 160, Alexander raises selfish motivations for generous behavior. He is aware of the broadness of the theory and writes (page 160): "If such conditions seem to render the propositions virtually untestable, that is simply a problem that we must solve if we are to deal in a better way with the unparalleled difficulty of understanding ourselves."

[15]Postlewaite (1998) provides an overview of one approach. This article also argues that there are advantages for explicitly incorporating the reasons why status (or other intangible arguments)

means that one can derive preferences like the ones proposed by Fremling and Posner from a more detailed model. Still, the lack of guidance about how status influences preferences is disconcerting. Fremling and Posner (1999) describe situations in which conspicuous spending increases status and others in which conspicuous thrift (for example, buying a modest car or wearing unpretentious clothes) enhances status, which makes me suspect that ex post adjustments in their status variable will make their model consistent with any observation.

While Fremling and Posner (1999) describe their model as a signaling model, it is not completely clear what is being signaled. Implicitly, status is important because it is an observable way for third parties to learn something important about an individual. In the formal model, the only characteristic that may be hidden is an individual's income. If status does signal income (as it does in many of the informal stories), then presumably it is not status that enters the utility function, but income (as perceived by third parties).[16] The function that transforms investment in status into perceived income would be determined as an equilibrium of a signaling game and need not have even the weak properties that Fremling and Posner (1999) posit.

3.4 Intrinsic Reciprocity

Game theory assumes that in strategic situations players act to maximize a preference relation over outcomes. As a result the process by which the outcome is reached does not matter. I argued that models of identity and generalized commodities permit preferences over outcomes to depend on the context in which the outcome was reached. Formally, this means only that α appears in the utility function. This subsection describes a framework in which the preferences that players optimize in a game depend on the game itself.

Context matters in a strategic situation if preferences over outcomes depend on the game being played. The idea is best introduced by an example. Consider two versions of the ultimatum game. In the first version, the proposer can make only two offers: an 80-20 split and a 20-80 split. In the second version, the proposer can make three offers: 80-20, 50-50, and 20-80. In both of these games, there exists an opportunity for the responder to choose between 80-20 division and a 0-0 division. If the responder's preferences depend only on the distribution of material payoffs available when she makes her decision, then her decision after she has been offered 20 cannot depend on whether player one could have proposed an equal split. Intuition

should matter into formal models rather than relying on reduced-forms.

[16]This interpretation suggests that differences in income across experimental subjects would be both relevant for experimental results and difficult to control.

suggests that the additional strategy might matter, with the responder more likely to reject the "unfair" 80-20 split when player one could have offered an equal division than when only unequal splits are available. The experimental results of Falk, Fehr, and Fischbacher (2003) confirm this intuition.[17]

Assuming equilibrium behavior, there are two reasons why adding strategies might influence the responder's choice in the ultimatum game. The first possibility is that adding strategies changes the equilibrium selection. This could only happen if the subgames in which responder decides whether to accept or reject a proposal have multiple equilibria. For appropriately defined preferences over outcomes, the responder may be indifferent between accepting or rejecting twenty units out of one hundred, but it is an implausible way to explain experimental behavior.[18]

The second possibility is that preferences depend on more than final payoffs. In the ultimatum game, if player 2's preferences depend only on final outcomes and she has a strict preference between accepting the 80-20 split and turning it down, then she must make the same choice on the equilibrium path whatever the other strategies may be available to player 1. Hence I concentrate on the possibility that preferences over outcomes at a decision point depend on more than just final payoffs. Player 2's rejection of the 80-20 offer when the equal split is unavailable is an example of destructive reciprocity. In situations where beliefs about the actions of others are irrelevant (for example, when the second player decides whether to accept or reject an offer in the ultimatum game), any descriptive model must permit allow preferences to depend on more than the distribution of income.

Rabin (1993) was the first to propose a specific model of equilibrium behavior in games where players take into account context to determine their behavior. His formulation uses the theory of psychological games introduced by Geanakoplos, Pearce, and Stacchetti (1989). Psychological games permit players' beliefs to enter into their preferences. In Rabin's model, the weight placed on an opponent's material payoffs depends on the interpretation of that player's intentions. He evaluated intentions by using beliefs (and beliefs about beliefs) over strategy choices. Rabin proposed that agent i pick his strategy s_i in a game to maximize a function of the form:

$$u_i(s_i; s^*) = v_i(O(s)) + \alpha_i^G(s^*)v_j(O(s)). \tag{6}$$

[17]Some of the experiments reported in Charness and Rabin (2002) have a similar flavor.

[18]If the responder's preferences over outcomes left her indifferent between 0-0 and 80-20, then for all popular parametrizations of (1), she would have a strict preference between 0-0 and 79-21. Consequently a model that explained the experimental results on the basis of equilibrium selection predicts that small perturbations in the set of feasible offers would dramatically change experimental outcomes.

In this expression, G is the game; $s^* = (s_i^*, s_j^*)$; $s = (s_i, s_j^*)$; $v_i(\cdot)$ denotes player i's utility function over outcomes;[19] and $O(s)$ is the outcome obtained if the players play s. Rabin interprets s_i as the strategy choice of player i, s_j^* as player i's beliefs about player j's strategy choice, and s_i^* as what player i believes that player j believes about player i's strategy choice. In equilibrium, beliefs are accurate, so that j actually plays s_j^* and $s_i = s_i^*$. $\alpha_i^G(\cdot)$ is a function of the beliefs of player i. It is this feature that makes the game a psychological game.

The utility function u_i expresses i's preferences over his own strategies (s_i) conditioned on expected behavior (s^*). A player seeks to maximize a weighted average of his material utility with that of his opponent. The weight $\alpha^G(s^*)$ depends on the game being played in addition to the strategy profile. The natural interpretation of $\alpha^G(s^*)$ is as a measure of the extent to which player i cares about player j's material welfare. The conventional formulation ($\alpha_i^G \equiv 0$) is a special case of this representation. When $\alpha_i^G(\cdot)$ is not constant, as is typically the case in Rabin's model, player i's preferences over strategies depend on more than his preferences over outcomes: the strategic context matters. Rabin presents a specific functional form for the coefficient $\alpha_i^G(\cdot)$ in (6). The form of the coefficient is less important than its content. $\alpha_i^G(\cdot)$ is positive if i thinks that j's behavior is nice and negative if he thinks that the behavior is nasty. In this way, (6) provides a model of intrinsic reciprocity. Kind (unkind) treatment raises (lowers) the weight placed on opponent's material payoff is preferences, making an agent more willing to sacrifice his own material payoff to increase (decrease) that of his opponent.

The optimization problem (4) faced by individual agents in the generalized consumption model appears to include the problem (6) as a special case. There is a subtle difference in the way that one closes the models to compute equilibria, however. In the strategic models, agent i takes context as given and chooses s_i to maximize (6) holding s^* fixed. In equilibrium one must have $s = s^*$. Models based on the generalized consumption good idea often require a consistency condition (the condition may describe the evolution of human capital or the magnitude of a status argument), but in general the decision maker controls s_i as it enters both directly and indirectly in (4). There is a technical implication of the difference. Assuming that preferences over outcomes are linear in probabilities, the standard assumption, existence of equilibrium follows from standard arguments in the game-theoretic models. Existence would not be guaranteed in the Akerlof-Kranton or Stigler-Becker settings without assumptions that lead to quasiconcavity of the reduced-form preferences in (4). The following example illustrates part of the problem.

[19] Rabin (1993) assumes that the outcomes are distributions of money x and that $v_i(\cdot)$ depends only on x_i. Charness and Rabin (2002) relax the second assumption.

Example 1 Consider the game:

	Fight	Opera
Fight	$2,1$	$0,0$
Opera	$0,0$	$1,2$

The matrix describes the material payoffs for a standard battle-of-the-sexes game. Assume that payoffs can be written in the form (6) with $v_i(\cdot)$ as given in the payoff matrix and $\alpha_i^G(\text{Fight}, \text{Fight}) = \alpha_i^G(\text{Opera}, \text{Opera}) = .8$ and $\alpha_i^G(\text{Fight}, \text{Opera}) = \alpha_i^G(\text{Opera}, \text{Fight}) = -.8$ for i = Row and Column. These weights have a natural interpretation: a player who anticipates coordination places positive weight on his or her opponent's material payoff, while one who anticipates a failure to coordinate blames the other player and places negative weight on that person's payoffs. Assume that the players expect the outcome s^* =(Fight,Opera). Then each player will place a negative weight on the other player's payoff and will prefer playing as expected than deviating: That is, (Fight,Opera) is an equilibrium outcome. When (Fight, Opera) is the expected outcome, the man believes that the women is planning to go to the opera even though she believes he is going to the fight. In this situation, he thinks that she is being nasty and is willing to give up the material value of coordination in order to lower her payoff.

Alternatively, one can view the payoffs as the result of Akerlof and Kranton's identity theory. Assume that preferences have the same representation as before. That is, they can be written in the form $u_i(s) = v_i(s) + \alpha_i^G(s)v_j(s)$, where $i \neq j$, $\alpha_i^G(\cdot)$ is the same as before, and the material payoffs $(v_1(s), v_2(s))$ are given in the payoff matrix. Coordination permits the man to feel in charge (if the outcome is (Fight,Fight)) or thoughtful (if the outcome is (Opera,Opera)); failure to coordinate leads to negative α because the man does not want to be viewed as selfish (if the outcome is (Fight,Opera)) or confused (if the outcome is (Opera,Fight)). Now, the outcome (Fight,Opera) is not an equilibrium because given that the column player is going to the opera, the man's best response is to go there as well. Doing so raises his material payoff and also changes his identity (α increases from -.8 to .8).

The difference between the two formulations is that, when considering a deviation, a player can change his "identity" but cannot change his view of his opponent's intentions. In the example, Rabin's model had a larger equilibrium set than Akerlof and Kranton's, but the in general there is no relationship between the two sets. □

The models of context-dependent preferences include the interdependent preference approach. The underlying preferences $v_i(\cdot)$ in (6) are defined over outcomes. If

an outcome specifies a material payoff to both players, it is permissible for v_i to depend on player j's material payoff. Charness and Rabin (2002) propose a functional form that assumes that the v_i are interdependent preferences that place positive weight on j's monetary payoff (with the weight changing to reflect concern for the player with the lowest monetary payoff). Falk and Fischbacher (forthcoming) also attempt to separate concerns for equity and concerns for intentions. Their model contains a "pure outcome concern parameter." This number measures the degree to which a player's preferences in the game depend only on the outcome and not on the context in which it was obtained. At one extreme, a player cares only about the outcome. In this case, Falk and Fischbacher's (forthcoming) model has the flavor of the models of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999).

Since the approach outlined in this subsection generalizes the interdependent preference approach, without restrictions, it has broader descriptive powers. To the extent that preferences over outcomes depend on the game, reciprocity models provide insight into the observations.

There is substantial experimental evidence that the distributional approach is not sufficient to explain and organize experimental findings, which suggests that it would be worthwhile investigating models of intrinsic reciprocity.

Binmore, McCarthy, Ponti, Samuelson, and Shaked (2002) compare one- and two-stage alternative-offer bargaining games in an effort to test whether experimental subjects obey backward induction for these games. The paper observes that subgame-perfect equilibrium strategies played by agents with interdependent preferences satisfy backward induction. They discover systematic departures from backward induction, which is evidence against the hypothesis of equilibrium behavior with interdependent preferences. Context-dependent models permit preferences over outcomes to depend upon how one reached the outcome. Hence the experimental results of Binmore, McCarthy, Ponti, Samuelson, and Shaked do not contradict these models. The fact that these models are consistent with experimental results is a tribute to their flexibility[20] rather than actual support for the formulation.

Costa-Gomes and Zauner (2001) analyzes data from ultimatum game experiments and estimate utility functions of the form $v_i + \lambda v_j$ under the assumption that agents play an equilibrium to a game with perturbed preferences. Their method contains an indirect test of the pure interdependent preference approach. In all terminal nodes in which the second player rejects the proposal, monetary payoffs are the same (zero for each player). If the coefficient λ does not depend on actions, then the estimated variance in the error term should be the same after all rejections. This is not the

[20]Segal and Sobel (2004b) show that dominance arguments have power to rule out some predictions in games governed by preferences represented by (6), but these restrictions are weak.

case, providing some evidence that preferences depend not just on the outcome, but on how the outcome was reached.

There are other ways to demonstrate that preferences depend on more than the final distribution of wealth. Bereby-Meyer and Niederle (forthcoming) compares outcomes of three-player games. The first player proposes a division of a fixed quantity between himself and the second player. The second player either accepts or rejects this proposal. The third player is non strategic. If the second player accepts, then the first and second players get paid according to the proposal (and the third player receives nothing). If the second player rejects, then (depending on the treatment) either the first or the third player receives a payment, while the other two players receive nothing. Bereby-Meyer and Niederle find that the second player's behavior depends both on the quantity of the payment given following a rejection and who receives this payment. The second player is more willing to reject a small offer when she knows that doing so will not lead to a high payoff for the proposer (either because the rejection payment is low or because the third player receives the payment). To explain this behavior with some form of interdependent preferences, the second player would not only need to have preferences that depended non trivially on the monetary payoffs of others, but she would need to weigh an opponent's monetary payoff differently depending on his role in the game. Models that permit preferences over strategies to depend on strategy choices provide a convenient way to capture the intuition that the second player might be willing to sacrifice her material payoff in order to punish player one when the first player makes a small offer. Naturally, the ability to punish (and therefore the second player's action) depends on whether the first or third player receives a payment after a rejection.

Cox (2004) compares the outcomes of three related games designed to separate predictions from different models designed to identify the source of non-selfish behavior in experimental trust games. In the basic experiment (Treatment A), subjects are divided into two groups. Those in the first group receive \$10 and decide how much to contribute to a member of the second group. The member of the second group receives three times the contribution. Finally, members of the second group can return any part of the transfer he or she received. (All transfers are anonymous.) In Treatment B, members of the first group decides on transfers as in the Treatment A while members of the second group do not have a move. In the Treatment C, members of the first group do not move. Instead, experimenters make the same contributions that were made in the Treatment A (and subtract the appropriate amounts from members of the first group). Members of the second group receive transfers as in the Treatment A, are told how the transfers are generated (and how they influence the endowment of first movers), and then decide how much to return.

The contrast between the results of Treatments A and C present the evidence most relevant to the power of the interdependent preferences to describe outcomes. If the second mover cares only about the distribution of payoffs, then contributions should be the same in these two treatments. They are not. The second mover tended to return more money in Treatment A. The difference between the amount returned is largest after large transfers, suggesting that in part the second mover acted to reward intentionally generous behavior.

In contrast to the models that focus on context, adding strategies to a game does not influence a player's preferences over the outcomes in the original game in Levine's (1998) set up. When players are uncertain about their opponents' preferences, however, players may try to use their actions to signal their preferences. The inferences one player draws about his opponent's payoffs depend on the strategies available. Consequently, observed preferences over outcomes depend on the strategic context in Levine's model (because changing the set of available strategies can change the signaling content of strategy choice).

Expanding the set of arguments in the utility function, as in the models of Sections 3.2 and 3.3, is consistent with standard decision theoretic methods. The new arguments in utility functions are externalities and fully captured by expanding the definition of commodity. The models in this subsection are not traditional. One attempt to return the models to the standard framework is to redefine the notion of an outcome. If the way in which one arrives at an outcome influences preferences, then the outcome should include that information. Formally, an outcome would need to include a description of the entire game and an anticipated strategy profile. Specifically, the 80-20 offer in the ultimatum game leads to an outcome that not only specifies that player one receives 80 and player two receives 20, but also the existence of other possibilities available to the first player. This transformation is logically possible, but hardly useful. Note further that the representation (6) specifies preferences over player i's own strategies conditional on a strategy profile s; it is not a utility function defined over outcomes. After redefining the outcomes, one would need to extend the preferences to this new space.

There are several problematic aspects of models based on context-dependent preferences. The most basic problem is the specification of α. Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fishbacher (forthcoming), and Rabin (1993) present explicit functional forms for α, all motivated by plausible intuitive arguments and appeals to selected experimental evidence, but no one has described observable behavioral assumptions on α that best describe behavior. Without a model of this parameter, Segal and Sobel (2004b) demonstrates that the theory that makes few definitive predictions about behavior.

Another problem is identification. It should be possible to separate preferences represented by v_i and u_i in (6) through revealed preference analysis. Observation of decisions made in a non-strategic setting determine u_i. Observations of decisions made in a strategic setting determine u_i. Adding incomplete information about preferences complicates this exercise, however. On the other hand, many different combinations of material payoffs $v_i(\cdot)$ and weight α will lead to the same behavior in strategic settings. Looking only at choice behavior in games it will not be possible to separate preferences for reciprocity from preferences over outcomes.

The third problem deals with the interpretation of mixed strategies. The behavioral interpretation of deliberate randomization is somewhat strained because it requires a player to select a precise weight on each pure strategy in spite of being indifferent over at least two pure strategies. Hence it is often attractive to interpret mixed-strategy equilibria as equilibria in beliefs. The important idea is that in some situations an agent must be uncertain about the pure-strategy choice of his opponent in equilibrium, but that uncertainty may be the result of incomplete information regarding the opponent's characteristics rather than due to conscious randomization on the part of the opponent. In standard game theory, the characterization of equilibrium does not depend on the interpretation of mixed strategies. In games where a player's preferences depend on the intentions of his opponent, the interpretation matters. A simple example, taken from Segal and Sobel (2004b) makes the point clearly.

Example 2 Consider the game:

	AM	PM
AM	$10, 10$	$0, 0$
PM	$0, 0$	$10, 10$
ALL	$7, 10$	$7, 10$

Column is a plumber and Row is a homeowner with a leaky faucet. The plumber can come in the AM or in the PM while the homeowner can arrange to be at home in the AM, in the PM, or ALL day. The plumber earns 10 if she coordinates with the homeowner, but nothing otherwise. The homeowner receives a payoff of 10 if he can meet the plumber and only cancel half of his appointments; he receives 7 if he stays home all day; he receives 0 is he fails to coordinate with the plumber. If players' preferences over strategies agreed with their preferences over outcomes, then the game has three equilibrium outcomes: (AM,AM), (PM,PM), and a continuum of equilibria in which the homeowner stays home all day and the plumber places probability of at least .3 on each pure strategy.

Now assume that the homeowner has preferences over strategies that lead him to put a positive weight on the plumber's payoff in response to nice behavior (apparent coordination) and a negative weight in response to nasty behavior. (Assume that the plumber cares only about her own payoffs.) Clearly, the two pure-strategy equilibria from the game with standard preferences will continue to be equilibria. But if randomization by the plumber is purposeful, then the homeowner may well think that a plumber who randomizes equally between AM and PM is nasty, because this behavior minimizes the probability of coordination. With a sufficiently negative weight on the plumber payoffs, the homeowner may prefer to play either AM or PM rather than to stay at home all day. Consequently, there may be an equilibrium in which both the homeowner and the plumber randomize equally between AM and PM, while (ALL, $\frac{1}{2}$AM + $\frac{1}{2}$PM) is no longer an equilibrium.

The above analysis depends on the interpretation of mixed strategies. In many applications, it is appropriate to treat the homeowner as if he is matched against a population of plumbers, some with a tendency to come in the morning, others with a tendency to come in the afternoon. If the homeowner does not attribute his uncertainty to a deliberate strategy of the plumber, then it is reasonable to assume that he does not place a negative weight on the plumber's payoff. In this case, however, there will be an equilibrium in beliefs in which the homeowner always stays at home (and he believes with probability greater than .3 that the plumber will come at any time). □

Many of the motivations for the importance of context and intentions are intrinsically dynamic, but dynamic considerations do not play a role in the basic formulation. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (forthcoming) study extensive-form games. They argue that the strategic-form approach of Rabin does not lead to good predictions in extensive-form games like sequential prisoner's dilemma games. Since it is possible to represent any extensive-form game in strategic form, the model applies to extensive-form games for standard reasons. Even for standard game theory, there is dispute about whether this transformation loses information. This issue is more complicated when intentions matter.

In standard examples, the notion of reciprocal preferences tends to increase the number of equilibria. Multiple equilibrium problems are possible, but less pervasive for the parametric models of extended preferences used in the literature. Coordination problems arise when preferences are interdependent. Players can get stuck in nasty equilibria in which they expect nasty behavior from their opponents and get it or in nice equilibria in which positive expectations are fulfilled.[21] While the ex-

[21]Segal and Sobel (2004b) formalize the intuition that allowing players' preferences to depend on

istence of multiple equilibria may be useful in applications, the possibility broadens the set of possible predictions from the theory so that equilibrium analysis places few restrictions on observable behavior.

3.5 Commitment

Allowing commitment changes the equilibrium concept rather than the specification of preferences. A player with commitment power selects his strategy to maximize a function of the form:

$$u_i(s_i, BR_{-i}(s_i)), \tag{7}$$

where $BR_{-i}(s_i)$ is the strategy profile of best responses for all other agents, taking i's strategy as given.[22] Commitment models are consistent with more general specifications of preferences. In the ultimatum game, if a selfish responder had full commitment power, then she would announce that she would only accept $s = 1$ and the proposer could do no better than to offer her the entire surplus. Experimental results in the ultimatum game are broadly consistent with the hypothesis that the responder has partial commitment power, but merely assuming commitment ability is not sufficient to handle a broader range of empirical regularities.

The solution s_i^* of problem (7) typically will not be a best response to the strategy choices of the other players. That is, commitment behavior is not consistent with Nash Equilibrium. I find it more useful to assume that all feasible commitment abilities are described in the specification of the strategic environment.

3.6 Repeated Games

The theory of repeated games provides a conventional framework in which to explain apparently unselfish behavior. In an infinitely repeated game, agents play a given static game, observe its outcome, and then play the game again and again, without end. Payoffs for the repeated game are discounted sums of the individual static-game payoffs. Strategies are rules that specify how to play in each repetition as a function of past behavior. Repeated games typically have large sets of equilibria. The folk theorem of repeated games states, roughly, that any feasible, individually rational payoff for the stage game can be obtained as a subgame-perfect equilibrium payoff for the associated infinitely repeated game.

---

context generally enlarges the set of equilibria.

[22] This formulation leaves open how the other $I - 1$ players arrive at best responses and what to do when the best response correspondence is not single valued.

To compare repeated games to the earlier approaches, consider a reduced-form description in which player i selects a stage-game action s_i to maximize

$$(1-\delta)u_i(s_i, s_{-i}) + \delta V_i(\theta(s)), \tag{8}$$

where $u_i(s)$ is player i's stage-game payoff function; $\delta \in (0,1)$ is a discount factor; θ describes the history of play; and $V_i(\theta)$ gives the continuation (average) value given history θ. In this formulation the continuation value function is endogenous. Thanks to the folk theorem, however, there are few restrictions on $V_i(\cdot)$. So although it is not possible to state that models of repeated interaction are formally identical to those based on general consumption goods, it is clear that if $V_i(\cdot)$ can take an arbitrary individually rational and feasible payoffs and δ is close to one, there is no reason that an equilibrium action choice for player i will be a myopic best response to the actions of his opponents.[23]

In a repeated interaction, an agent exclusively interested in his material consumption could rationally forgo short-term utility in order to obtain future benefits. If an apparently unselfish action is part of a repeated interaction between patient agents, then the folk theorem of repeated games can explain the observation as a part of equilibrium behavior between selfish agents. The logic of the folk theorem is the logic of instrumental reciprocity.[24] Individuals forgo their short-term selfish gains because being nice (or, more precisely and more generally, playing their equilibrium strategy) will lead to nice treatment in the future. Punishing nasty behavior serves to discourage nasty behavior, but punishment only occurs because players fear that a failure to punish will lower their future payoffs. This argument requires that the actions an agent takes today influences his future payoffs and that the influence is sufficiently great to counter short-term incentives.

Embedding an interaction into a repeated-game setting is a powerful and accepted way to describe behavior that, when viewed with a static perspective, appears to be inconsistent with selfish behavior. Like the approach of generalizing consumption, repeated game theory forces the observer to ask: "What does that player stand to gain?" in a way that often leads to useful insights.

The approach has limits. I discuss some of them now.

Because laboratory experiments carefully control for repeated-game effects, these results need a different explanation.

[23] That is, if equilibrium strategies specify the actions s^* in a given period, then there is no reason to expect that s_i^* solve: $\max_{s_i} u_i(s_i, s_{-i}^*)$.

[24] The anthropologists (for example Sahlins (1968) and Service (1966)) studying exchange clearly recognize this motivation.

In order for conventional repeated-game arguments to apply, the future must be important. Agents must be patient and there must be opportunities to reward and punish today's behavior. When these conditions fail, theory predicts a return to myopic selfish behavior. Relationships do end, and while it is easy to find evidence of myopic self interest at the end of relationships, it is also easy to cite examples of employees who do not shirk as retirement approaches and families that stand by dying relatives. Fitting this behavior into an individualistic repeated-game framework is possible, but awkward.

Repetition increases the range of equilibrium behavior because it creates the possibility of punishment. Punishment may be costly for the punisher as well as the punished. If so, the question arises: Why should anyone punish? The theoretical answer is: People punish because otherwise they will be punished themselves. Of course, the argument must be repeated to ensure that people are willing to punish the punishers of the punishers and so on. Theory provides elegant justification for this answer.[25] The theory is less direct and, perhaps, less convincing than the answer supplied by models of intrinsic reciprocity: Punishment arises because people get utility directly from lowering the welfare of people who have hurt them.

Repeated-game theory incorporates strategic context, not by changing preferences but by changing the way people play. In order to obtain equilibria distinct from repetitions of equilibria of the underlying static game, the history of play must influence future play. History does not influence preferences, but it does influence expectations about behavior. The principle of subgame consistency would require play in a subgame to be independent of where the subgame arises in a larger game. One must abandon subgame consistency to in order to predict repeated-game behavior distinct from repetitions of static equilibria.

Conditioning on history is so descriptive that there is little resistance to abandoning subgame consistency. There is some support for adopting a weaker principal. Some histories may trigger punishments that are bad for all players. For example, the "grim trigger strategy" specifies that players never again cooperate following a non-cooperative action in the prisoner's dilemma. The principal of renegotiation proofness identifies a set of possible (renegotiation-proof) equilibrium payoffs with the property that no payoffs in the set are Pareto-dominated by other payoffs in the set. The logic behind this definition is that if players are able to "renegotiate" at the beginning of each period on the equilibrium that they will play, then they will never agree to play an equilibrium that is Pareto inferior to another equilibrium. Renegotiation proofness says, in effect, that history can influence future play, but no history can induce players to select inefficient continuation. The idea is contro-

[25]Classic treatments are Abreu (1988) and Fudenberg and Maskin (1986).

versial (the weaker position that players will not play an equilibrium whose payoffs are Pareto dominated by another equilibrium's payoffs is already controversial for one-shot games). It is not straightforward to define renegotiation proofness in infinitely repeated games[26] and the restrictions imposed by renegotiation proofness do not lead to more descriptive predictions.

Finally, repeated game theory provides predictions consistent with many observations and also gives strong intuitions about qualitative features that increase the possibility of cooperation. But repeated games have too many equilibria and the selection process is often tailored to particular examples. The theory as it is typically used does not produce interesting refutable hypotheses.

4 Using the Models

This section describes several economic settings in which narrow notions of self-interested behavior provide limited insight. These examples provide further evidence in support of developing models that assume extended preferences and illustrates the way in which the different approaches of Section 3 provide alternative predictions.

4.1 Charity

It is difficult to rationalize charitable contributions as optimizing behavior from an individual who cares only about material wealth. Nevertheless, people do make charitable contributions.

Contributions appear to be sensitive to the economic environments in ways that are consistent with conventional theory.[27] For example, changes in tax laws that reduce the marginal cost of giving increase the amount of giving. Social psychologists have discovered factors that influence contributions that are less obvious consequences of standard economic assumptions. For example, Cialdini and Trost's (1998) review essay suggests that contributions increase in response to small gifts from the charity. There is also evidence that contributions are an increasing function of the contributions of others. These predictions are less obvious consequences of standard economic assumptions.

The models of Section 3 suggest different reasons why people give to charity. These models make different predictions about what influences charitable giving and

[26] Abreu, Pearce, and Stachetti (1993) and Farrell and Maskin (1989) are two approaches.

[27] Andreoni (2001) is a survey.

in principle can be distinguished. This section discusses some theoretical models of charitable giving and relevant experimental evidence.

Altruism, modeled as one agent placing value on the material welfare of others, predicts positive contributions that decline with the contributions of third parties. This is the crowding-out effect. The crowding-out effect should be strengthened if agents have distributional preferences: If agents are motivated to give because they wish to raise the income of the poor, then the more that the poor receive from others, the lower the value of direct transfers. The distinguishing feature of this kind of explanation is that people care about the distribution of income, but they do not care about how the nature of transfers that determine income distribution.[28]

Andreoni's (1990) warm-glow theory of charitable giving fits within the tradition of preferences over general consumption goods. Andreoni assumes that agents obtain utility by contributing to others. Agent i's utility depends on agent i's material consumption and how much he contributes to other agents. Preferences depend on not just the distribution of material goods, but on how one arrives at the distribution of material goods. Andreoni's model is qualitatively different from modeling altruism by simply assuming preferences place positive weight on other people's consumption as in the models of interdependent preferences. In distributional models, i cares about j's income, but not the source of the income. If j's income increases, then (under the standard assumption of diminishing marginal utility) i would be less willing to give money to j. In Andreoni's model, i derives utility from contributing. In the purest formulation, i's contribution would be independent of j's income, in contrast to the prediction of models based on altruism.

Sugden (1984) provides a model of charitable giving based on intrinsic constructive reciprocity. Sugden assumes that agents feel an obligation to contribute. With this (non-standard) assumption, Sugden is able to analyze his model using conventional methods. The most interesting result is that, because obligations are assumed to be increasing functions of the contributions of others, there will be a positive relationship between one's contribution and the (expected) contribution of others.

Hence one obtains a different prediction about the relationship between the contributions of others and one's own contributions depending upon the underlying theory. Altruism and more general distributional approaches predict that an individual contributes less when others contribute more. Warm-glow models predict that an individual's contribution does not depend on the contribution of others. Sugden's model of giving based on reciprocity predicts that an individual's contribution

[28]If j gains income from third parties, the marginal value of a contribution by i to j should not increase. The literature assumes that contributions will actually decrease; this prediction would not hold for linear preferences.

increases with the contributions of others. Empirical evidence from Cialdini and Trost (1998) and Croson's (1999) experiments therefore provide evidence against the simplest models of altruism or the warm-glow hypothesis.[29]

4.2 Incentives and Effort

The hypotheses of greed and equilibrium form the basis of powerful theories of contracting, which in turn make predictions about how firms should design internal compensation schemes and trade with suppliers. These predictions do not hold in simple experiments designed to reflect natural setting and do not appear to hold in many natural settings. Consequently there is a role for the kinds of models described in Section 3.

4.2.1 The Hold-Up Problem

The hold-up problem[30] has become an important toy model of bargaining between workers and firms. It has been used as the foundation for theories of the internal organization of the firm. Yet there is substantial reason to be skeptical of the predictions of the model.

In the simplest hold-up model, the first agent makes a costly investment. The investment increases the value of an object to the second agent (but not the object's value to the first agent). The second agent makes a take-it-or-leave it offer to the first agent. Assuming that agents maximize their monetary payoff, the subgame-perfect equilibrium predicts that there will be no investment. The second agent will purchase the item at the first agent's reservation price. Typically, positive investment is efficient.[31] The literature focuses on institutional arrangements that reduce or eliminate the inefficiency.

The standard analysis points out that if agents can commit to a complete contract, the first agent can be given proper incentives to invest. The literature on incomplete contracts argues that complete contracts are infeasible if agents have asymmetric information or unenforceable if agents are free to re-negotiate and third

[29]Andreoni (1995) and Palfrey and Prisbrey (1997) study experiments designed to distinguish giving due to altruism, warm-glow effects, and mistakes. Both papers find evidence of warm-glow giving. Palfrey and Prisbrey's design convincingly demonstrates that apparently altruistic behavior is due to mistakes.

[30]See Hart (1995) for an overview of the problem and its implication.

[31]These predictions are sensitive to modeling assumptions. Ellingsen and Robles (2002) point out that adaptive dynamics do not avoid efficient outcomes. Gül (2001) suggests that the standard results are sensitive to the information structure and assumptions about the bargaining process.

parties lack information needed to enforce the contracts. The literature then argues that ownership patterns and internal organization of firms arise to solve or lessen hold-up problems. Assuming that the predictions of the hold-up model with selfish agents are valid, the standard approach provides a coherent framework in which to study the nature of firm boundaries when complete contracting is not feasible.

There have been many attempts to study properties of simple hold-up games in experimental environments. The game underlying the hold-up problem has the same structure as the gift-exchange game introduced in articles by Fehr, Kirchsteiger, and Riedl (1993) and Fehr, Kirchler, Weichbold, and Gächter (Fehr, Kirchler, Weichbold, and Gächter 1998). In the first stage, the firm offers the wage w to the worker. In the second stage, the worker can either reject the offer (leading to zero monetary payoff to each player) or accept the offer. If the worker accepts the offer, he must then choose a level of effort, e. There is a monetary cost associated with effort, but it increases the profits available to the firm. For example, monetary payoffs conditional on an accepted offer could be $(v-w)e$ for the firm and $w-c(e)$ for the worker, where v is a redemption value set by the experimenter and $c(\cdot)$ is the worker's cost of effort. If the worker maximizes his material payoff, then he would choose $e = 0$, since the firm cannot condition wages on output.[32] Experimental subjects violate standard theory. In the laboratory, wage offers are typically positive and effort supplied is an increasing function of the wage.

Charness and Haruvy (2002) attempt to identify the cause of the increasing relationship between wages and effort. They compare the worker's behavior when wages are set by the firm, as in the original gift-exchange model, or by an external process. In the treatments where wages are determined through an external process, they were drawn from a bingo cage or set by the experimenter. In Charness and Haruvy's (2002) study, effort increases with wages when the firm sets wages, but is relatively flat otherwise. This provides evidence that models based on interdependent preferences are not sufficient to describe the worker-firm relationship. Preferences depend on the process by which workers receive their wages – not on the wages themselves.

These experiments suggest a reconsideration of simple contracting environments. Fehr and Schmidt (2000) present an experimental study of a simple contracting environment. The agent's effort net of wages determines the principal's monetary payoff. The agent earns her wages net of effort costs. Fehr and Schmidt permit the principal to offer two different kinds of contract. An explicit contract specifies a wage, a target effort level, and a fine. The agent receives her wage and if her effort is less

[32]When the monetary payoff to the firm is $(v-w)e$, the firm's choice of wage is not determined in equilibrium.

than the target, then she pays the fine with positive probability. An implicit contract specifies a wage, a target effort, and a bonus. The agent receives her wage. After observing the agent's effort choice, the principal decides whether to pay the bonus. Fehr and Schmidt assume that the principal must pay an additional (small) cost to propose an explicit contract. Assuming that players maximize their monetary payoffs, in non-trivial specifications subgame-perfect equilibrium predicts that the principal will offer an explicit contract. If the principal offers an implicit contract, the agent will expect to receive no bonus independent of her effort choice, and will therefore not agree to work. An explicit contract typically can induce positive effort levels. Fehr and Schmidt's principals opted for an explicit contract in fewer than 15% of the trials, with the frequency declining over time. Principals made more money when they offered implicit contracts, which induced positive effort levels and positive bonuses. Fehr and Gächter (2002) compare the performance of explicit to implicit contracts in a similar environment, where the choice of contract is a treatment rather than a choice variable of the principal. They find that the implicit contracts perform better than explicit contracts in an especially strong sense: agents cooperate less at a given level of compensation if they face a contract that fines them for shirking.[33]

The results of the contracting experiments are consistent with observational studies. Homan ((1953) and (1954)) observed a group of workers who systematically exceeded minimum work standards without apparent economic incentive for doing so. Bewley (1999) conducted extensive surveys of managers. His informants (managers) emphasize the importance of maintaining morale. Reducing wages is costly to firms because it leads to lowered worker morale, which in turn reduces workers' productivity. Bewley argues that high-powered incentives (wages and bonuses) are not an effective means of motivating workers. Bewley (Bewley 1999, page 407) finds little support for the standard paradigm in his interviews with managers. He describes the hold-up model as "fanciful" and states bluntly that "the hold-up problem hardly exists; ... it is not a consideration in labor relations." Part of the explanation for the failure of the hold-up model is that productivity depends upon the attitudes that workers have towards each other and to the firm, and that incentive schemes are selected with this in mind.

Employers opt for contracts that create the risk of hold up, but that workers do not take full advantage of this opportunity. Inefficiency arises, but not to the extent that theory predicts. Furthermore, it does not pay to make contracts as complete as possible. At the margin, there appears to be an important trade off between contractual incompleteness and apparently unselfish behavior.

Theories of internal organization of firms could benefit from an emphasis on inten-

[33]Cabrales and Charness (2000) obtain similar results.

tions rather than on contracts and information. Akerlof (1982) is a good example of the potential of this approach. Akerlof uses Homan's ((1953) and (1954)) studies of clerical workers in the early 1950's to motivate a model of labor markets that shares many features with the identity model that he later developed with Kranton (2000). In the model, the firm offers a fair wage; the wage generates good feeling among the workers, who in turn work more than the minimum required by the firm. Workers' utility may be increasing in effort in this model due to a desire to conform to a norm of behavior. The firm's decision to pay a fair wage is a profit maximizing response to the strategic environment.[34]

4.2.2 Intrinsic versus Extrinsic Rewards

Economic theory emphasizes the importance of incentives. Providing rewards contingent upon effort or positive performance directly encourages positive effort. While there is substantial evidence consistent with this point of view, there is also an argument that explicit rewards may have the adverse consequence of crowding out the agent's intrinsic motivation for performing the task.[35]

The social psychology literature contains experiments that demonstrate the possibility that people provided with extrinsic rewards will devote less effort to a task than people not provided with explicit incentives. A representative study (Lepper, Greene, and Nisbett (1973)) measured how the willingness of children to participate in a drawing activity depended on whether the activity had been rewarded in the past. Children given extrinsic rewards for the drawing activity were less likely than other children to participate in the activity after the reward had been withdrawn. Social psychologists interpret this result as a sign that the existence of an explicit reward inhibits the subject's ability to get intrinsic enjoyment from the activity. The experimental evidence on contracting is preliminary evidence that similar results arise in economic environments.

Two approaches from economics provide a framework for intrinsic rewards based on informational asymmetries. The work of Holmström and Milgrom (1991) on multi-dimensional agency problems suggests that high-powered incentive schemes may encourage agents to devote effort into activities that lead to immediate or observable outputs. In this model, agents respond to extrinsic rewards by allocating effort inefficiently.

Benabou and Tirole (2003) assume that the principal has information relevant to the agent. Any incentive contract offered by the principal has the potential to

[34]Solow (1979) also argues that sociological factors contribute to downward wage rigidity.

[35]Deci and Ryan (1985) surveys the literature in social psychology.

convey this information to the agent. For example, if the principal has superior information about the difficulty of the job, then a contract that promises a high-reward contingent on success of a project might convey the message to the agent that the job is distasteful. Benabou and Tirole present models in which the incentive offered by the principal may signal to the agent that the task is an onerous one. In their basic model, increasing compensation increases the probability that an agent will supply effort, but also signals to the agent that the job is distasteful or that effort is unlikely to lead to success.

Gneezy and Rustichini ((2000b) and (2000a)) present evidence for explicit incentives having counter-intuitive influences in laboratory and natural experiments. Gneezy and Rustichini (2000a) shows that the imposition of an explicit penalty for failing to pick up a child on time at a day-care center leads to a decrease in the number of people who pick up their child on time. Gneezy and Rustichini (2000b) describes experiments that show the level of performance in a task is not monotonic in monetary rewards. Gneezy and Rustichini suggest that a mechanism similar to the one described by Benabou and Tirole is at work. The incentive scheme conveys information to the agent, which leads to a counter-intuitive response. In the day-care setting of Gneezy and Rustichini (2000a), for example, imposing a modest fine for late pick ups could lower the agent's subjective probability that an even more severe penalty would be imposed.

These models of negative aspects of extrinsic rewards focus on information asymmetries. In Holmström and Milgrom (1991), rewarding observable performance may induce suboptimal allocation of effort. In Benabou and Tirole (2003), explicit incentives provide information that may lead to a rational agent to update his opinion about the attractiveness of a task.

The psychology literature suggests that crowding out does not depend on the ability of incentive schemes to convey information about the task, but instead argues that incentive schemes change preferences in systematic ways. Bewley (1999) and Kreps (1997) support this point of view. Specifically, Bewley discusses three ways by which managers motivate workers. Direct exchange is standard monetary (extrinsic) incentives. Indefinite exchange describes implicit, but material, payoffs. Internalization is a change in preferences brought about by the work environment. If firms treat workers nicely, then the workers start to care about the objectives of the firm. Employees become part of the team and work to advance the firm's interests. Bewley suggests that firms do not lower wages because doing so would destroy the basis for voluntary cooperation. Intrinsic reciprocity focuses attention on the role that incentive schemes play in preference formation. Models based on intrinsic reciprocity permit a firm's behavior to induce its workers to obtain satisfaction from

directly contributing to the firm's profitability.

4.3 Markets and Selfishness

Do markets cause selfish behavior? Some have argued that they do. Bowles (Bowles 1998, page 89) observes that the more the "situation approximates a competitive (and complete contracts) market with many anonymous buyers and sellers, the less other-regarding behavior will be observed." In his comparative study of the development of markets in Indonesian villages, Geertz (Geertz 1963, page 34) writes that "the general reputation of the bazaar-type trader for 'unscrupulousness,' 'lack of ethics,' etc., arises mainly from" asymmetry of roles in retail markets. Balance is more difficult to maintain in asymmetric transactions, so cooperation based on reciprocity is more difficult to sustain. The reciprocity that facilitates cooperation in symmetric transactions does not thrive in market settings.[36]

Results of experiments designed to model competitive situations are consistent with predictions of self interest. Prasnikar and Roth (1992) studied a variation of the ultimatum game with many proposers. The proposers simultaneously make an offer. If the responder accepts the offer p_i from proposer i, then proposer i earns $\$10 - p_i$, the responder earns p, and the other proposers earn 0. If the responder rejects all offers, then all players earn nothing. Hence the design adds competition between proposers. Subgame-perfect equilibrium predicts that the responder will receive (nearly) all of the surplus. Experiments confirm this prediction.

Andreoni, Brown, and Vesterlund (2002) identify a related experimental environment in which standard predictions are confirmed and compare the result to similar games in which experimental results are not consistent with narrow versions of self interest. They consider a two-player game in which the monetary payoff of player i is $T - c_i + \gamma_i G(f(c_1, c_2))$, where T is an initial endowment; c_i is player i's contribution; $\gamma_i > 0$ is a measure of a player's marginal utility of the consumption good; $f(\cdot)$ aggregates the contributions; and $G(\cdot)$ is an increasing, concave transformation of the total contribution. Andreoni, Brown, and Vesterlund (2002) assume that $\gamma_1 > \gamma_2$ and distinguish three games. In the first, $f(c_1, c_2) = c_1 + c_2$ and players move simultaneously. The second treatment uses the same $f(\cdot)$ but is a sequential game: player

[36] On the other hand, the experiments that Henrich and his associates (2001) performed in different societies led these authors to conjecture that cooperation in simple experimental games is positively related to the degree of market integration. They argue that experimental subjects apply strategies that succeed outside the laboratory to novel experimental games. Subjects more exposed to situations where there are substantial gains from cooperation or where application of fairness norms are important for success will be more likely to exhibit cooperative behavior in the laboratory.

one moves first and then player two, knowing the first player's contribution, moves second. The third treatment is like the second except that $f(c_1, c_2) = \max(c_1, c_2)$.[37] In the first treatment, the Nash equilibrium predicts that the second player will contribute nothing, while in the second and third treatments, the first player contributes nothing in a subgame-perfect Nash equilibrium.[38] Andreoni, Brown, and Vesterlund (2002) find that players make positive and roughly equal contributions in treatment one, contributions are positive for player one and slightly greater for player two in treatment two, while in the third treatment experimental behavior roughly conforms to the subgame-perfect Nash equilibrium theory (with the first player making no contributions). In the third treatment, the second player is much less likely to punish a free rider (by responding to a zero contribution with a zero contribution) than in the game in which contributions are additive.

From the perspective of intrinsic reciprocity, the third treatment is qualitatively different from the first two. Since the second player receives a higher marginal utility from the public good, she can see that in the third treatment any "sensible" contribution made by the first player will be wasted – the second player will want to contribute even more, making the value of the first contribution zero. Hence the magnitude of player one's contribution should not influence player two's preferences over final outcomes. It is reasonable to expect player two's material preferences to determine her behavior. In the first and second treatment, player one's strategy can influence player two's material payoff.[39] It is not hard to specify preferences in the form (6) that are consistent with experimental results. Agents act as if they were selfish in the third treatment because the form of the game prevents player from exhibiting the kind of nice or nasty behavior that might trigger deviations from selfishness.

The evidence that markets behave as standard theory predicts in experimental settings motivates research to identify weaker assumptions under which the predictions of these models continue to hold. Becker's (1962) observation that budget-constrained individuals with randomly generated demands lead to downward sloping demand curves and Gode and Sunder's (1993) work on auction performance with "zero-intelligence" agents are examples.

Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) show how the "com-

[37]This treatment is a variation on the best-shot game of Harrison and Hirshleifer (1989).

[38]The predictions for the second and third treatments require that γ_1 is not too much greater than γ_2.

[39]Since the second player responds differently when player one fails to make a contribution in the second and third treatment, theories that rely on preferences that depend only on the distribution of monetary payoffs will not make predictions consistent with the experiments.

petitive" prediction of the game studied by Prasnikar and Roth (1992) continues to hold under the assumption that some individuals in the population have utility functions that depend on the distribution of monetary payoffs. These results follow because in the market, it only takes two selfish bidders to drive the price up to the competitive level. Assuming that some of the proposers or the responder cares about the distribution of payoffs does not change the equilibrium outcome.

The broad observations of Bowles and Geertz and the specific experimental findings are also consistent with models of intrinsic reciprocity. In these models, unselfish behavior arises in response to the behavior of others. An agent may be generous to another in order to avoid a spiteful response or in order to provoke an altruistic response. The power of spiteful behavior is weak in market environments. If a consumer refuses to buy an object because its price is "too high," a firm will just sell to another customer (and, in equilibrium, there will be another customer willing to pay the price). In some situations, the cost of generosity is high. In a competitive equilibrium, a firm that drops its price risks the possibility of facing large excess demand. Meeting the demand might drive average cost above price. Hence, in certain market environments it may be impossible to distinguish between the decisions of agents who are maximizing their narrow self interest from unselfish agents who make the same decisions. Added to the anonymous nature of market interactions (that make it difficult to identify the kind and the unkind), the outline of a model of how markets might influence behavior begins to appear. Segal and Sobel (2004a) provides such a model. This paper makes two contributions. First, it shows that it in market environments more general than the ones studied in experiments, equilibrium outcomes are competitive even when agents are willing to punish nasty behavior and reward nice behavior. The result follows because in a market, agents can only influence the payoffs of others by changing the equilibrium price and there are limited opportunities to do this. When there is no opportunity to help or hurt others, narrow self interest determines behavior. Second, it demonstrates for the kind of separable functional forms used in the study of interdependent preferences and reciprocity ((1) and (6)), price-taking behavior cannot be distinguished from selfish behavior. Hence agents will appear selfish whenever the price-taking assumption makes sense (for example, in large economies).

The existence of markets may not change preferences, but it may remove incentives for reciprocal behavior. Indeed, the theory suggests the possibility of substitution between well organized economic institutions and the observance of other-regarding behavior.[40] As Dasgupta (2000) notes, there is nothing mysterious in many

[40]The interaction can go both ways. Yamagishi (1988) and Yamagishi, Cook, and Watabe (1998) demonstrate how well developed sanctions procedures enhance trust in Japanese society. Fehr,

acts of reciprocity and "there is no reason to invoke the idea that there is greater innate generosity and fellow-feeling among poor people in poor communities than exists among members of modern urban societies."

Assuming that the predictions of standard models hold in more general settings, it would be worthwhile investigating the extent to which the properties of these outcomes continue to hold. One can show the existence of competitive equilibrium in an environment where agents have exotic preferences, for example, but it is a separate question whether the unselfish behavior leads to good outcomes. The fundamental theorems of welfare economics provide conditions under which selfish behavior leads to efficiency. There is no general result suggesting that outcomes in games played by agents with the kinds of extended preferences described in this paper will be efficient – either with respect to underlying material preferences or with respect to preferences that take into account attitudes towards the behavior and intentions of others.[41]

4.4 Repeated Interaction

Game theory relies on two methods to explain cooperative behavior in repeated interactions. The folk theorem of repeated games provides a rich theory that is consistent with all manner of behavior in repeated interactions. If future gains are large enough relative to the gains from cheating today, then (assuming a mild technical assumption holds) any individually rational, feasible payoff can be supported as an equilibrium payoff. The literature on reputations works differently. Players initially are uncertain about the motives of their opponent. Cooperation arises if players can infer from past behavior that their opponent is likely to be trustworthy.

In the folk theorem, players must be forward looking. In each period, there is typically short-term benefit from cheating. Players refrain from cheating in order to gain future benefits. In reputation stories, players use experience to determine whether they believe their opponent is reliable.

In the purest form, these stories differ in their predictions about breaches of cooperation. In the simplest repeated-game stories, there are no deviations. (When

Gächter, and Kirchsteiger (1997) and McCabe, Rassenti, and Smith (1998) present additional experimental evidence that providing more opportunities to punish increases cooperation in situations where selfishness and subgame perfection predict that they would have no influence on behavior. It is possible that effectiveness of punishment options differs systematically across cultures.

[41] Kranton's (1996) theoretical work demonstrates the stability of both market based and reciprocal systems of exchange. In her model, market-based exchange is more efficient than bilateral reciprocal trading arrangements. Nevertheless, a system of reciprocal exchange may be self-sustaining. When more agents opt for reciprocal exchange, markets thin. It becomes optimal for agents to engage in personal exchange.

there is a deviation, a punishment ensues.) In the reputation stories, a reputation develops only because of learning. Learning takes place only because cheating arises with positive probability. One expects to see breakdowns in some relationships (when the interests of the players are not matched). Models of intrinsic reciprocity provide a third way to describe cooperation in repeated interactions. After a sequence of good outcomes, players' interests become more closely linked. A history of positive interaction with someone leads you to care about that person's welfare.

Allowing players' preferences to exhibit intrinsic reciprocity typically enlarges the equilibrium set in one-shot games. Therefore it is perhaps surprising that repeated games played by agents with intrinsic preference to reciprocate may have equilibrium sets that are smaller than when players have conventional preferences over outcomes. Take a simple example. Let s^* be a strategy profile that maximizes the sum of (material) payoffs in the stage game. Suppose that by playing s_1^* repeatedly, player 1 leads player 2 to play to maximize the sum of material payoffs. This could happen for the appropriate specification of preferences. It follows that if player 1 is sufficiently patient, he can guarantee an average payoff equal to the outcome of s^*. Hence, under certain conditions, average repeated game payoffs **must** be efficient (in the set of material payoffs) if individuals have intrinsic preferences towards reciprocity. On the other hand, if individuals have conventional preferences efficient payoffs are only guaranteed to be elements of a large set of equilibrium payoffs. The result requires strong assumptions and it suggests that agents will cooperate in the final periods of long, repeated interaction, a prediction that is not consistent with all available evidence. It does provide an alternative to conventional modeling of repeated interaction.

5 Origins

It is natural to ask where preferences come from. The question takes on a greater importance when one argues that incorporating extended preferences may lead to more useful models. A better understanding of the origins of preferences may add structure to modeling efforts by placing restrictions on the type of utility functions that people can have and the circumstances under which utility functions can change. This section provides an overview of attempts to model the evolution of interdependent preferences and reciprocity.[42]

[42]Sethi and Somanathan (2003) also reviews this literature. The January 2004 issue of *Journal of Economic Behavior and Organization* contains a review essay by Henrich (2004) on evolutionary models of prosocial preferences and a number of commentaries.

In economic contexts, utility maximizing behavior can be confused with selfish behavior. If one is intrinsically motivated by the desire to help others, then one can question whether making a material sacrifice to aid others is truly altruistic.[43] The evolutionary perspective might make it possible to distinguish clearly between selfish behavior and more general preferences. There is no dispute that at the lowest level of selection, fitness is the appropriate measure of success. Unselfish genes do not exist. The notion of self interest leads to confusion even in biological models when one thinks about evolution of individual organisms (instead of individual genes). A trait that appears to be altruistic at the level of an organism will arise as the result of conflicting "self interests" of different genes. Traits that fail to maximize the fitness of the organism are common. At this level of abstraction, theoretical explanations of prosocial behavior are easy to generate, but difficult to evaluate. In biology, reduced-form models sometimes abstract from gene-level conflicts and model evolution at the level of the individual. In economics, I am aware of no model that treats selection at a level lower than the individual. I therefore limit attention to selection for individual traits and describe conditions under which it is in the best long-term interest of an individual to maximize something other than his short-term material payoff. The interaction between individuals and groups will play an important role in this discussion.

5.1 Reciprocal Altruism

Trivers (1971) introduced the idea of reciprocal altruism. His theory parallels standard repeated games arguments that justify forgoing short-term gains for long-term benefits in repeated relationships. This type of behavior arises in non-primates.[44] There is a large literature on the evolutionary foundations of cooperative behavior in repeated games. This literature provides a foundation for the appearance of behavior that is inconsistent with short-term rationality. Reciprocal altruism is instrumental reciprocity. Actors may repay kindness with kindness, but only because they anticipate future benefits in return.

Arguments that generate cooperative behavior in evolutionary settings using reciprocal altruism depend on the same kind of assumptions that are necessary for the folk-theorem in repeated games. The future must be important; there must be an

[43]Sober and Wilson (1998) present a careful discussion of these issues.

[44]See Dugatkin (1997) for extensive examples. De Waal (1996) describes more elaborate forms of reciprocity that arise in higher primates. Kroptokin (1902) presents many examples of apparently non-selfish behavior in animals. Ridley (1996), however, argues that elaborate forms of reciprocity are uniquely human. Boyd and Richerson (1988) briefly survey additional evidence supporting the position that cooperation in large group settings is a characteristic of human behavior.

ability and incentives to punish opportunistic behavior; and it must be possible to identify deviators. From the game-theoretic perspective, a weakness of the repeated-game arguments is multiplicity of equilibria. Indeterminacy arises in the evolutionary literature because strategically stable strategies often fail to exist in repeated games.

5.2 Green Beards

Dawkins (1982) describes a general mechanism that creates the possibility of cooperative behavior. Imagine a population that will play a prisoner's dilemma game in pairs. A subset of the population has an observable feature (a "green beard") that is perfectly correlated with discriminatory cooperative behavior: people with green beards cooperate with other green beards, but with no one else. Individuals with green beards can thrive because they gain the benefits of cooperative pairings without running the risk of getting cheated. Frank (1988) exploits a version of this argument. He introduced a model in which people decide whether to play a prisoner's dilemma game with an opponent or opt out. The population contains one group of agent that always cooperates and another that always defects.[45] Group members give off different signals. Frank shows that if the signals are sufficiently informative, players can use them to determine whether they wish to play with their opponent. A significant degree of cooperation can be supported in equilibrium provided that the signal that cooperators emit is sufficiently informative. The green-beard argument provides a powerful reason, nicely modeled by Robson (1990), to believe that inefficient outcomes may fail to be evolutionarily stable.

A well-understood limitation of this approach is that an individual who could grow a green beard without cooperating would have an advantage. This individual would be able to reap the gains of generous behavior when he meets others with green beards without paying the cost.

The same mechanism forms the basis for a general approach to the problem of selection of preferences in a strategic setting. The basic framework begins with a game. The payoffs of the game are assumed to be the players' material payoffs (or, in a strict evolutionary framework, reproductive fitness). Players may alter their utility functions within a parametric class. For example, a player may replace his material payoff with a weighted average of his material payoff and the material payoff of his opponent. Altering payoffs creates a new game. Assuming equilibrium play in the new game generates a predicted equilibrium outcome, which in turn generates material payoffs for the players. The evolutionary dynamic operates on the material payoffs.

[45] Frank models this as a fixed characteristic of the agent, but alternatively one could assume that different groups have different preferences.

Players can choose to play games using different strategies than fitness-maximizing strategies. Further, the presence of non-fitness maximizers in the population may influence the behavior of others. Assuming that the fraction of agents in the population with particular preferences grows in proportion to relative fitness, the literature asks whether there are situations in which non-fitness maximizers remain a positive fraction of the population.

There are many games in which a player gains by acting as if he is not motivated by his material payoffs, provided that his opponents' know his preferences. Güth and Yaari (1992)[46] initiated the study of this process under the assumption that strategy choice is observable.[47]

These approaches study specific games; they are open to the criticism that the successful preferences are finely tailored to the strategic setting. Koçkesen, Ok, and Sethi (2000b) provide a version of these results that applies more broadly. They identify a more general class of games in which agents who choose a utility function that is increasing both in their monetary payoff and their relative monetary payoff (the ratio of monetary payoff to the average monetary payoff) receive higher monetary payoffs in equilibrium than players who maximize monetary payoffs. The critical assumption in Koçkesen, Ok, and Sethi (2000b) (and also in a related example of Possajennikov (2000)) is an assumption of supermodularity (or strategic complementarity). To get an intuition for the result, imagine a symmetric two-player game in which one player cares only about monetary payoffs and the second cares about relative payoffs as well. Koçkesen, Ok, and Sethi (2000b) observe that for these games there is no equilibrium in which the second player has a lower monetary payoff than the first player (because by mimicking the first player, the second player increases both absolute and relative payoff).[48] Using similar techniques, Sethi and Somanathan (2001) show that preferences similar to those introduced by Levine (1998) can survive in a general family of games.[49] As many public goods games exhibit strategic substitutes, these results may help organize some experimental results.

The approach does not provide specific guidance about the nature of the interdependent preferences that survive evolutionary pressures. The analysis identifies circumstances in which commitment ability is valuable. The papers give conditions under which certain types of non-selfish preferences can persist in a population dom-

[46]Güth (1995a), Güth and Yaari (1992), Guttman (2000), and Possajennikov (2000) all provide results with the same general flavor, with an emphasis on social preferences broadly consistent with those introduced to describe experimental outcomes.

[47]Schelling (1960) gives vivid, early illustrations of the approach in the context of bargaining.

[48]The result requires more assumptions in n-player games. Koçkesen, Ok, and Sethi (2000b) prove a related result for games with strategic substitutes.

[49]Koçkesen, Ok, and Sethi (2000a) contains a related result.

inated by selfish individuals. The analysis does not provide a systematic theory of the distribution of preferences that would survive without a priori restrictions on preferences. Due to the complexity of the analysis, the papers study the advantages of a particular, intuitive sort of non-selfish behavior, rather than identify the end result of an evolutionary process. This approach is the appropriate first step in a research program designed to provide a rationale for social preferences, but supports only the conclusion that evolutionary models do not demand all agents have fitness maximizing preferences. Future work may provide a more precise description of stable preferences.

A more basic problem is the central assumption that preferences be observable. In all of the models described, it is (at least weakly) to an agent's advantage to convince his opponent that he has non-selfish preferences while actually having fitness maximizing preferences. These agents gain the strategic advantages of commitment to non-selfish behavior (changing the behavior of their opponents), but do not pay the cost (making decisions that fail to maximize fitness). In this way, the models suffer from precisely the same criticism as the green-beard models. One would expect evolutionary pressures to favor the kind of duplicitous behavior found in standard models: Agents would arise who look like non-selfish agents, but whose true preferences are traditional. These strategies are not feasible in commitment models because of the assumption that there is complete information about preferences.

Ok and Vega-Redondo (2001) and Ely and Yilankaya (2001) examine the evolution of preferences when preferences are not directly observable.[50] Consequently, these papers provide a formal model that studies the critique of green-beard mechanism. The central idea is that modifying preferences only increases fitness to the extent that the preference change can modify the behavior of other agents. When preferences are not observable, there will be selective advantage to imitating the behavior of self-interested agents. The evolutionarily stable outcomes of the selection process must agree with the equilibrium outcomes of the underlying game played by selfish agents. These models therefore warn that the assumption that preferences are observable in commitment models is critical. It provides a framework confirming the intuition that "green-beard" arguments rely on the assumption. It naturally leads one to ask about the existence of mechanisms by which agents can credibly signal their true preferences.

As explanations for the existence of cooperative behavior all of the green-beard models raise the same question: What prevents defectors from learning how to fake signals? Even if the answer is "nothing," it is likely that selection will favor members of the population who are capable of creating difficult to imitate signals or who are

[50] Although in Ok and Vega-Redondo (2001), players can infer it under some matching conditions.

able to distinguish sincere from insincere signals. Green-beard arguments have power if there is a hard-to-break link between signal and behavior. While these links cannot be deduced from general principles, biologists argue that humans have developed the capacity to evaluate the intentions of others.[51] It seems likely people differ in their abilities to deceive others and to be deceived.[52] These characteristics vary with individual preferences and have implications for the kinds of tasks people are suited to perform.

The end results of these conflicting pressures is not clear. Alexander (1987) and Ridley (1993) argue that the pressures lead to increases in social, mental, and emotional complexity.[53] What remains is a complicated picture of the preferences of individuals, but one which includes the possibility of non-selfish behavior intrinsically included in preferences.

5.3 Kin and Group Selection

Arguments supporting altruistic behavior, in the sense that individuals reduce their own reproductive fitness to benefit others, can be based on inclusive fitness. Hamilton's (1964) notion of inclusive fitness provides an explanation for altruistic behavior in animals. Hamilton shows that an action that lowers the probability that an individual survives could increase that individual's total fitness (genetic contribution in future generations) if it increases the probability that relatives survive. Hamilton's ideas provide a way to understand a wide range of animal behavior. While these ideas provide strong support for prosocial behavior when individuals interact in small groups with closely related individuals, these conditions surely do not apply to laboratory experiments and are inadequate to explain the existence of prosocial behavior in common natural settings.

[51]For example, Trivers (1971, page 51) states that "there is ample evidence to support the notion that humans respond to altruistic acts according to their perception of the motives of the altruist. They tend to respond more altruistically when they perceive the other as acting 'genuinely' altruistic." De Waal (1996, page 116) writes that "a person who lies without blushing, who never shows remorse, and who grabs every opportunity to bypass the rules just does not strike us as the most appealing friend or colleague. The uniquely human capacity to turn red in the face suggests that at some point in time our ancestors began to gain more from advertising trustworthiness than from fostering opportunism." Ekman (2001) argues that there are characteristic facial expressions that provide credible signals of honest behavior.

[52]Abraham Lincoln [or possibly P. T. Barnum] made (most of) this point more gracefully: "You may fool all the people some of the time; you can even fool some of the people all the time, but you can't fool all of the people all the time."

[53]Alexander even conjectures that consciousness is the outcome of the need to evaluate and interpret the motivations of others.

There are coherent theoretical models that extend kin selection to groups of unrelated individuals. Sober and Wilson (1998) describe how group-selection models can lend support to non-selfish behavior (see also Bergstrom (2002) and Samuelson (1993)). These models are natural generalizations of arguments based on inclusive fitness. Since Hamilton's work, it has been apparent that non-selfish behavior can have selective advantage in closely related groups: One agent should be willing to sacrifice individual fitness if by doing so there is a large enough increase in the fitness of closely related individuals. In this way, the individual's genes (although not necessarily the individual) gain. More generally, one can imagine that non-selfish preferences in a subset of the group may make the average fitness of the group higher than that of an entirely selfish group. Within the group, selfish members do better than non-selfish members. This is the case for standard public-goods models. If there is only one group, these conditions lead to the extinction of the non-selfish members. Imagine instead that there is another population, consisting entirely of selfish individuals. The proportion of non-selfish agents in the entire population may increase from one generation to the next if the relative increase of the group containing non-selfish agents (at the expense of the other group) compensates for the relative decline of the non-selfish agents within their own group. If groups remain stable over time, this argument only postpones the extinction of the non-selfish agents as first the selfish group dies, and then the selfish agents take over the remaining group. If groups re-form in each period, then there are conditions under which non-selfish behavior can survive.

An attractive model using group selection arguments to explain the persistence of non-selfish preferences in an economic environment is due to Herold (2003). Herold examines a two-player game of perfect information in which the leader can decide whether to give a gift to the follower. Giving the gift maximizes total surplus, but is costly to the leader. The follower can do nothing, reward, or punish the first player. Herold looks at three situations: two in which only one of the unselfish preference types is available and one in which all three types of preferences may enter the population. Rewards and punishments lower the material payoff of the follower. Herold investigates whether players with non-selfish preferences can be evolutionarily stable in an environment where the game is played in (anonymous) small groups, fitness (material payoffs) determines reproductive success, and groups re-form after each generation. Players know the distribution of preferences in their group, but do not know the preferences of their partner. Herold observes, as in Ok and Vega-Redondo (2001) and Ely and Yilankaya (2001), the leaders behave as if they maximize expected fitness in any stable outcome. Their behavior could involve giving gifts, but only if doing so brings rewards or avoids punishment. There are

stable outcomes in which a fraction of the followers has interdependent preferences.

Consider the case where only rewards are possible. In a world in which every follower is selfish, none of the leaders are generous. Consequently, followers with preferences for rewarding generous behavior can enter the population without sacrificing fitness. Moreover, if a positive fraction of the population of followers give rewards to generous leaders, then there is a positive probability that a large enough fraction of them will be concentrated in a group to induce the leaders in this group to be generous. All members of the cooperative group obtain higher payoffs than the rest of the population and, because there must be a high concentration of followers who give rewards in this group, the fraction of the entire population who gives rewards can grow.

Next consider the case in which followers are either selfish or get utility from punishing. It is also possible for followers who like to punish to be represented in a stable population, but the argument differs in an important respect. It is costly for a follower who punishes greedy behavior to enter a population in which all followers are selfish (because in this population leaders will not be generous and therefore the punisher will be required to engage in costly punishment). Herold (2003) shows that in this setting a monomorphic equilibrium of selfish followers is stable, but there is also a monomorphic equilibrium with only followers who punish. Selfish followers cannot gain a foot hold in the punishment equilibrium because if there were enough selfish followers to create a group in which leaders are greedy, then the followers in this group – with a disproportionate share of selfish followers – would do badly relative to the population and therefore be less common in the next generation. Finally, when Herold permits all three types of preferences, a stable outcome in which all followers punish exists and, under some conditions, the equilibrium in which both some followers do nothing and some reward exists.[54]

A straightforward analysis of the Price (1970) equation provides an understanding of the mathematical conditions needed for a group selection argument to support the spread of prosocial behavior. Henrich (2004) and Sober and Wilson (1998) discuss these conditions carefully. A restrictive condition is that the variation within groups is significantly smaller than the variation between groups. If there is a lot of variation within a group, then individuals who are not fitness maximizing within the group are at a large disadvantage relative to other group members. If different groups are similar, then one group is unable to do significantly better than another. If there is free mixing between groups, then the variance between groups falls.

There is general agreement among biologists that group selection depends on a degree of genetic variation across groups that is not consistent with migration

[54]Gintis (2000) presents a related models.

patterns. Relatively small amounts of intermarriage is sufficient to destroy between group variation necessary for biological group selection arguments. Alexander (1987, page 37 and pages 168–70) is representative of the consensus.[55]

While human groups are genetically similar, they are culturally diverse. These differences make the transmission of prosocial behavior through cultural channels more probable than purely genetic transmission. Boehm (1993) and (1997) presents and defends a mechanism that supports the rise of altruistic behavior. Boehm argues that human forager societies have characteristic social organization that facilitates group selection for prosocial behavior. Boehm (1993) cites evidence that foragers and other small-scale societies create egalitarian cultures in which all households have comparable social and economic status. This structure comes about through the ability to sanction both the shirkers who attempt to share in the group's resources without contributing and bullies who attempt to monopolize the resources. A consequence of the egalitarian structure is a reduction of the behavioral variation within groups. Boehm then points out that there is variation across groups in the way that they respond to emergency conditions (for example, famine).

Boehm (1997) argues that these features of small-scale societies (in addition to the ability and willingness of groups to sanction deviations) strongly support the development of behaviors that favor group survival. Because variation within a group is low and deviations from group behavior are sanctioned, there is within group pressure to retain traits that are good for the group. Because different groups behave differently, successful groups grow. Hence the structure imposed by human culture facilitates the development of traits that are beneficial to group survival.[56]

In summary, genetic group selection arguments are not likely to be an important reason for the development of altruistic behaviors, but that the ability of groups to design cultural practices or institutions that reduce intra-group variation and maintain inter-group variation does provide a powerful, consistent, and empirically justifiable explanation of the development of prosocial behavior.

[55]Sober and Wilson (1998) argue that group selection is an important explanation for the evolution of prosocial behavior in humans. Smuts's (1999)'s insightful review of Sober and Wilson's book accepts the logical validity of group-selection arguments, but argues that attempts to identify altruism in organisms should not lose sight of the fact that selection leads to fitness maximization at the level of the gene.

[56]Boyd and Richerson (1985) argue that cultural group selection can lead to the development and retention of prosocial behavior. They describe the importance of imitation as a mechanism that reduces intra-group variation, while maintaining inter-group variation.

5.4 Evolutionary Evidence of Decision Biases

I have described evolutionary models based on strategic interaction. Evolutionary psychology suggests a variety of mechanisms consistent with natural selection to describe social behavior. Some of this work provides reasons for action rules that violate standard material-utility maximizing behavior in favor of actions consistent with social norms.[57] In this section I discuss one example that the routines are context specific. Cosmides and Tooby (1992) review experimental evidence on Wason's problem.[58] In its original form, the Wason (1966) selection task, subjects examine four cards. On the visible side of the cards, they see:

$$\text{(E)} \quad (4) \quad \text{(K)} \quad (7).$$

Subjects know that each card has a letter on one side and a number on the other. Subjects are asked to turn over precisely those cards that need to be turned over to determine the truth of the statement:

If a card has a vowel on one side, then it has an even number on the other side.

Subjects make systematic errors in analyzing conditional statements when given the problem in an abstract form.[59] The error rate goes down when the problem is reformulated so that the task asks whether someone has violated a social norm. In a representative reformulation, the visible faces of the four cards are:

$$\text{(Beer)} \quad (24) \quad \text{(Coke)} \quad (17).$$

Subjects know that each card has a beverage on one side and an age on the other. Subjects are asked to turn over precisely those cards that need to be turned over to determine whether there are any violations of a law forbidding people under 21 from drinking alcoholic beverages.

Cosmides and Tooby (1992) interpret these experiments as evidence that human cognitive processes have evolved to identify violations of conditional statements when these statements can be interpreted as cheating on a social contract. They go on to speculate that people have mental algorithms that lead them to punish cheaters.

[57] In a speculative essay, Varela (1999) suggests that effective people rely on "crazy wisdom." He suggests that ethical behavior requires a mixture of rational calculation and spontaneity and that cognitive limitations increase the likelihood of ethical behavior.

[58] Social and cognitive psychologists conducted the experiments. The experimental designs violate the accepted practices of experimental economics: the task was poorly specified; subjects had no financial incentive for providing accurate answers; and controls on the context were incomplete.

[59] Subjects should check the (E) and (7) cards to verify the statement. Most people examine either the (E) card only or the (E) card and the (4) card.

Cosmides and Tooby's experimental results are stimulating, but alternative explanations of the experimental findings[60] that have no relationship to reciprocity or interdependent preferences are available.[61]

5.5 Learning

Game-theoretic models of the evolution of preferences take a simple view of the evolutionary process. Reality is more complicated. This subsection speculates on several directions that may lead to alternative explanations for non-selfish behavior.

Animals (including humans) operate in a variety of different strategic environments. Cognitive constraints make it impossible for anyone to optimize in every natural environment. These limitations create at least three reasons why people would fail to exhibit self-interested preferences.

The first, and most obvious, observation is that individuals will not be fully selfish because they are unable to perform the needed calculations. There is no doubt that cognitive limitations prevent people from solving complex optimization problems, but in general the critique applies equally well for all objective functions. Cognitive limitations are a good argument for some models of bounded rationality, but without a theory of complexity, provide no systematic evidence that people optimize interdependent preferences or incorporate intentions of others in optimizing behavior.

Another idea is that people have a limited ability to distinguish one situation from another. Instead they use experience and easy-to-process signals to sort the problems that they face into a small number of categories. For each category they apply a preference relationship (or behavioral rule of thumb) that is well suited to representative members of the category. According to such a view, an agent may have several preference relationships. Non-selfish preferences would be likely to appear in some environments (for example, those resembling repeated interaction with close associates) than in others.

Finally, preferences appear to be especially fluid during the earlier years of life and especially susceptible to the influence of others. Most children find themselves surrounded by supportive, cooperative informants. The hypothesis that they are

[60] Cheng and Holyoak (1985) and Davies, Fetzer, and Foster (1995) offer critiques and alternative interpretations.

[61] Even if one accepts Cosmides and Tooby's interpretation of their data, their hypothesis only states that humans are equipped to identify certain types of cheating behavior. Standard models in economics typically assume that agents have this ability. Cosmides and Tooby's experiments do not explain why agents would lower their material payoffs to respond to cheaters.

playing in cooperative strategic environments with cooperative players gets reinforced. The existence of a supportive environment is plainly essential: The child will die without food. The *recognition* of the nature of the environment is also essential for the development of certain skills, notably language. The ability of humans to learn natural language presumably requires operating under hypotheses that inputs are reliable. What mommy calls a ball really is a ball. At least, it is something that everyone calls a ball. The ability to acquire language may be more important than the ability to avoid being duped in economic exchanges later in life and (perhaps) biological hardware may bias individuals to follow cooperative strategies even when they are not fitness maximizing.[62]

Parents and schools attempt to teach children to be non-selfish. Parents' dominant position may enable them to induce their children to internalize preferences that benefit the parents. Since conflicts of interest exist between parents and children, it will not be in the best interest of parents to have selfish children. In particular, parents stand to gain from having children who are willing to repay kindness with kindness.[63] To the extent that changes in their preferences operate against their self interest, children should be expected to resist the changes. Since children have so much to gain from trusting their parents, and since the trust is frequently well placed, efforts to internalize preferences may be effective nevertheless. Experimental evidence that preferences are age dependent (for example the Harbaugh, Krause, and Liday (2003) study of the ultimatum game) are consistent with the possibility that preferences are fluid at some stages in life and that people learn to adopt prosocial preferences.

5.6 Summary

This section reviewed literature aimed at providing evolutionary foundations for non-selfish preferences. The results are mixed. If one takes the position that pressures that lead to selection of genetic material determines an individual's preferences, then there is no reason why different selection pressures faced by different genes in a single individuals will direct the individual to maximize fitness. The literature in economics does assume that selection operates at the level of the individual. In this setting, the clearest theoretical setting for the survival of non-selfish preferences is an environment

[62]Simon ((1990) and (1993)) argues that people should be receptive to social influences because the benefits from access to cultural wisdom are greater than the costs associated with following society's suggestions to help others.

[63]Alexander (1987, page 103) writes that a basis for moral systems is the strong incentive for individuals to influence others to be "beneficient" to others. Parents may be particularly effective teachers.

in which preferences are observable or, more generally, it is costly for selfish agents to appear otherwise. This condition sets the stage for an evolutionary "arms race" in which there is co-evolution of both abilities to misrepresent preferences and to detect misrepresentation. At the other extreme, when there is incomplete information about strategies, there is theoretical support for the view that equilibrium behavior will coincide with fitness maximizing equilibrium behavior for all agents.

I draw three conclusions from the work described. First, the analysis is consistent with the idea that predictions based on the assumption of maximizing material self-interest need not be accurate in small-group settings. In small groups, agents are more likely to know each other. Preferences are more likely to be observable (either directly or through a signal that is linked to preferences). Sanctions based on expulsion are feasible.[64] These sanctions can reduce the fitness of selfish agents.

Second, in all of the models, there is scope for fitness maximizers. The crude interpretation of this observation is that self-interested behavior, narrowly defined, will always be with us. To the extent that other preferences appear in the population, they are balanced by traditional economic preferences.[65]

That different people have different preferences is hardly controversial,[66] but it raises an important question. Under what conditions on the strategic environment (or economic institution) do standard predictions remain valid when only a fraction of the agents maximize material payoffs? Experimental results in auctions and best-shot bargaining suggest that standard predictions continue to hold in competitive environments. This insight awaits a complete characterization.

Third, the evolutionary approach does not impose structure on preferences. The cautious conclusion from current research is only that there is no strong argument for ruling out all behavior that does not maximize material payoffs.

[64]Expelled agents may have difficulty gaining entrance to other groups.

[65]A serious evolutionary study that permits heterogeneity of preferences must confront the issue of whether spatially isolated people facing different conditions might evolve differently. People from different areas may have genetic predispositions towards different preferences. For example, hunters face qualitatively different strategic situations than gathers. Hunters (of large game) must cooperate in order to gather food successfully. Gatherers can usually collect food without cooperation. One could imagine that if the style of food gathering was the essential arena for the evolution of preferences, then there would be different levels of fitness maximizers within a population of hunters than in a population of gatherers. On the other hand, to the extent that these differences ever existed, mixing of populations would diminish them. Diamond (1997) argues convincingly that it is unnecessary to invoke genetic differences to explain broad patterns of world history.

[66]The proposition does shake the foundation of the methodology put forth in Stigler and Becker (1977), who argue that economists should seek to explain behavior under that assumption that "tastes neither change capriciously nor differ importantly between people."

6 Closing Arguments

> If you would like to be selfish, you should do it in a very intelligent way. The stupid way to be selfish is . . . seeking happiness for ourselves alone. . . . The intelligent way to be selfish is to work for the welfare of others.
>
> The Dalai Lama

Economics, which frequently relies on the joint hypotheses of intelligence and self interest, should be open to models in which agents are selfish in intelligent ways.[67] This paper reviews reasons why more narrowly conceived models may be insufficient, describes a variety of alternatives, and suggests possible applications. This section contains response to arguments against the approaches discussed in the paper.[68]

There are at least three problems with this list. First, it treats "standard theory" as a single object, rather than many different approaches. Second, I could not find clear statements of criticisms in the literature. I may have ignored or weakened the strongest criticisms. Third, I did not give the criticisms precise mathematical formulations. This makes the boundaries of the argument unclear and makes it impossible for any of the arguments or any of the counter arguments to be decisive. Some of the criticisms lend themselves to formal analysis.

6.1 If It Is Not Broken, Do Not Fix It

Argument. Standard theory works. There is no need to change it.

Response. Conventional theory does work well in many situations.[69] The paper reviews evidence that it fails to account for many interesting findings without being stretched and strained. Enlarging the toolkit to include more general models could expand the range of useful economic analysis.

Adopting a broader perspective also permits us to view apparent successes of standard theory as evidence that assumptions are too strong. When laboratory models confirm predictions of conventional models, for example in market settings, we can look for less restrictive assumptions that make the same predictions.

[67] I found the Dalai Lama's quotation in the review essay on Sober and Wilson (1998) by primatologist Barbara Smuts (1999). She argues that many types of apparently unselfish behaviors are viable at the level of an organism.

[68] My list overlaps the list in Section III of Conlisk (1996), which provides a parallel defense of bounded-rationality modeling.

[69] See Lazear (**?**) for a compelling review of the successes of the Chicago school.

6.2 Complexity

Argument. Allowing extended or context-dependent preferences leads to models that impose unrealistic demands on agents' abilities to reason and modelers' abilities to characterize equilibria.

Response. Traditional rationality assumptions do not impose any limits on the computational abilities of agents. There is no justification for imposing limits at an arbitrary point that separates traditional models from the ones described in this paper.

There are several examples, notably Bolton and Ockenfels (2000) and Fehr and Schmidt (1999), of tractable models involving interdependent preferences. Arguably, none of the proposed models of reciprocity in games is as simple as these models, but further work may change that. To the extent that the context-dependent models capture a genuine intuition about behavior and help to organize observations and hypotheses, they are useful.

6.3 Only the Selfish Survive

Argument. A careful study of the origin of preferences proves that only selfish agents survive.

Response. With sufficient freedom to define "selfish" this statement is a tautology. With a narrow conception of selfish behavior, it is possible to provide formal models that support the assertion, but the models rely on strong assumptions. There are plausible reasons to believe that preferences for non-selfish behavior and reciprocity have survival value and have survived. Until economics becomes a special case of molecular biology there will be reason to examine reduced-form models that permit non-selfish behavior.

6.4 Generality

Argument. The standard tools of economic analysis apply to a wide range of problems; no other approach has the same range. Or, in the words of Stigler and Becker (1977, pages 76-7): "this traditional approach of the economist offers guidance in tackling these problems and that no other approach of remotely comparable generality and power is available."

Response. Relaxing the assumptions of the traditional approach creates a theory of more generality and more power.

6.5 Discipline

Argument. Economics needs the discipline provided by the assumption of self-interested behavior to generate behavioral hypotheses and predictions based on well understood general principles.

Response. What John Conlisk (1996, page 685) wrote in defense of bounded rationality modeling is appropriate here: "Discipline comes from good scientific practice, not embrace of a particular approach. Any approach . . . can lead to an undisciplined proliferation of hypotheses to cover all facts." Even within conventional economic theory, individual greed can mean many things. We are comfortable abandoning risk neutrality to study gambling and insurance behavior. We are comfortable abandoning myopic optimization to study dynamic interactions. We are comfortable inserting unmarketed goods into utility functions to model externalities. In the wrong hands, these modifications reflect a lack of discipline. Properly used, they reflect a willingness to extend and revise the principles of equilibrium and optimization in order to explain behavior.

6.6 Definite Outcomes

Argument. Standard models provide clear predictions. Expanding the domain of preferences makes it impossible to obtain definite answers to economic questions.

Response. Traditional economic methodology owes its power to its generality and its flexibility. In markets and strategic settings indeterminacy is the rule. If preferences need only satisfy standard neo-classical assumptions, then there are essentially no restrictions on aggregate excess demand functions and therefore no restrictions on market-clearing prices.[70] In game-theoretic environments, even when preferences are specified, multiple equilibrium problems arise. There is no generally accepted way to select among multiple equilibria, especially in repeated-game environments where the folk theorem places no serious restrictions on what can be observed. Clear predictions come only as a result of imposing strong assumptions on preferences or action sets.

[70]For a careful statement of the results, see Debreu (1974), Mantel (1974), and Sonnenschein (1973). Hildenbrand's (1994) approach provides some limitations on aggregate behavior.

Not only is there indeterminacy once a model has been specified, there is no limit to the number of different models one can propose that have a plausible connection to an economic problem. Coming up with a model to explain observations is not difficult ex post, the challenge is to come up with a useful model that applies to more than one situation.

6.7 Parsimony

Argument. Models of extended preferences introduce too many free variables. The theory explains everything, therefore it explains nothing.

Response. It is important to distinguish the set of all predictions that come from a theory from the set of predictions obtained from a particular specification within the theory. Any of the approaches described in Section 3 provide an imaginative modeler sufficient scope to summarize empirical regularities. For example, the folk theorem of repeated games guarantees that practically any outcome can be supported as an equilibrium of a repeated interaction between patient players. While the theorem does have assumptions, it is hard to imagine any outcome from a dynamic interaction that could not be described as an equilibrium of some repeated game. We judge the value of a particular repeated game model on its explanatory power, generality, plausibility, and elegance. But these criteria are informal, artistic categories rather than logical ones. A test of a class of models is its ability to provide useful descriptions and predictions. The new models of interdependent preferences (of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999)) or reciprocity (Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fishbacher (forthcoming),and Rabin (1993)) put forth specific parametric versions of their models. These models supply refutable implications. Parsimony demands that we obtain our descriptions from a relatively small collection of available parameterizations.

7 Conclusion

The bully who boasts that he can beat his foes with one hand tied behind his back will prove his claim by picking his foes wisely. But he will look awkward when he wins and look foolish if he loses. His boast is less a signal of strength than an attempt to intimidate. Restricting theory to use only a subset of available tools is not discipline. It is a handicap.

The hypothesis that reciprocity is an instrumental motivation for human behavior is overwhelming. There are good reasons to reciprocate in dynamic interactions

as cooperation generates future cooperation and retaliation may serve to inhibit exploitation. There is strong evidence that the desire to reciprocate is an intrinsic aspect of preferences. There is strong social pressure to internalize a preference for reciprocity. There are plausible stories that describe why these motivations persist. While the body of evidence cannot establish the truth of the hypothesis, the evidence is sufficiently strong and the advantages sufficiently clear, to justify continued development of the modeling tools that I have discussed. A philosophical refusal to consider extended preferences leads to awkward explanations of some phenomena. It limits the questions that can be asked and restricts the answers. It is a handicap.

Extending the arguments of preferences and permitting the preferences to change with context in a systematic way enables theorists to continue to use economic theory to predict and explain the impact of parameter changes, while expanding the scope of the theory. These models promise a language for the study the effect of market institutions and contracts on economic performance.

To take the interdependent preference theory seriously, work should proceed on three fronts. We need to develop foundational theory to identify general properties of extended preferences. We need to apply the theory to specific problems and develop restrictions leading to tractable models that efficiently summarize what we observe and generate interesting hypotheses. We need experimental work to investigate these hypotheses and provide evidence about whether preferences are stable across games, roles, and individuals.

References

Abreu, D. (1988): "Towards a Theory of Discounted Repeated Games," *Econometrica*, 56, 383–396.

Abreu, D., D. Pearce, and E. Stacchetti (1993): "Renegotiation and Symmetry in Repeated Games," *Journal of Economic Theory*, 60, 217–240.

Akerlof, G. (1982): "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 97, 543–569.

Akerlof, G., and R. Kranton (2000): "Economics and Identity," *Quarterly Journal of Economics*, 103, 715–753.

Alexander, R. (1987): *The Biology of Moral Systems*. Aldine de Gruyer, New York.

Andreoni, J. (1990): "Impure Altruism and Donations to Public-Goods — A Theory of Warm-Glow Giving," *Economic Journal*, 100, 464–477.

——— (1995): "Cooperation in Public Goods Experiments: Kindness or Confusion?," *American Economic Review*, 85, 891–904.

——— (2001): "The Economics of Philanthropy," in *The International Encyclopedia of the Social and Behavioral Sciences*, ed. by N. Smelser, and P. Baltes, pp. 11369–11376. Elsevier, Oxford.

Andreoni, J., P. Brown, and L. Vesterlund (2002): "What Produces Fairness? Some Experimental Results," *Games and Economic Behavior*, 40, 1–24.

Becker, G. S. (1962): "Irrational Behavior and Economic Theory," *Journal of Political Economy*, 70, 1–13.

Bénabou, R., and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, 489–520.

Bereby-Meyer, Y., and M. Niederle (forthcoming): "Fairness in Bargaining," *Journal of Economic Behavior and Organization*.

Bergstrom, T. C. (2002): "Evolution of Social Behavior: Individual and Group Selection," *Journal of Economic Perspectives*, 16, 67–88.

Bewley, T. (1999): *Why Wages Don't Fall During a Recession*. Harvard University Press, Cambridge.

Binmore, K., J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked (2002): "A Backward Induction Experiment," *Journal of Economic Theory*, 104(1), 48–88.

Boehm, C. (1993): "Egalitarian Behavior and Reverse Dominance Hierarchy," *Current Anthropology*, 34(3), 227–240.

——— (1997): "Impact of the Human Egalitarian Syndrome an Darwinian Selection Mechanics," *The American Naturalist*, 150, S100–S121.

Bolton, G., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity and Competition," *American Economic Review*, 90, 166–193.

Bowles, S. (1998): "Endogenous Perferences: The Cultural Consequences of Markets and Other Economic Institutions," *Journal of Economic Literature*, 36, 75–111.

Boyd, R., and P. J. Richerson (1985): *Culture and the Evolutionary Process*. University of Chicago Press, Chicago.

——— (1988): "The Evolution of Reciprocity in Sizable Groups," *Journal of Theoretical Biology*, 132, 337–356.

Cabrales, A., and G. Charness (2000): "Optimal Contracts, Adverse Selection, and Social Preferences: An Experiment," Discussion Paper 478, Universitat Pompeu Fabra.

Charness, G., and E. Haruvy (2002): "Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach.," *Games and Economic Behavior*, 40(2), 203–231.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics*, 117(3), 817–869.

Cheng, P. W., and K. J. Holyoak (1985): "Pragmatic Reasoning Schemas," *Cognitive Psychology*, 17, 391–416.

Cialdini, R. C., and M. R. Trost (1998): "Social Influence: Social Norms, Conformity, and Compliance," in *The Handbook of Social Psychology*, ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey, vol. 2, chap. 21, pp. 150–192. McGraw-Hill, Boston, 4th edn.

Conlisk, J. (1996): "Why Bounded Rationality?," *Journal of Economic Literature*, 34, 669–700.

Cosmides, L., and J. Tooby (1992): "Cognitive Adaptations for Social Exchange," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, ed. by J. H. Barlow, L. Cosmides, and J. Tooby. Oxford University Press, Oxford.

Costa-Gomes, M., and K. Zauner (2001): "Ultimatum bargaining behavior in Israel, Japan, Slovenia, and the United States: A Social Utility Analysis," *Games and Economic Behavior*, 34(2), 238–269.

Cox, J. C. (2004): "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46(2), 260–281.

Croson, R. T. A. (1999): "Theories of Reciprocity and Altruism: Evidence from Linear Public Goods Games," Discussion paper, University of Pennsylania.

Dasgupta, P. (2000): "Economic Progress and the Idea of Social Capital," in *Social Capital: A Multifaceted Perspective*, ed. by P. Dasgupta, and I. Serageldin. World Bank, Washington.

Davies, P. D., J. H. Fetzer, and T. R. Foster (1995): "Logical Reasoning and Domain Specificity: A Critique of the Social Exchange Theory," *Biology and Philosophy*, 10, 1–37.

Dawkins, R. (1982): *The Extended Phenotype*. Oxford University Press, Oxford.

De Waal, F. B. M. (1996): *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press, Cambridge.

Debreu, G. (1974): "Excess Demand Functions," *Journal of Mathematical Economics*, 1, 15–21.

Deci, E. L., and R. M. Ryan. (1985): *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum, New York.

Diamond, J. (1997): *Guns, Germs, and Steel: The Fates of Human Societies*. Norton, New York.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

Dugatkin, L. A. (1997): *Cooperation Among Animals: An Evolutionary Perspective*. Oxford University Press, New York.

Ekman, P. (2001): *Telling Lies*. W. W. Norton, New York.

Ellingsen, T., and J. Robles (2002): "Does evolution solve the hold-up problem?," *Games and Economic Behavior*, 39(1), 28–53.

Ely, J., and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Falk, A., E. Fehr, and U. Fischbacher (2003): "On the Nature of Fair Behavior," *Economic Inquiry*, 41(1), 20–26.

Falk, A., and U. Fischbacher (forthcoming): "Modeling Fairness and Reciprocity," in *Strong Reciprocity*, ed. by S. Bowles, R. Boyd, E. Fehr, and H. Gintis. MIT Press, Cambridge.

Farrell, J., and E. Maskin (1989): "Renegotiation in Repeated Games," *Games and Economic Behavior*, 1, 327–360.

Fehr, E., and S. Gächter (2000): "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14, 159–181.

——— (2002): "Do Incentive Contracts Crowd out Voluntary Cooperation," Working Paper 34, University of Zurich, http://www.iew.unizh.ch/wp/iewwp034.pdf.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, 65, 833–860.

Fehr, E., E. Kirchler, A. Weichbold, and S. Gächter (1998): "When Social Forces Overpower Competition: Gift Exchange in Experimental Labor Markets," *Journal of Labor Economics*, 16, 324–351.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, 108, 437–459.

Fehr, E., and K. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

——— (2000): "Fairness, Incentives and Contractual Choices," *European Economic Review*, 44, 1057–1068.

——— (forthcoming): "Theories of Fairness and Reciprocity — Evidence and Economic Applications," in *Advances in Economics and Econometrics: 8th World Congress*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky.

Frank, R. H. (1988): *Passions within Reason*. Norton, New York.

Fremling, G., and R. Posner (1999): "Market Signaling of Personal Characteristics," Discussion paper, University of Chicago.

Fudenberg, D., and E. Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or With Incomplete Information," *Econometrica*, 54, 533–556.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Ecoonomic Behavior*, 1, 60–79.

Geertz, C. (1963): *Peddlers and Princes*. University of Chicago, Chicago.

Gintis, H. (2000): "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology*, 206, 169–179.

Gneezy, U., and A. Rustichini (2000a): "A Fine is a Price," *Journal of Legal Studies*, 29, 1–18.

Gneezy, U., and A. Rustichini (2000b): "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 115, 791–810.

Gode, D. K., and S. Sunder (1993): "Allocative Efficiency of Markets as a Partial Substitute for Individual Rationality," *Journal of Political Economy*, 101, 119–137.

Goeree, J. K., and C. Holt (2000): "Asymmetric Inequality Aversion and Noisy Behavior in Alternating-Offer Bargaining Games," *European Economic Review*, 44, 1079–1089.

Gül, F. (2001): "Unobservable Investment and the Hold-Up Problem," *Econometrica*, 69, 343–376.

Güth, W. (1995a): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24, 323–344.

——— (1995b): "On Ultimatum Bargaining Experiments — A Personal Review," *Journal of Economic Behavior and Organization*, 27, 329–344.

Güth, W., R. Schmittberger, and B. Schwarze (1982): "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 3, 367–388.

Güth, W., and M. E. Yaari (1992): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change – Approaches to Evolutionary Economics*, ed. by U. Witt. University of Michigan Press, Ann Arbor.

Guttman, J. M. (2000): "On the Evolutionary Stability of Preferences for Reciprocity," *European Journal of Political Economy*, 16, 31–50.

Hamilton, W. (1964): "The Genetical Evolution of Social Behavior," *International Journal of Theoretical Biology*, 7, 1–17.

Harbaugh, W., K. Krause, and S. Liday (2003): "Bargaining by Children," Discussion paper, University of Oregon.

Harrison, G., and J. Hirschliefer (1989): "An Experimental Evaluation of the Weakest Link, Best Shot Models of Public Goods," *Journal of Political Economy*, 97, 201–225.

Hart, O. (1995): *Firms, Contracts, and Financial Structure*. Clarendon Press.

Henrich, J. (2004): "Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation," *Journal of Economic Behavior and Organization*, 53, 3–35.

Henrich, J., R. Boyd, S. Bowles, H. Gintis, E. Fehr, R. McElreath, and C. Camerer (2001): "Cooperation, Reciprocity and Punishment: Experiments in 15 Small-Scale Societies," *American Economic Review*, 91, 73–78.

Herold, F. (2003): "Carrot or Stick? Group Selection and the Evolution of Reciprocal Preferences," Discussion paper, University of Munich.

Hildenbrand, W. (1994): *Market Demand: Theory and Empirical Evidence*. Princeton University Press, Princeton.

Holmström, B., and P. Milgrom (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7, 24–52.

Homans, G. (1953): "Status Among Clerical Workers," *Human Organization*, 12, 5–10.

——— (1954): "The Cash Posters," *American Sociological Review*, 19, 724–733.

Koçkesen, L., E. A. Ok, and R. Sethi (2000a): "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior*, 31, 301–310.

——— (2000b): "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory*, 92(2), 274–299.

Kranton, R. (1996): "Reciprocal Exchange: A Self-Sustaining System," *American Economic Review*, 86(4), 830–851.

Kreps, D. (1997): "Intrinsic Motivation and Extrinsic Incentives," *American Economic Review, Papers and Proceedings*, 87, 359–364.

Kropotkin, P. (1902): *Mutual Aid: A Factor of Evolution*. William Heinemann, London.

Lazear, E. (2000): "Performance Pay and Productivity," *American Economic Review*, 90, 1346–1361.

Lepper, M. R., D. Greene, and R. E. Nisbett (1973): "Undermining Children's Intrinsic Interest with Extrinsic Reward," *Journal of Personality and Social Psychology*, 28, 129–137.

Levine, D. (1998): "Modelling Altruism and Spitefulness in Game Experiments," *Review of Economic Dynamics*, 1, 593–622.

Mantel, R. (1974): "On the Characterization of Aggregate Excess Demand," *Journal of Economic Theory*, 7, 348–352.

McCabe, K., S. Rassenti, and V. Smith (1998): "Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining," *Games and Economic Behavior*, 24, 10–25.

Ok, E. A., and F. Vega-Redondo (2001): "On the Evolution of Individualist Preferences: An Incomplete Information Scenerio," *Journal of Economic Theory*, 97, 231–254.

Palfrey, T., and J. Prisbrey (1997): "Anomalous Behavior in Public Goods Experiments: How Much and Why," *American Economic Review*, 87, 829–846.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior and Organization*, 42, 125–129.

Postlewaite, A. (1998): "The Social Basis of Interdependent Preferences," *European Economic Review*, 42, 779–800.

Prasnikar, V., and A. Roth (1992): "Considerations of Fairness and Strategy: Experimental Data from Sequential Games," *Quarterly Journal of Economics*, 79, 355–384.

Price, G. (1970): "Selection and Covariance," *Nature*, 227, 520–521.

Rabin, M. (1993): "Incorporating Fairness into Game Theory," *American Economic Review*, 83(5), 1281–1302.

Read, D., G. Loewenstein, and M. Rabin (1999): "Choice Bracketing," *Journal of Risk and Uncertainty*, 19, 171–197.

Ridley, M. (1993): *The Red Queen: Sex and the Evolution of Human Nature*. Viking, London and New York.

——— (1996): *The Origins of Virtue*. Penguin, New York.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144, 379–396.

Roth, A. (1995): "Bargaining Experiments," in *Handbook of Experimental Economics*, ed. by J. Kagel, and A. Roth. Princeton University Press, Princeton.

Sahlins, M. (1968): *Tribesmen*. Prentice-Hall, Englewood Cliffs, New Jersey.

Samuelson, P. (1993): "Altruism as a Problem Involving Group versus Individual Selection in Economics and Biology," *American Economic Review, Papers and Proceedings*, 83, 143–148.

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

Segal, U., and J. Sobel (2004a): "Markets Make People Look Selfish," Discussion paper, University of California, San Diego.

——— (2004b): "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," Discussion paper, University of California, San Diego.

Service, E. R. (1966): *The Hunters*. Prentice-Hall, Englewood Cliffs, New Jersey.

Sethi, R., and E. Somanathan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

——— (2003): "Understanding Reciprocity," *Journal of Economic Behavior and Organization*, 50, 1–27.

Simon, H. (1990): "A Mechanism for Social Selection and Successful Altruism," *Science*, 250, 1665–1668.

——— (1993): "Altruism and Economics," *American Economic Review, Papers and Proceedings*, 83, 156–161.

Smuts, B. (1999): "Multilevel Selection, Cooperation, and Altruism: Reflection on *Unto Others*," *Human Nature*, 10, 311–327.

Sober, E., and D. Wilson (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge.

Solow, R. (1979): "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics*, 1, 79–82.

Sonnenschein, H. (1973): "Do Walras' Identity and Continuity Characterize the Class of Community Excess Demand Functions?," *Journal of Economic Theory*, 6, 345–354.

Stigler, G., and G. Becker (1977): "De Gustibus Non Est Disputandum," *American Economic Review*, 67, 76–90.

Sugden, R. (1984): "Reciprocity: The Supply of Public Goods Through Voluntary Contributions," *Economic Journal*, 94, 772–787.

Thaler, R. H. (1999): "Mental Accounting Matters," *Journal of Behavioral Decision Making*, 12, 183–206.

Trivers, R. L. (1971): "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, 46, 35–58.

Varela, F. (1999): *Ethical Know-How*. Stanford University Press, Stanford.

Wason, P. (1966): "Realism and Rationality in the Selection Task," in *Thinking and Reasoning: Psychological Approaches*, ed. by J. Evans. Routledge and Kegan Paul, London.

Yamagishi, T. (1988): "The Provision of a Sanctioning System in the United States and Japan," *Social Psychology Quarterly*, 51(3), 265–271.

Yamagishi, T., K. S. Cook, and M. Watabe (1998): "Uncentainty, Trust, and Commitment Formation in the United States and Japan," *American Journal of Sociology*, 104, 165–194.