Social Preferences and Public Policy: Are good laws a substitute for good citizens?

Samuel Bowles
Santa Fe Institute and University of Siena[1]
20 July, 2005

Abstract

In a second best world of incomplete contracting, laws and policies designed to harness self-regarding preferences to public ends may fail. These failures occur, I will suggest, when conventional self-interest-based policies limit the effectiveness of governance processes that go beyond the usual fiat and contract approach to implementation, and that rely on informal enforcement strategies and the activation of social preferences Experimental evidence indicates that incentives that appeal to self interest may reduce the salience of other-regarding preferences and other civic motives. Historical cases suggest that for this and additional reasons, institutional crowding out occurs. The evidence for these processes is reviewed and a model illustrating the possibly counter productive nature of the conventional approach is presented.

JEL: D64, D52, H41, H23, Z13, C92

Keywords: Social preferences, implementation theory, incentive contracts, incomplete contracts, framing, behavioral experiments, motivational crowding out, institutional complementarities, ethical norms, constitutions

---

[1] Draft for discussion at the workshop on Happiness and Public Economics, Princeton University, 24-25 May, 2005. Thanks to the Behavioral Sciences Program of the Santa Fe Institute and the University of Siena for financial support of this project, and the Certosa di Pontignano (Siena) for providing an ideal research environment. I would like to thank Margaret Alexander, Iris Bohnet, James Boyce, Juan Camilo Cardenas, Ernst Fehr, Simon Gaechter, John Geanakoplos, Amara Moore-Levy, Elinor Ostrom, Michael Kosfeld, Paul Seabright, Rajiv Sethi, E. Somanathan, Tim Taylor and Elisabeth Wood for their contributions to this research.

> Lawgivers make the citizen good by inculcating habits in them, and this is the aim of every lawgiver; If he does not succeed in doing that, his legislation is a failure. It is in this that a good constitution differs from a bad one.
>
> Aristotle (1962):103, *Nicomachean ethics* (350 b.C)

1. Introduction

The classical thinkers from Aristotle to Thomas Aquinas, Jean-Jacques Rousseau, and Edmund Burke recognized the cultivation of civic virtue not only as the test of good governance, but also as its essential foundation.. All stressed the other-regarding, rule abiding and norm-enforcing behavior documented in recent behavioral experiments. Nicolò Machiavelli's *The Prince* (1513) , and Thomas Hobbes' *Leviathan* (1651) represented a sharp break with this Aristotelian tradition. These founding works of modern political philosophy took self-interest as a fundamental behavioral assumption and asked how the potentially destructive consequences of the autonomous pursuit of individual gain might be constrained by the authority of a sovereign ruler.

The response of the classical economists was that good laws are those that harness selfish motives for public ends. This was the key insight of Bernard Mandeville's *Fable of the Bees*. The subtitle of the 1714 edition of the *Fable* ((Mandeville (1924))) announced that the work contained "..several discourses to demonstrate that human frailties...may be turn'd to the advantage of civil society, and made to supply the place of moral virtues," with the result, he explains in the text (p.24), that "the worst of all the multitude did something for the common good." In contrast to the Aristotelian view that good laws make good citizens, Mandeville suggested that the right rules of the game governing social interactions might harness shabby motives to elevated ends.

Thus in his *Essays*: *Moral, Political and Literary* (1742), David Hume (1964):117-118 recommended the "maxim" that

> in contriving any system of government ... every man ought to be supposed to be a *knave* and to have no other end, in all his actions, than private interest. By this interest we must govern him, and, by means of it, make him, notwithstanding his insatiable avarice and ambition, cooperate to public good.

In similar spirit, Jeremy Bentham (1970 [1789]) offered his *"Duty and Interest* junction principle: Make it each man's *interest* to observe ... that conduct which it is his *duty* to observe." His *Introduction to the Principles of Morals and Legislation* laid out the public

policy implications of Hume's maxim.[2] Adam Smith's invisible hand and its refinement with the early 20th century welfare economics of Marshall and Pigou provided the economic foundations of this claim.

As a result, most political theorists and constitutional thinkers since the late 18th century have taken the self-regarding *Homo economicus* as their fundamental assumption about behavior, and partly for this reason, have stressed competitive markets, well-defined property rights, and efficient, (and since the 20th century) democratically-accountable states as the critical ingredients of governance. Good rules of the game thus came to displace good citizens as the *sine qua non* of good government. Prices would do the work of morals.

The classical constitutional challenge posed by Bentham, Hume, Smith and others is this: what laws would simultaneously facilitate peoples' the pursuit of their own ends, while inducing each to take adequate account of the effects of their actions on others?[3] They correctly identified the source of the market failures that to this day provide the primary rationale for government interventions, though economics has since considerably sharpened our understanding of what it means to that people 'take adequate account of the effects of their actions on others' and why Pareto-inefficient allocations result when they do not.

If the "others" were our kin, neighbors, or friends, our concern for their well- being or our desire to avoid social sanction might induce us to take account of the effects of our actions on them. Reflecting this fact, an important response to the constitutional challenge – one that long predates the classical economists and that now seems utopian – is that caring for the well-being of others need not be confined to intimates but ought to be extended to all of those with whom one interacts. However, with the increasing scope of markets over the last half a millennium, individuals have come to interact not with a few dozen, but with hundreds and indirectly with millions of strangers. And so, with the maturation of capitalism and growing influence of economic reasoning, the burden of good governance shifted from the task of cultivating civic virtue to the challenge of designing institutions that work tolerably well in its absence. Prices, not ethics would ensure that actors took account of the effects of their actions on others.

This remains the canonical model of policy-making in economics. Hume's maxim is

---

[2] The quote is from Harrison (1983):118 citing Pauper VII 380.

[3] Smith's most famous endorsement of 'natural liberty' at the end of book IV of the *Wealth of Nations* is the following: "Every man, as long as he does not violate the laws of justice, is left perfectly free to pursue his own interest his own way." In the same paragraph he makes it clear that justice requires "every member of the society [to be protected] from the injustice or oppression of every other member of it."

beautifully illustrated by implementation theory and mechanism design (Laffont (2000), Maskin (1985), Hurwicz (1975)). These approaches seek to determine the contracts, property rights and other social rules – in short, constitutions – that induce individuals with conventional self-regarding preferences to implement (as a Nash equilibrium of a non-cooperative game) an outcome which is not sought by any of the individual participants, but which is socially valued. Moreover, because they believe that our institutions do this job tolerably well, many economists would not take issue with the philosopher David Gauthier (1986):96 when he writes that: "morality has no application to market interactions under the conditions of perfect competition."

Recent advances in behavioral economics and cognitive psychology point to a number of shortcomings of this approach. First, while self-interest is a powerful motive, significant fractions individuals of most populations studied violate the predictions of the standard model of self-regarding individuals. Many people have other-regarding references; their actions are explained not only by the own-states that they will bring about, but also take some account of the states experienced by others (Camerer and Fehr (2004), Camerer (2003)). Second, evaluations of states are not independent of the processes that produce the state -- intentions of others matter as well as their actions (Blount (1995)). For this and other reasons behaviors are sensitive to framing (Kahneman and Tversky (2000), Frey and Jegen (2003)). Finally, preferences evolve under the influence of the tasks and social interactions that make up the social lives of individuals (Bowles (1998), Henrich, Boyd, Bowles, Fehr, and Gintis (2004)). I will use the term social preferences to refer to their other-regarding, process-dependent, and endogenous aspect.[4]

Social preferences attest to the internalization of ethical norms. This is suggested by the fact that people are willing to reduce their own material gains in order to reduce the gains of those who have acted selfishly or have exploited the cooperation of another, even when the target of the punishment has not harmed them personally (Kahneman, Knetch, and Thaler (1986), Fehr and Fischbacher (2004)).

The ingenious classical response to the constitutional challenge – let prices to the work of morals – was not motivated by the belief that economic actors are amoral. Hume, in the

---

[4] Bowles (2004). By preferences I mean reasons for behavior, that is attributes of individuals -- other than beliefs and capacities -- that account for the actions they take in a given situation. Thus preferences are not summary statements of what the individual *does* or what the individual *wants* (the two standard views in economics). I do not explicitly address comparison-based utility functions (Layard (1980), Fehr and Schmidt (1999), Bowles and Park (2005)) though much of what is said below is applicable in these cases as well. Comparison-based utility is addressed in the two companion papers by Robert Frank and Richard Layard and Steve Nickell.

sentence immediately after the quote above, mused that it is "strange that a maxim should be true in politics which is false in fact."[5] Instead they reasoned that when large numbers of strangers interact, ethical behavior would be an insufficient basis for good government. What I call "Mandeville's mistake" was not that the classicals ignored moral behavior, but instead that assumed it would be unaffected by incentive-based policies designed to harness self-interest. Along with civic virtue, explicit incentives and constraints could thus contribute additively , as it were, to good government. As a result of this implicit additivity assumption' they failed to take account of the conditions under which civic virtue can flourish and favorably affect aggregate outcomes.[6]

In the pages that follow I suggest that in a second-best world of incomplete contracting, policies designed to harness self-regarding preferences to public ends may be counter-productive.[7] These failures occur when conventional self-interest-based policies limit the effectiveness of governance processes that go beyond the usual fiat and contract approach to implementation, and that rely also on informal enforcement strategies and the activation of social preferences. The importance of these governance processes has long been recognized and has recently been documented empirically with respect to the management of common pool resources (Ostrom (1990), Baland and Platteau (1996)), tax compliance (Pommerehne and Weck-Hannermann (1996), Andreoni, Erand, and Feinstein (1998)) and generalized obedience to law (Kahan (1997)).

The sometimes counter-productive effects of what economists might consider improved incentives operate through institutional crowding out. I review historical and other non-experimental cases of this process in the next section. In section 3, I review experiment evidence, showing that the classical additivity assumption is often violated. The experiments

---

[5] Smith (1976 [1759]):3: "How selfish soever man may be supposed, there are evidently some principles in his nature that interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it."

[6] This is a surprising oversight, as the proposition that preferences are endogenous was common among the classicals (Mill (1867[1848]), Bentham (1970 [1789]) Smith (1937 [1776])) and the idea that markets might degrade morals was voiced by many of their contemporaries (Burke (1955 [1790]), de Tocqueville (1958 [1830]), Marx (1959 [1847]):29). Also common at the time was the "*doux commerce*" view expressed by William Robertson (1769) that "commerce...softens and polishes the manners of men."

[7] Related views have been advanced by Titmuss (1971), Taylor (1976), Ben-Porath (1980), Hirschman (1985), Bowles (1989), Frohlich and Oppenheimer (1995), Frey (1997) Cooter (1998),, Ostrom (2000) and others.

suggest that it fails because explicit incentives provide cues that self-interested behavior is appropriate or alter the information structure of a game in ways that reduce the salience of fair-mindedness, reciprocity, and other motives. (I use 'incentives' without adjective to mean incentives appealing to self-regarding preferences.) In section 4 I then propose a utility function that captures much of the motivation exhibited in experiments and use it to illustrate the reasons for the failure of the additivity assumption. In section 5 I model the evolution of preferences and equilibrium selection among 'culture-contract' equilibria, showing how policies to improve contracts may either crowd in or crowd out other regarding motives. The concluding section proposes a revision of Hume's maxim.

2. Institutional complementarity and crowding out

Institutional complementarity exists when the beneficial effects of one institution are enhanced by the presence of another institution (Aoki (2001), Milgrom and Roberts (1990)). Institutional crowding out occurs when one institution reduces the effectiveness or viability of another institution. The economics literature subsequent to Warr (1982) has recognized that under some conditions governmental provision of public goods may induce fully offsetting reductions in private contributions (though neither field nor experimental evidence bears out this full crowding out hypotheseis Bolton and Katok (1998)).But the problem is far more general.

As an illustration, suppose that the sustainable use of a common pool resource is accomplished through peer monitoring. The common ownership of the natural asset increases the expected duration of interactions among community members because those who leave surrender their claim on it, thereby providing conditions for effective disciplining of defectors in the management of the common pool resource. Privatizing the asset will enhance exit options and may crowd out the communal management based on long-term repeated interactions. In this case, common ownership and peer monitoring of resource use are complementary institutions

A less hypothetical example of institutional complementarity is provided by the lobster fishermen on the coast of Maine (U.S.) who have for decades regulated their catch by limiting access to a defined fishing territory. Only those belonging to a particular so-called harbor gang -- those fishing from a particular harbor who have been granted membership -- are by local custom are allowed to set their traps in the territory. (Acheson (1988)). Boundary violators are likely to find the buoys cut from their traps which are then impossible to locate. Intruders have been fired upon. Infringements of environmental regulations or the local norms of the particular gang are also sanctioned by other gang members. In recent years, the State of Maine has formalized the gang system by recognizing the territories of the harbor gangs and setting up democratically elected councils with powers to regulate limits on number of traps and numbers of days fishing. State officials occasionally intervene when conflicts exceed the

enforcement capacities of the local communities, as they did during the near collapse of the fishery during the 1920s, or when violence between gangs erupts. But the State employs only six officers to enforce environmental regulations along the entire 4342 mile coastline and to oversee the fishing of 6,800 lobstermen. In recent years, fishing yields have grown and the lobstermen have prospered. The effectiveness of the state's regulations is greatly enhanced by their informal enforcement by the gangs, while the gang's effectiveness is conditioned on the availability of the state as the enforcer of last resort.

The mismanagement of the Himalyan forests of Kamaun and Garhwal districts in Uttar Pradesh (India) provides a contrasing example.[8] Before the 20th century, large well-defined tracts of forests were considered the exclusive property of each village. Access was regulated by the village *panchayats*, and should unauthorized outsiders remove forest products, fighting might break out or fines be levied. To this point forestry management resembled the decentralized regulation by Maine's harbor gangs. But during the First World War, the British colonial administration took over the forestry management, seeking to meet the demand for railroad ties and other wood products. The colonial intervention disrupted the regulation by the local communities and evoked incendiary protests that destroyed large stands of pine. The government, in retreat, awarded access to the less valuable oak forests to "all *bona fide* residents of Kumaun" thereby obliterating the traditional boundaries of village forests and making local regulation virtually impossible. For example, in 1932 a group of villagers from Papdev prevented their neighbor, Jeet Lal, from harvesting grass from the forest, because he had not contributed to the construction of fencing the grass preserve. Jeet Lal took his neighbors to court and *they* were fined, the punishment being upheld on appeal because, according to the new regulations, Jeet Lal had an unconditional right of access.

The government's destruction of the community's capacity to regulate access illustrates the opposite of complementarity, namely *institutional crowding out*. This occurs when the presence of one institution undermines the functioning of another. Another example of crowding out comes from nearby Palanpur (also in Uttar Pradesh) where the extension of the labor market (and increased geographical mobility) appears to have reduced the costs of exit and hence the value of one's reputation, with the effect that the informal enforcement of lending contracts has been undermined (Lanjouw and Stern (1998):570). The development of modern labor markets in central highland Peru appears to have degraded the system of communal labor by which communities produced local public goods (Mallon (1983)). Similar cases in which greater mobility and anonymity of traders compromised preexisting systems of contractual enforcement come from long distance traders in early modern Europe (Greif (1994), Greif (2002)) and shoe manufacturers in Brazil and Mexico.(Woodruff (1998), Schmitz (1999))

[8] This account is based on Sethi and Somanathan (1996) and Somanathan (1991)

My final example is a quasi-natural experiment (Gneezy and Rustichini (2000a)). In Haifa, at six randomly chosen day care centers, a fine was imposed parents who were in picking up their children at the end of the day (in a control group of centers no fine was imposed). Parents responded to the fine by significantly greater tardiness: the fraction picking up their kids late more than doubled. When after 16 weeks the fine was revoked, their enhanced tardiness persisted, showing no tendency to return to the *status quo ante*. Over the entire 20 weeks of the experiment, there were no changes in the degree of lateness at the day care centers in the control group. The counter-productive imposition of the fines appears to illustrate crowding out: using a market mechanism (the fine) seems either to have signaled a low cost of their tardiness or to have undermined the parents' sense of personal obligation to avoid inconveniencing the teachers (Gneezy (2003)).

While it is difficult to establish precise causal connections in non-experimental settings, these cases suggest that institutional crowding out occurs. They are far from unique; Table A1 provides citations to other non-experimental cases in which interventions approximating the canonical welfare economics prescription appear to have crowded out other valued systems of governance.[9] The examples do not establish that crowding out is welfare-reducing, only that it occurs. The perfection of third party contractual enforcement by the Genovese state and traders and the consequent eclipse of the 'collectivist' enforcement strategies described by Greif (1994), for example, surely yielded net economic benefit.

3. Explicit incentives vs civic motives?

If only self-regarding motives are at work, the additivity assumption cannot fail. The reason is that the policy maker is then working with a *tabla rasa*: the mobilization of self regarding motives towards some public end cannot extinguish other motives that might also have contributed to the public benefit. But in a great many experiments this is not the case.

As a benchmark, I will begin with a case of additivity or perhaps even super-additivity. To explore the effects of explicit incentives in the laboratory, Gaechter, Kessler, and Konigstein (2004) implemented a "gift exchange game" (Fehr, Gachter, and Kirchsteiger (1997)) in which principals (employers) make a wage offer with a stipulated desired level of effort on the part of the agent (worker). The agent may then choose an effort level, with costs to the agent rising in effort. In the 'stranger treatment' the pairs were shuffled every period, so that each period was a one-shot interaction for the participants who were certain they would

[9] Distinguishing between experimental and non experimental studies is somewhat arbitrary. For example, Upton (1974) and Gneezy and Rustichini (2000a) could easily be classed as experiments. The criterion I used is that in experiments the subject pool were recruited specially rather than being a natural group.

not encounter any partner more than once. The best response for a self regarding subjects in this treatment is for agents to provide minimal effort irrespective of the wage, and for principals, inducting this, to offer the minimal wage. In the 'partner treatment' the two remained paired over ten periods, and this set of ten periods was itself repeated three times. Because the interaction was repeated with the same partner, subjects with self-regarding preferences in this treatment have reasons to provide higher wages and effort than the minimum, even if they believe their partner also to be self-regarding.

As in earlier experiments with this game, 'employers' made wage offers far more generous than the minimum required to illicit the one unit of effort. The effort offered in return in the stranger treatment is much higher (four times higher) than would have been optimal for a self-regarding 'employee' so we can conclude that social preferences of some sort were at work. Repeated interaction resulted in much higher levels of effort than the stranger treatment, and effort rose over the three sets of treatment and also (except for a dramatic end of game drop off) within the three sets of play. The fact that repetition contributed to cooperation as well as the sharp effort reduction during the last two periods of play show that self interested incentives were effective. (The end game fall off is not likely due to learning as the decline is not monotonic within the sets and high (indeed higher) levels of effort are restored when the second and third set are initiated.). But the repeated interaction did more than to activate self-regarding motives: the reciprocal response to generous wages was 60 percent greater in the repeated treatment than in the one shot. The fact that end of set effort did not fall to the level of the stranger treatment also suggests that while repetition engaged the self regarding motives it also tapped social preferences that the stranger treatment did not evoke.

In this case the provision of incentives for self-regarding subjects improved performance and did not degrade (even enhanced) other regarding preferences. But this is not generally the case. Fehr and Gaechter (2000) implemented the same gift exchange game. In their "trust" treatment, the interaction ends when the agent chooses an effort level, as in the stranger treatment above.. In the "incentive" treatment, following the agent's choice of an effort level, the employer may fine the worker, presumably using this option if the worker's effort level is thought to be inadequate. By contrast with the trust treatment, the incentive treatment links pay to performance and hence represents a more complete contract. In this experiment, the total surplus from the interaction is principal's profits plus the agent's wage minus the cost of effort (and the fine where applicable.)

As above, in the trust treatment, a self-regarding agent would choose the minimum feasible level of effort irrespective of the principal's wage offer, and, anticipating this, a self-regarding principal would offer the minimum wage. As in other experiments of this type subjects did not conform to this expectation: Employers made generous offers and workers' effort levels were strongly conditioned on these offers, high wages being reciprocated by high levels of effort. The introduction of explicit incentives, however, had a negative effect:

average effort levels by agents were substantially *lower*. The additivity assumption failed in this case because under the incentive (fine) treatment, initially generous offers by employers were not reciprocated by higher employee effort, and once employers understood this, they made low offers. Thus he explicit incentive (the threat of the fine) appears to have extinguished reciprocal motivations.

Inequality aversion among the agents may also have been involved. The experiment was constructed so that had subjects responded optimally on the basis of self-regarding preferences, the surplus would have been more that twice as great under the incentive treatment as under the trust treatment. But the total surplus was higher in the trust treatment, by 20 percent in those cases where the principal offered a contract such that the expected fine for shirking exceeded the cost of working (the no shirking condition was fulfilled), and by 53 percent where the principal's contract did not meet the no shirking condition.

An important result of this experiment emerges if we compare the distribution of the surplus under the trust treatment and the incentive treatment. In the incentive treatment (that is, confining our attention to the cases in which the principal's contract fulfilled the no shirking condition) profits are more than double the profits in the trust treatment, while the net payoffs to the agent are less than half. The incentive treatment allowed employers to save enough in wage costs to offset the reductions in work effort. Summarizing this result, the authors write: "the incentive opportunities in the incentive treatment allow principals to increase their profits relative to the trust treatment, but ...this is associated with an efficiency loss."

Perverse incentive effects also occurred in a field experiment in Colombia conducted by Juan Camilo Cardenas (Cardenas, Stranlund, and Willis (2000)).The experiment captured the logic of a common pool resource extraction problem (over-exploitation of forests) faced by the rural people who participated. In the absence of explicit incentives the subjects selected extraction levels not far above the social optimum and much less than what would have been the Nash equilibrium level assuming individual optimization with self-regarding preferences. But when monitoring of the subjects' extraction levels (by the experimenter) and the prospect of a fine for over-extraction were introduced, subjects extracted more rather than less. After a few rounds, their extraction levels approximated the new (self-regarding) Nash equilibrium level (taking account of the fine). The subjects had switched from other-regarding to self-regarding behavior, apparently as a result of the imposition of punishment. Like the fine imposed on the tardy Haifa parents, the effect of "improving" the incentive structure apparently was to diminish the salience of the other-regarding motives that had been in force in the absence of the incentives.

A related experiment may provide some insight into how why the additivity assumption fails (Frohlich and Oppenheimer (1995).) Subjects played 5-person public goods

games under two conditions: one group played the standard contribution game and the other played a modified game in which a randomized assignment of payoffs made it optimal to contribute the maximal amount to the public good. Half of the subjects (in each treatment) were allowed to engage in discussion prior to each play (of course the discussion should have had no effect on the outcome of the standard game, as the dominant strategy is to contribute nothing). After 8 rounds of play, another 8 rounds were conducted, this time with the same groups but with all playing the standard game. Among those who had been permitted discussion, those who had experienced the incentive-compatible (veil of ignorance) game contributed significantly less in the final 8 rounds, and (in subsequent questionnaires) expressed less concern with questions of fairness.

The authors' explanation is that the incentive-compatible mechanism rewarded those contributing to the public good, thus making self-interest a good guide to action, while those experiencing the standard game gained high payoffs only to the extent that they evoked considerations of fairness as a distinct motive among their group-mates.. They conclude

> The failure of the ... (incentive compatible) mechanism to confront subjects with an ethical dilemma appears to lead to little or no learning in ethical behavior in the subsequent period. ... It is an institution, like other incentive compatible devices, which can generate near optimal outcomes. ... However from an ethical point of view it is not only unsuccessful as pertains to subsequent behavior; it appears to be actually pernicious. It undermines ethical reasoning and ethically motivated behavior. (Frohlich and Oppenheimer (1995):44)

This interpretation is consistent with a large literature on the effects of performing various kinds of tasks on subsequent (sometimes seemingly unrelated) values (Breer and Locke (1965).) Other experiments have documented these dynamic crowding out effects (Irlenbusch and Sliwka (2004), Gaechter, Kessler, and Konigstein (2004)). In these two experiments, as in the case of the fines for tardiness at the Haifa day care centers, the negative effects of incentives persisted even after the incentives are no longer operative.

Fehr and List (2004) offered a different interpretation of counter productive incentives found in their trust experiments with Costa Rican businessmen and students. The highest level of trustworthiness was elicited when the principal was *permitted* to fine the agent for untrustworthy behavior, but had *pre-committed not to use it*, evidently a signal by the principal of trusting behavior that was then reciprocated by the agent. By contrast "explicit threats to penalize shirking, backfire by inducing less trustworthy behavior." They conclude: "the psychological message that is conveyed by incentives – whether they are perceived as kind or hostile – has important behavioral effect." Subjects in the identical experiments of Fehr and Rockenbach (2003) exhibited the same behavior. Trustees in the trust experiments

by Falk and Kosfeld (2005) acted less trustworthy (they returned less of the Truster's transfer) when the Truster opted to impose a minimum return rate. In post-play interviews, most agreed with the statement that the imposition of the minimum was a signal of distrust; among the 57 per cent of Trustees who had reduced their transfer after the imposition of the minimum, this view was virtually unanimous (93 percent of them agreed with the statement).

Fines or other negative incentives may have strongly positive effects, however. We know from public goods with punishment experiments (Fehr and Gachter (2000), Ostrom, Walker, and Gardner (1992), Yamagishi (1988)) that defecting members of a group substantially increase their contributions if other members have paid to reduce the defector's payoffs. Bowles, Carpenter, and Gintis (2001) show that punishment is effective even when it is not sufficient to make positive contributions a best response (defined over the game payoffs) and Barr (2001) and {Masclet, 2003 #3370} find that the simple expression of disapproval by fellow members without any material punishment is effective. However, when (as occasionally occurs) high contributing members are punished by peers, they reduce their contributions in subsequent rounds (Bowles, Carpenter, and Gintis (2001)). These results are consistent with the view that negative incentives in the form of expressed disapproval (with or without payoff consequences) may evoke shame (if the subject feels guilty about his contribution) and other aspects of preferences not captured in the game payoffs. In his case negative incentives may 'crowd in' other-regarding motives. However when the target of punishment does not feel guilt, the result of punishment is spite rather than shame.

Experiments (mostly by psychologists) have demonstrated conditions under which extrinsic rewards, such as monetary payment for performance of a task, may diminish one's intrinsic motivation to do the task (Deci, Koestner, and Ryan (1999)). While these experiments continue to generate controversy (Cameron, Banko, and Pierce (2001), Eisenberger and Cameron (1996)), my reading of the evidence is that these crowding out effects appear for interesting rather than boring tasks and when the reward is expected in advance and closely tied to the task performance. One may conclude that performance-based pay in work places may diminish employee's motivation to do tasks which they initially found intrinsically interesting or challenging. But the evidence is also consistent with an important role for explicit (extrinsic) incentives in motivating individuals to do tasks in which they have little intrinsic interest (that is to say, a great many jobs).

These extrinsic reward experiments differ from most economic experiments in two relevant ways. First, the public goods, gift exchange, and other games favored by economists are structured so that a purely self-regarding individual will contribute the minimal amount permitted; the finding of interest is that most experimental subjects do not behave this way. The intrinsic motivation experiments by psychologists consider activities that the subjects initially enjoy doing (painting pictures, for example) and show that pay for performance may degrade these initial positive motivations. Second, the incentives (extrinsic rewards) in the

psychological experiments are typically devised by the experimenter, not as in many of the economics experiments by one or more of the strategically interacting subjects (as the gift exchange or other principal agent game). Thus the extrinsic incentives are not viewed as a signal of the type or intent by another subject. The Colombian experiments by Cardenas et al share this property with the psychological experiments: fines were imposed by the experimenter not by other experimental subjects.

Additional evidence of non additivity is found in other experiments, some of which are summarized in Table A2 (see also Frey and Jegen (2003)). I have not listed here the substantial literature on the 'crowding out of intrinsic by extrinsic motives' as that is adequately surveyed in the works just above.

Drawing general conclusions from these experiments is difficult. The studies are not entirely comparable and have been designed to answer somewhat different questions. Some of the results are consistent with quite contrasting interpretations. For example, the paper by Fischbacher, Fong, and Fehr (2005) found that while low offers by proposers in a dyadic bargaining (Ultimatum) are often rejected by respondents, this occurs much less frequently when there is competition among respondents. This result could be interpreted as showing that market competition crowds out fair-mindedness. But the Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis (2005) study of 15 simple societies could be interpreted as showing the opposite: in that study the likelihood that low ultimatum game offers are rejected was significantly greater in the more market-integrated societies. The experiments and historical cases surveyed thus do not allow any simple interpretation. But a few lessons may be suggested.

4. Why additivity fails.

Incentives work. This is particularly true of positive incentives and applies also to negative incentives that avoid conveying negative information about the type or intentions of those with whom the individual is interacting. In some experiments, the response to variations in a given incentive structure (variations in a piece rate or gain share, for example) very closely approximates what one would expect based on self-regarding preferences (for example, Anderhub, Gaechter, and Konigstein (2000) Irlenbusch and Sliwka (2004)).

The experiments also suggest that when additivity fails, this may be the result of one of the following causes.

Framing. Incentives may signal appropriate behavior shifting the frame from ethical and other-regarding to instrumental and self-regarding, or because the incentives provide a signal of the cost (to another) of the individual's behavior, in which case self-regarding behavior modified by the incentive would seem appropriate behavior. The information

conveyed by these framing effects may induce irreversibilities in behavior.

Information about intent or type. Incentives imposed by one's partner convey information about the partner. It may provide a negative signal about the partner's type or intentions, either in the form of lack of concern about one's well being or lack of trust. However, punishment by peers may stimulate a positive response if it evokes shame rather than spite. Positive incentives may also make it more difficult to evoke reciprocity by signaling good will to another (who may not be able to distinguish generosity from a self regarding response to the positive incentive).

Self-determination. Where intrinsic motivation is present, incentives may 'overjustify' the activity and reduce the individual's sense of autonomy.

Inequality aversion. Where a principal's use of incentives reduces the agent's share of the the surplus, the agent may retaliate in ways that reduce the surplus.

Endogenous preferences. Incentives may alter the duration, information structure, degree of assortation, face-to-face ness and other aspects of social interactions, leading to a long term shift in equilibrium endogenous preferences. These dynamic learning effects may operate across institutions, contributing to institutional crowding out and complementarity, and for failures of the additivity assumption. Included are effects on the costs and effectiveness of peer punishment of transgressions of ethical behavior.

The first four mechanisms are static effects involving given preferences, framing, and beliefs, while the fifth works through endogenous preferences.[10] I will address the learning dynamics associated with endogenous preference in the next section. In this section I model an interaction with social preferences, showing how incentives may degrade performance due to the failure of the classical additivity assumption.

Suppose members of the a population engage in a common project that has the form of a public goods problem. They have the following motivations. They have an *intrinsic motivation* to contribute to the project. They are *self interested* and thus care about their own material payoffs.They are unconditionally *altruistic* or *spiteful* and thus place some weight, positive or negative on the payoffs of others players, independent of their beliefs about the others' types or past behavior. They are *reciprocators* and thus the value they place on the payoff of others (positively or negatively) depends on their beliefs about the others' type . They

---

[10] Framing does not imply endogeneity. Framing makes behavior situationally dependent, but it is consistent with time-invariant situationally specific behavior: over time, one acts the same way in the same situation. By contrast, preferences are endogenous if one's experiences result in durable changes in behavior in given situations.

have norms about how much they should contribute; If they violate the norm they experience *guilt*. Finally, they experience *shame* if they violate their own norms and are sanctioned by others for this behavior. Each of these motives is subject to the framing and other static effects listed above; they may also change in the long run in response to the preference endogeneity to be studied in the next section

These motives (excepting spite) may induce team members to take more adequate account of the effects of their actions on fellow team members. The altruism and reciprocity of the members may lead them to value the payoffs of team members and thus to contribute more on their behalf. Reciprocity motives may induce a member to punish those contributing little to the team's output. Shame may enhance the effects of being punished by others. Finally, guilt may induce a more nerly socially optimal level of contribution.

Consider a team with two members, i and j. (The model is readily generalized to a team of n members.) The output of the team varies linearly with the contributions of the members, and it is divided such that each member receiving an amount $\phi < 1$ times the sum of the contributions. Each takes an action contributing a fraction of one unit of $a_k \in [0,1]$ for $k = i,j$ to the team and the remainder to a private project yielding benefits $(1- a_k)$.. After each has made an allocation, the contributions of each to the project are made known to the other, and i may impose a penalty μ_{ij} on j, while j may impose μ_{ji} on i, at a cost $c\mu^2/2$. Abstracting from the cost of one's own punishing of others for the moment, the material payoff to member i is thus

$$\pi_i = 1- a_i + \phi(a_i + a_j) - \mu_{ji}$$

To capture intrinsic motives suppose that contributing up to some level, $\underline{a}$, is intrinsically pleasurable independently of one's norms. But beyond that it is onerous, so that the utility associated with contributing level a is $\delta(a - \underline{a})^2$ where $\delta < 0$. To incorporate guilt and shame we say that each member suffers a psychic cost $\gamma(a^* - a)^2$ if his contribution deviates from his contribution norm (a^*) which is assumed to exceed $\underline{a}$. It may seem odd that the member experiences guilt in contributing too much, but contributing less than $1-a^*$ to the private project may violate a norm (the private project may be care of one's own children, for example.) Below I assume that members contribute less than their norm, but this is just a simplification to facilitate interpretation of the results. The weight β_{ij} ("benevolence") placed by i on j 's payoffs depends on both unconditional altruism (or spite) and reciprocity. In the spirit of the model of reciprocal preferences due to Levine (1998), member i's benevolence towards j is

$$\beta_{ij} = \alpha_i + \lambda_i(a_j - a_i^*)$$

where $\alpha_i \in [-1,+1]$ is i's unconditional spite or altruism and λ_i his degree of reciprocity $\in [0,1]$.

The level of reciprocal motivation therefore depends on the extent to which *j* has deviated from *i*'s contribution norm: if *j* has contributed to their joint project more than *i*'s norm, and $\lambda_i > 0$, then *i* experiences good will toward *j,* and positively values his payoffs. But if *j* and contributed less than a_i^* then *i* may experience malevolence toward *j* ($\beta_{ij} < 0$) and enhance his utility by paying to reduce *j*'s payoffs. I do not include in *i*'s valuation of *j*'s payoffs, the costs to *j* of punishing *i,* because it seems implausible that *i* will increase his contribution because he cares about *j* and realizes that *j* will have to bear the costs of punishing him if he (*i*) contributes too little.

Finally, to reflect the fact that shame is a social emotion, evoked by the contempt of ones associates as expressed by their willingness to incur costs to punish a behavior, shame is measured as

$$s_i = \sigma_i (a_i^* - a_i)\mu_{ji}$$

Thus σ is a measure of one's susceptibility to shame. Punishment by others thus inflicts both material costs and subjective costs, the total being $\mu_{ji}(1 + \sigma_i(a_i^* - a_j))$. If both members have the same contribution norm and abstracting from spite, it will not occur that a member who has exceeded his own norm will nonetheless be punished. To avoid this complication in the numerical case I consider below, I assume $a_i^* = a_j^*$, and α_i and α_j are both non-negative.

Combining the above terms we have the utility of the *i*th individual

$$u_i = \delta(a - \underline{a})^2 + \pi_i + \beta_{ij}\pi_j - c\mu_{ij}^2/2 - \gamma(a_i^* - a_i)^2 - \sigma_i(a_i^* - a_i)\mu_{ji}$$

Utility is thus the sum of the intrinsic value of contributing, the individual's own material payoffs (including the cost of being punished) plus the valuation of the others material payoffs minus the cost of punishing *j*, minus the subjective valuation of guilt and shame. An analogous function describes *j*'s utility. Note that *i* makes two choices; first choose a_i, then in light of what *j* has contributed, decide what if any punishment to direct at *j*.

If *j* is contributing an amount such that $\beta_{ij} = \alpha_i + \lambda_i (a_j - a_i^*) < 0$, member *i* will choose to punish *j*. This is because *i* benefits from reducing *j* 's payoffs and the marginal cost of punishing is zero when $\mu = 0$. The utility maximizing level of punishment is given by

$$c\mu_{ij} = -\beta_{ij} = -\{\alpha_i + \lambda_i(a_j - a_i^*)\}$$

This instructs *i* to choose the level of punishment that equates the marginal cost of punishment (the left hand side) to the marginal benefit of punishment, namely the valuation *i* places on reducing the payoff of *j* (as long as this level is positive, and choose zero punishment otherwise). Where punishment is positive, it is clearly increasing in λ and decreasing in α, as

one would expect.

We assume that *i* knows that the punishment by *j*, if positive, will be $\mu_{ij} = -\beta_{ij}/c$. Substituting this value into utility function, *i* will choose the level of contribution to satisfy:

$$2\delta(a-\underline{a}) -1 + \phi(1+\beta_{ij}) + \lambda_j/c + 2\gamma_i(a_i^* - a_i) + \sigma_i\{-\beta_{ji}/c + (a_i^* - a_i)\lambda_i/c\} = 0.$$

This condition requires that a_i be chosen so as to equate the marginal cost and benefits of contributing. The first term is the marginal intrinsic benefit or cost of the contribution, while the -1 is the opportunity cost of contributing less to the private project. The term $\phi(1+\beta_{ij})$ gives the increment both to *i*'s own material payoffs as well as *j*'s, the latter valued by *i*'s benevolence towards *j*, while λ_j/c is the marginal reduction in punishment occasioned by contributing more. The next term is marginal reduction in guilt, followed by a term giving the reduction in shame occasioned both by more closely approximating one's norm and by invoking less punishment. Member *j*'s utility maximization yields the analogous first order condition.

Recalling that $\beta_{ij} = \alpha_i + \lambda_i(a_j - a_i^*)$, for $\lambda_i > 0$, total differentiation of the first order condition reveals that $da_i/da_j > 0$, so *i*'s contribution increases with *j*'s contribution. It is also true that for $a_i^* > a_i$, $da_i/d\gamma_i > 0$ and $da_i/d\sigma_i > 0$, so an increase in guilt motives and the susceptibility to shame raise *i*'s contribution.

One can rearrange the first order condition to give a closed form expression for a_i as a function of a_j and the parameters introduced above. This is member *i*'s best response function. (It is cumbersome, and unnecessary, as the comparative statics are readily inferred from the first order conditions). The best response functions given by *i* and *j*'s first order conditions are shown in Figure 1. The dashed lines in the figure illustrate two comparative static effects: a downward shift in *j*'s best response function induced by an decrease in susceptibility to guilt, $\Delta\gamma_j$, and a leftward shift in *i*'s best response function induced by an decrease in *i*'s guilt, $\Delta\alpha_i$.

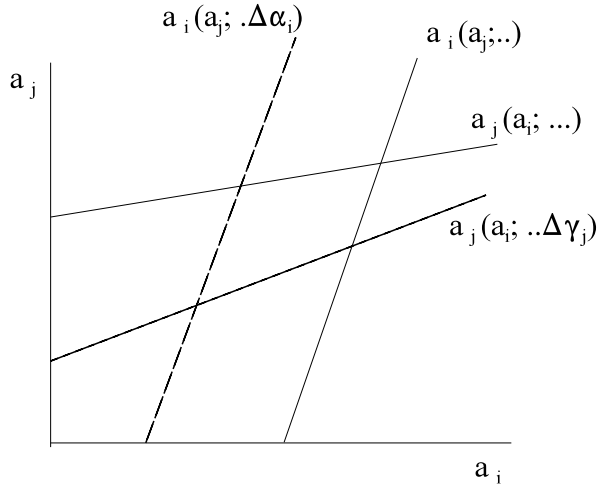

Figure 1. Equilibrium contributions to the team project, with social preferences. The dashed lines show the effects of decreased altruism by *i* and decreased guilt by *j*.

Were social motives absent, neither member would contribute (this is because the marginal material benefit is less than the marginal cost of contributing, as long as $\phi < 1$). But significant levels of reciprocity will induce members to punish their low-contributing mates, and this alone, or in combination with shame may support high levels of contribution. Alternately altruism or guilt can also support high levels of contribution. As the interaction is somewhat complex, it is a good idea to check that a plausible Nash equilibrium exists, and to illustrate how the model works. Assume *i* and *j* are identical, and dropping subscripts, suppose: $\phi = 0.6$, $\alpha = 0.0$, $a^* = 0.5$, $\lambda = 0.3$, $\gamma = 0.6$, $\sigma = 1.2$ and $c = 1$. Then $a^N = 0.5$, that is, members implement the common contribution norms, and as a result, they experience no shame or guilt and do not punish one another. As a result, both gain 0.1 in material benefits net of their contribution from the project (that is 0.6(0.5+0.5) - 0.5.)

Recall that lacking social preferences they would not have contributed at all, so the fact that *in equilibrium* they do not experience shame, guilt or benevolence toward the other does not imply that these motives are unimportant. As confirmation, consider the same two individuals in a disequilibrium state, at which *j* is contributing 0.4 and *i* is contributing only 0.1. By shirking, *i* captures 0.2 in net material benefits from the project (that is, 0.6(0.1 + 0.4) - 0.1). But *j* would as a result experience strong malevolence towards i ($\beta_{ji} < 0$) and so would punish *i* heavily, inflicting 0.06 in material costs and inducing additional subjective costs of 0.03 in shame on *i*. These, along with *i*'s subjective costs of guilt (0.09) would reduce *i*'s net benefits to virtually zero. In this situation, *i*'s best response is an increase in contribution.

There is no reason why social preferences will not be experienced in equilibrium (though it seems unlikely that high levels of shame, guilt and *mutual* punishment would be persistent.) To see how this might arise, suppose the two members held different contribution norms, with $a_j^* > a_i^*$. Both might adhere to their own norms when in equilibrium, and hence not experience guilt or shame. But at these equilibrium values, the fact that *i* was a shirker according to *j*'s norms might induce *j* to punish *i* and this punishment would be part of the incentives accounting for *i* meeting his own norm.

Why does additivity assumption fail in this model? How might the presence of other-regarding incentives influence the net effect of the introduction of self-regarding incentives in the form of subsides, taxes, fines? This depends on the nature of the incentive intervention, of course, but the above studies suggest a number of possibilities. An explicit incentive might increase the return on ones own contribution, raising φ. But explicit incentives may undermine intrinsic motives, reducing the level of contribution that is intrinsically pleasurable ($\underline{a}$). The cost of effective punishment may increase if the incentive structure supports more anonymous and competitive interactions, and improves the exit options. The valuation of others' material payoffs may be attenuated for two reasons. The incentives might suggest that a self-regarding rather than ethical mode of behavior is appropriate. Or they might degrade the available signals of the others' generosity and good will (an increase in φ means that a given

contribution signals less generosity). The frame shift may also lower the individuals contribution norm, thereby enhancing his valuation of the others payoffs (conditional on the others behavior and given parameters of the benevolence function). But for the same reason the frame shift also reduces guilt and shame, and may independently of its effect on α^* may also reduce the generic susceptibility of shame and guilt. For the numerical values used in the above computation of the Nash equilibrium, the latter (negative) effect greatly outweigns the former (positive) effect, as one would expect. Table 1 summarizes some possible effects, assuming as above that none of the individuals is contributing more than their own norm.

Table 1. Some possible effects of incentives on social preferences

Aspect of preferences	*Notation*	*Some possible effects of incentives*
own material payoffs	$1- a_i + \phi(a_i + a_j)$	Increased marginal net benefit of contribution
intrinsic motivation	$\delta(a - \underline{a})^2$	May reduce pre-existing intrinsic motivation
i) punishment by others (μ_{ij}): ii) cost of punishing others:	$-\beta_{ij}/c$ (if $\beta_{ij}<0$) $-\beta_{ij}^2/2c$ (if $\beta_{ij}<0$)	Increased marginal cost and reduced marginal benefit of punishment
value of others payoffs ($\beta_{ij}\ \pi_j$)	$\{\alpha_i + \lambda_i(a_j - a_i^*)\}\pi_j$	Less reciprocity and altruism toward others
guilt	$\gamma_i\ (a_i^* - a_i)^2$	Degraded norm, less generic guilt
punishment-induced shame	$\sigma_i\ (a_i^* - a_i)\mu_{ji}$	Less generic and punishment-induced shame

5. The coevolution of institutions and preferences

Non-additivity may arise for additional reasons not captured in Table 3. Incentives may also inhibit the evolution of social preferences that could contribute to good governance. Such dynamic preference endogeneity effects are suggested in Gaechter, Kessler, and Konigstein (2004), and Frohlich and Oppenheimer (1995) for example. But a dynamic version of crowding in may also occur: governmental interventions may also support the evolution of these social preferences.

The acquisition of new preferences may be modeled as a standard cultural evolution process in which individuals periodically update their behavioral norms (perhaps frequently, perhaps only during adolescence) after having taken into account information about the frequency distribution of various behavior in the population, the payoffs associated with various behaviors in recent periods, and other facts (Bowles (1998)). Equilibrium (that is, stationary) preferences will depend on the nature of the updating rules and the structure of social interactions given by the society's institutions. The latter are important as they determine

who meets who to do what tasks and with what benefits. Among the institutions making up this cultural environment are the structure of markets, contracts, legislation and other aspects of society affected by public policy. A consequence is that differing economic institutions support different equilibrium preferences.

This cultural evolution model provides a dynamic setting for the processes studied by Brown, Falk, and Fehr (2004). They designed a market experiment to explore the effects of contractual incompleteness on the pattern of trading. The good to be exchanged varied in quality, with higher quality more costly to provide. In the complete contracting condition, the level of quality promised by the supplier was enforced by the experimenter, while in the incomplete contracting condition the supplier could provide any level of quality (irrespective of any promise or agreement with the buyer). Buyers and sellers knew the identification numbers of those they were interacting with, so they could use information they had acquired in previous rounds as a guide to whom they would like to interact with, the prices and quality to offer, and the like. Buyers had the opportunity to make a private offer (rather than broadcasting a public offer) to the same seller in the next period, thus attempting to initiate an on-going relationship with the seller.

Very different patterns of trading emerged under the complete and incomplete contracting conditions. In the first, 90 percent of the trading relationships lasted less than three periods (most of them were single-shot). By contrast, only 40 percent of the relationships were this brief under the incomplete contracting condition and most traders formed trusting relationships with their partners. Buyers in the incomplete contracting condition offered prices considerably in excess the supplier's cost of providing quality. When Buyers were disappointed by the quality supplied, they terminated the relationship, thereby withdrawing the implied rent from the supplier. The differences in behavior under the two treatments were particularly pronounced in later rounds of the game, suggesting that the traders learned from their experiences, and updated their behaviors accordingly.

These results suggest that trust reciprocity may depend on the form of the contract, contractual incompleteness sometimes supporting of trusting and reciprocal behaviors. The converse is also true: expectations of lower levels of trust and reciprocity would plausibly lead those designing contracts to be willing to pay more for more complete contracts. The question then arises: how will changes in the legal and policy environment that more closely approximate the ideal of complete markets and costless contracting affect the evolution of social preferences? The model that follows suggests that the effects may be deleterious.

I model an interaction like that in Fehr, Klein, and Schmidt (2001) embedded in a dynamic cultural evolutionary environment. As in Bohnet, Frey, and Huck (2001), the underlying process jointly determines the distribution of contracts and the distribution of behavioral norms in the population, a dynamic sometimes termed the *co-evolution of*

institutions and preferences

Consider a population of buyer and sellers who are paired randomly for a single interaction. They trade a good whose quality (high (H) or low (L)) is determined by the seller and is costly for the buyer to determine *ex ante*. Buyers offer a contract, following which sellers determine the quality of the good they will provide. The buyer may offer one of two contracts. If the complete (C) contract is offered, the seller receives a fixed compensation just sufficient to offset the costs of providing low quality. These are C-type buyers. According to the incomplete (I) contract, the buyer pays the cost of producing low quality, plus half of the net profits resulting from the transaction. These are I-type buyers.

Sellers are also of two types. R-type sellers interpret the I-contract as a sign of trust on the part of the buyer, and reciprocate by providing high quality, incurring an additional cost of δ_H as a result. When offered a C-contract, however R-type sellers feel mistrusted, experiencing a subjective cost δ_C, and they retaliate, provide low quality. S-type sellers are completely self-regarding and provide low quality irrespective of the contract. The net surplus of the transaction (net of compensating the seller sufficient to offset the cost of low quality) are π^H and π^L for high and low quality respectively. Those offering a C-contract must pay a cost of μ for monitoring and contractual enforcement, while those offering an I-contract make themselves vulnerable to an expected loss of κ through theft should they interact with a S-type seller who gains an additional amount κ as a result.

seller → / buyer ↓	Reciprocator: (R)	Selfish(S)
Incomplete contract (I)	$\pi^H/2$, $\pi^H/2 - \delta_H$	$\pi^L/2 - \kappa$, $\pi^L/2 + \kappa$
Complete contract (C)	$\pi^L - \mu$, $-\delta_C$	$\pi^L - \mu$, 0

Table 2 Payoffs among reciprocal and self interested sellers exchanging a variable quality good with buyers offering complete and incomplete contracts.

To exclude cases in which only one of the pairs of contracts and preferences is viable, I further assume that $\pi^H > 2(\pi^L - \mu)$, $\pi^H - \pi^L > 2(\kappa+\delta_H)$ and $\pi^L > 2(\kappa - \mu)$. The payoffs appear in Table 2. From these assumptions we know that {I,R}, that is, the I-contract matched with the R-seller is the joint surplus maximizing outcome (and is also Pareto efficient). Thus buyers will offer I contracts if there are sufficiently many R-sellers in the population. But that does not guarantee that {I,R} will be observed in practice in a dynamic setting. Writing the fraction of the sellers who are reciprocators as ω, the expected payoffs to buyers offering the I- and C-contracts are:

$$v^I = \omega\pi^H/2 + (1 - \omega)(\pi^L/2 - \kappa)$$

$$v^C = \omega(\pi^L - \mu) + (1-\omega)(\pi^L - \mu) = \pi^L - \mu$$

Similarly, writing the fraction of the buyers offering incomplete contracts as ϕ, the expected payoffs to the R- and S-sellers are

$$v^R = \phi(\pi^H/2 - \delta_H) + (1-\phi)(-\delta_C)$$

$$v^S = \phi(\pi^L/2 + \kappa)$$

What kinds of contracts and behaviors would we expect to observe in this population. One's intuition is that likely outcomes would include a high frequency of both incomplete contracts and reciprocating sellers or the opposite: a predominance of both complete contracts and self interested sellers. These correct intuitions are readily formalized. The dynamical system we want to study concerns the state space defined by all possible combinations of contractual and behavioral strategies or $\phi \in [0,1]$ and $\omega \in [0,1]$. We wish to explore the movement of both ϕ and ω over time. Suppose that both suppliers and buyers periodically update their strategies by switching to strategies with higher payoffs. It is easily shown that this process gives the familiar replicator dynamic equations

$$d\phi/dt = \phi(1-\phi)(v^I - v^C)$$

$$d\omega/dt = \omega(1-\omega)(v^R - v^S)$$

The stationary values of ϕ and ω in this dynamic are:

$d\phi/dt = 0$ for $\phi = 0$, $\phi = 1$ and $\omega = \omega^* = (\pi^L + 2(\kappa-\mu)/(\pi^H - \pi^L + 2\kappa)$, and

$d\omega/dt = 0$ for $\omega = 0$, $\omega = 1$ and $\varphi = \varphi^* = 2\delta/(\pi^H - \pi^L - 2\kappa)$

The resulting dynamical system is illustrated in Figure 3, with the arrows indicating the out of equilibrium adjustment given by the equations immediately above. The point (ϕ^*, ω^*) is stationary but it is a saddle: small movements away from ϕ^* or ω^* are not self-correcting. The asymptotically stable states are {I,R} and {C,S}, confirming the above intuition.

In this deterministic setting, the initial state determines which of these two asymptotically stable states occurs. But a more realistic dynamic would include stochastic influences on payoffs and occasional idiosyncratic updating of preferences (that is, acquiring lower rather than higher payoff preferences). In a plausible version of this process (e.g. Young (1998), Naidu and Bowles (2004)), both {I,R} and {C,S} will occur, but over a very long time period the system would

spend more (and depending on the technical details, possibly virtually all) of the time in the one with the larger basin of attraction.

Public policy may affect the long run outcome, therefore, by altering the two critical values ϕ^* and ω^*. Consider two possibilities. First, the government imposes the rule of law, reducing theft and setting $\kappa=0$. Because $d\varphi^*/d\kappa > 0$ and $d\omega^*/d\kappa > 0$ the effect will be to lower the critical fractions of I-types and R-types required to propel the population the surplus maximizing outcome. This is a case of crowding in: an institutional innovation (the rule of law) generates a cultural environment in which reciprocal preferences and hence efficient incomplete contracts may proliferate. Suppose, as a second example, that judicial reforms reducing the cost contract enforcement lowered μ. Because $d\omega^*/d\mu < 0$ and $d\varphi^*/d\mu = 0$, the effect of this improvement in the contractual enforcement environment is to increase the critical fraction of suppliers who are R-types necessary to propel dynamic to the efficient equilibrium. In this case crowding out has occurred.

6. Conclusion. Implementation theory in light of behavioral economics

The above evidence and reasoning does not recommend abandoning Hume's objective

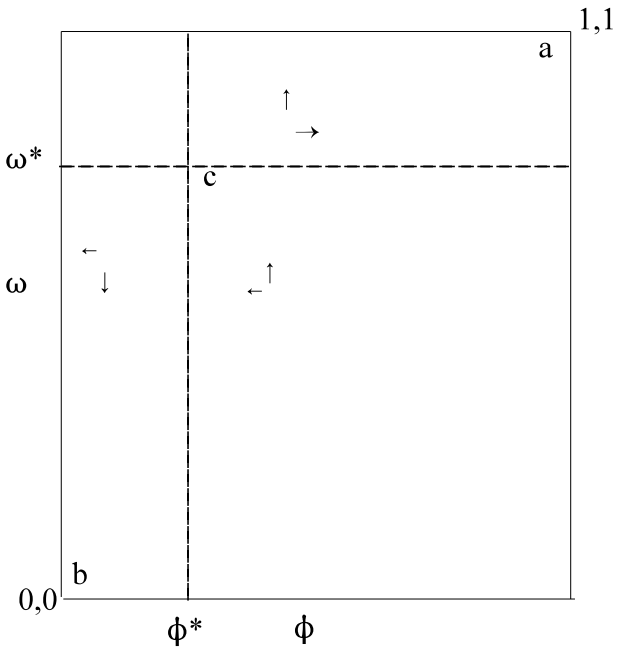

Figure 3 The coevolution of contracts and behaviors. The arrows indicate the directions of change implied by the assumed dynamic. States a, b, and c are stationary; c is a saddle.

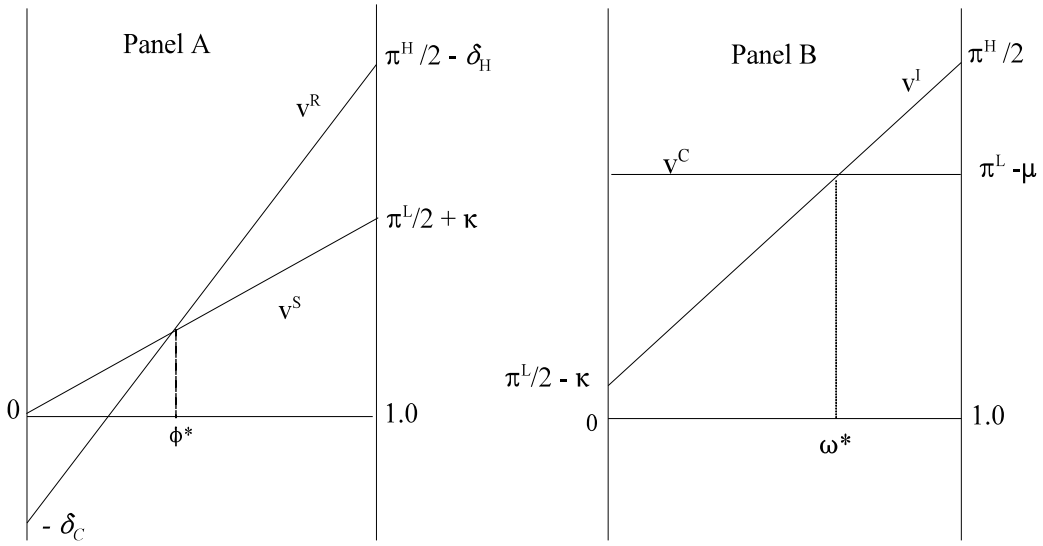

Figure 2.Payoffs to reciprocal and self interested behaviors (panel A) and incomplete and complete contracts. ϕ is the fraction of buyers offering complete contracts, ω is the fraction of suppliers who are reciprocators.

of harnessing self-regarding preferences to public ends: self interest remains a powerful motive and there is ample evidence that conventional incentive- based contracts and policies often work very well (Laffont and Matoussi (1995), Lazear (2000)). Rather two more modest recommendations follow.

First, the conventional economic prescription for perfecting market institutions – fewer impediments to mobility and competition – may not be the best way to harness self interest. Note that none of the historical examples of institutional crowding out (in section 2) hinges critically on the presence of social preferences (though mostly likely they were involved). Rather they illustrate a general principle about the relationship among institutions. Where contracts are incomplete, efficient exchange or other forms of mutually beneficial cooperation may occur if the structure of the interaction provides incentives that internalize the effect of one's actions on others. This may be done even with conventional self-regarding preferences, as is well known, if repeated interactions and small group size allow for effective retaliation against defectors (Fudenberg and Maskin (1986)), positive assortation by type (Bergstrom (2002), Axelrod and Hamilton (1981),Grafen (1979)), and for the establishment of reputations (Kreps (1990), Shapiro (1983)). These conditions facilitating cooperation among self-regarding individuals when contacts are incomplete are typically eroded by the policies seeking to create ideal-type Walrasian markets characterized by ephemeral contact among anonymous individuals. As a result, barring complete contracting, there is likely to be a tension between the allocative gains to be had by eliminating 'market imperfections' in favor of 'flexibility' on the one hand and on the other the informal contractual enforcement benefits of the mechanisms that allow mutually

beneficial interaction in the absence of complete contracts.[11]

Second, social preferences are a fragile resource for the policy maker, one that may be either empowered by legislation and public policy, or irreversibly diminished.. This suggests an extension of Hume's maxim: Good policies and constitutions are those that support socially valued ends not only by harnessing selfish preferences, but also by evoking, cultivating and empowering public-spirited motives. This will be particularly important where contracts are incomplete; for it is in these cases that as Arrow (1971):22 put it: "norms of social behavior, including ethical and moral codes (may) ...compensate for market failures." Where this is the case, conventional incentive-based interventions may be worse than ineffective, motivating a norm-related analogue to the second best theorem due to Lipsey and Lancaster (1956-1957): *where contracts are incomplete (and hence norms may be important in attenuating market failures), public policies and legal practices that more closely approximate idealized complete contracting may exacerbate the underlying market failure (by undermining socially valuable norms such as trust or reciprocity) and may result in a less efficient equilibrium allocation*. A constitution for knaves, Bruno Frey (1997) observed, may produce knaves, just as Michael Taylor (1976) had earlier suggested that the Hobbesian state may produce Hobbesian man.

In seeking to implement a socially desired outcome, one must check that the preferences necessary to implement the outcome are sustainable under the policies, contracts or rules used in the implementation. The problem is more difficult than Hume suggested, not only because preferences are endogenous, but also because populations are heterogeneous and individuals are versatile. The problem, then is not to find a way to induce a homogeneous population of self-regarding individuals to implement a socially desirable outcome. Rather, it is to devise rules such that in cases where cooperation is socially desirable, individuals with other-regarding preferences will have opportunities to express their pro-sociality in ways that induce all or most to cooperate, as in the public goods with punishment experiments. And as Adam Smith stressed, in situations where competition rather than cooperation is essential to socially valued outcomes, the task is exactly the opposite. He hoped that under competitive conditions self-interest would serve as a solvent of the solidaristic motives that sustained cartels and other 'conspiracies against the public.'

For instrumental reasons, then a policy maker adopting this approach would not be indifferent to the preferences that the rules empowered and promoted. But once the effect of policies on preferences is recognized for practical reasons it will also arise as an ethical issue. The reason is that the philosophical bedrock of liberalism is, *contra* Aristotle, that one ought not to pass judgements on people's preferences, and (*a fortiori*) one ought not to seek to change them in the public interest.

---

[11] Platteau (1994) makes a related point, based on anthropological data from Africa. See also Ben-Porath (1980)

Table A1: Institutional Crowding

Citation	*Group studied*	*Interaction*	*Results*	*Comment*
Somanathan (1991)	Uttar Pradesh forest users (India)	Early 20th century colonial intervention	..undermined village-level management	See also Thompson, Feeny, and Oakerson (1986) on Sahel forests
Jodha (1990)	Common property resource users in India	Access to markets and governance of CPRs.	...raised the 'transactions costs of enforcing social disciplne' (A72)	See also Arnold and Campbell (1986):438
Mallon (1983)	Early 20th c central highland Peru	Communal labor contributions to community projects	...became unenforceable due to labor mobility	
Lanjouw and Stern (1998):570	Palanpur, Uttar Pradesh	Enforcement of credit contracts	Labor mobility undermines enforcement	Mobility reduced value of reputations
Andre and Platteau (1997):32	Rural Rwandans	Customary obligations to share with needy kin	...do not apply to land acquired by purchase	
Frey and Oberholzer-Gee (1997)	Swiss citizens	Local nuclear reactor site (compensated /not)	Compensation reduced acceptability of location by half	Payment possibly a signal of degree of risk
Upton (1974)	U.S blood donors	Paid donations or uncompensated	Highly motivated givers respond negatively to incentives	Substantiates Titmuss (1971) See: Bliss (1972), Arrow (1972)
Greif (1994)	Maghribi and Genovese traders	Collectivist contract enforcement	Long distance trade eclipsed 'collectivist' enforcement	
Greif (2002)	Early modern European traders	Self policing of communities of traders to protect joint reputation	Success of 'community responsibility' contract enforcement in promoting impersonal exchange	increased anonymity, size, and heterogeneity of merchant communities raised enforcement costs
Gneezy and Rustichini (2000a)	Haifa daycare parents	Fine imposed for lateness	...increased lateness which persisted after fine was withdrawn	fine signaled 'how bad' lateness was, shifted 'from a communal to an exchange' relationship

Gneezy and Rustichini (2000b)	Israeli students	Payment for soliciting contributions to social causes	Payment may reduce the performance of the solicitors	
Garcia-Barrios and Garcia-Barrios (1990)	Rural Mexicans	Maintenance of irrigation systems	Erosion of patron client relationship by modern contracts degraded maintenance	See also Jodha (1990)
Sengupta (2001)*	Farmers in Magarh, Bihar	Interaction of colonial state and *Ahar-pyne* irrigation	Degradation of community-based irrigation systems	
Bohannan (1959)	Tiv community (Nigeria)	Introduction of money in a system of three unlinked exchange processes	Linking via money eroded multi-centric institutions and their moral distinctions	'Money ...creates its own revolution." facilitated comparability of once exclusive social spheres
Woodruff (1998)	Producers and buyers of shoes in Mexico 1990s	Enforcement of incomplete contracts between sellers and buyers	Trade liberalization provided low cost exit options for buyers	Destroyed existing infor-mal enforcement system based on joint retaliation against cheaters
Schmitz (1999)	Italian shoe producers, Sinos Valley, Brazil (1960s)	Pooling of financial resources, tools, and information	Introduction of an auction system by export agents to allocate orders	Heightened competition and inequality among firms eroded inter firm cooperation
Gurven, Hill, and Kaplan (2002)	Ache community (Paraguay)	Risk reduction through food sharing	...does not apply to money or to purchased goods	Increased role of market exchange degrades co-insurance system

Table A2: Explicit Incentives and Social Preferences: Experiments

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Fehr and Gaechter (2002)	Swiss students	Gift Exchange (GE)	Explicit incentives reduce effort (especially if negative), redistribute surplus to principal.	Framing and IA effects Incentives eliminate the positive effects of generosity (31)
Gneezy (2003)	U.S students	Proposer-Responder	W-curve: Non-monotonic effects of fines and rewards.	Discontinuity at zero reflects shift from moral to a strategic mode? See Gneezy and Rustichini (2000b)
Fehr and List (2004)	CEO's & students (Costa Rica)	Trust Game (TG) with optional punishment	Not using the punishment option when it is available results in high performance	Key: "the psychological message .. conveyed by incentives – whether ... kind or hostile..."
Bohnet, Frey, and Huck (2001)	U.S. students	Contract enforcement	Compliance is non-monotonic in degree of enforcement	"Monetary" crowd out "Honest" preferences where enforcement is moderately strong
Fehr and Rockenbach (2003)	German students	Trust Game with optional punishment	Not using the punishment option when it is available results in high performance	Forgoing the punishment option is a signal of good will and trust
Cardenas, Stranlund, and Willis (2000)	Colombian rural poor	Common pool resource with fines	Fines induce more self-interested behavior & pool over exploitation	Fine induced a shift from moral to self interested frame ?
Schotter, Weiss, and Zapater (1996)	U.S. students	Ultimatum and Dictator Games	competitive threats to survival induced lower offers	"..[market] offers justifications for actions that in isolation would be unjustifiable" p.38
Fehr, Gachter, and Kirchsteiger (1997)*	Swiss students	Gift Exchange (effort non-contractible)	Monitoring and fines reduced effort	
Frohlich and Oppenheimer (1995)*		Prisoners' Dilemma (PD)	Incentive compatible option reduced performance in subsequent play	IC option 'undermines ethical reasoning and ethically motivated behavior.' p.44

Falk and Kosfeld (2005)	Swiss students	Trust Game	Ps who impose a minimum return rate on trustees receive less than trusting Ps	imposed minimum understood by S's as a sign of distrust by Ps
Hauser, Xiao, McCabe, and Smith (2004)	U.S. students	Trust Game	Weak sanctions by Truster or by Nature induce less 'trustworthiness' .	"Extrinsic incentives ...can ...change subjects' frame from ethical to income-maximizing."
Fischbacher, Fong, and Fehr (2005)	Swiss students	"Bargaining" vs "Market" Ultimatum Game	Competition among respondents reduced rejections	Competition made punishment of 'unfair' offers less certain
Bohnet and Baytelman (2005)	senior executives in U.S.	TG: one shot, repeated, w/o & w punishment, communication ("institutions")	institutionals increase amount sent and (conditional on that) returned; option of punishment reduces offers of other-regarding trustees	"punishment [option] destroys intrinsic trust and...controlling for expectations of trust, lowers..willingness to reward trust"
Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis (2005)	Members of 15 simple societies	Ultimatum Game	Offers and rejection of low offers were greater in more market-integrated societies	"*doux commerce*"?
Fehr, Klein, and Schmidt (2001)	German students	Gift exchange with piece rate and incomplete contracts	Incomplete (bonus) contracts yield higher returns to both P and A and are more common.	'existence of fairminded As may [explain] why many contracts are ...left incomplete'
Galbiati and Vertova (2005)	Italian students	Public goods game with rewards and penalties	stated contribution norm raises contributions independently of self-regarding incentives.	Contributions respond to socially determined 'obligations' (crowding in)
Tyran and Feld (2004)	Swiss students	Public goods with mild and strong sanctions	'compliance is much improved if mild law is endogenously chosen i.e. self imposed'	self imposed punishment does not indicate hostile intent
Gaechter, Kessler, and Konigstein (2004)	Swiss students	Gift exchange with fine, bonus, and trust	Cooperation is reduced in rounds subsequent to an incentive treatment; larger effect for fine than bonus	"Irreversibility: .. Incentives have a lasting negative effect on voluntary cooperation"
Hoffman, McCabe, Shachat, and Smith (1994)	U.S. students	Ultimatum game	Market 'labels' (Exchange game) reduced offers and raised acceptance levels	Market framing induces self-regarding preferences

Irlenbusch and Sliwka (2004)	German students (Erfurt)	Gift exchange (wage-effort) with piece rate option	Piece rates lower effort when they are in force, and after they are abandoned.	"..incentive [suggests] an individual maximization frame rather than a cooperative frame"
Rustrom (2002)	U.S. students	Creative task ('tower of Hanoi') with large, small and no penalties and rewards	Penalties degraded performance; large rewards induced better performance than small (but no better than the no-incentive treatment)	Penalties 'distracted' S's
Tenbrunsel and Messick (1999)	U.S. students	social dilemma with weak, strong and no sanctions	Ss evaluated sanction treatment as 'business' rather than 'ethical' Weak sanctions decreased expectations others would cooperate.	Weak (strong) sanctions reduce (increase) cooperation; no effect of sanctions for those adopting an ethical frame
Gaechter and Falk (2002)	Austrian students	One shot and repeated gift exchange game	Reciprocity stronger in repeated game; repetition induces selfish agents to imitate reciprocators	Repetition does not reduce reciprocal motives and "crowds in" 'imitated' reciprocity
Bowles, Carpenter, and Gintis (2001)	U.S. students	Public goods with punishment	Peer punishment induced defectors to contribute more, even when defection remained a best response	Punishment activated guilt, crowding in shame induced cooperation.

Note: P is principal, IA is inequality aversion, S is subject. * indicates that my research on this case is not complete.

Works cited

Acheson, James. 1988. *The Lobster Gangs of Maine*. Hanover, N.H.: New England Universities Press.

Anderhub, Vital, Simon Gaechter, and Manfred Konigstein. 2000. "Efficient Contracting and Fair lay in a Simple Principal Agent Experiment." *Institute for Empirical Research in Economics*.

Andre, Catherine and Jean-Philippe Platteau. 1997. "Land Relations Under Unbearable Stress: Rwanda Caught in the Malthusian Trap." *Journal of Economic Behaviour and Organization*, 34:1, pp. 1-55.

Andreoni, James, Brian Erand, and Jonathan Feinstein. 1998. "Tax Compliance." *Journal of Economic Literature*, 36:2, pp. 818-60.

Aoki, Masahiko. 2001. *Towards a comparative institutional analysis*. Cambridge: MIT Press.

Aristotle. 1962. *Nicomachean ethics*. Indianapolis: Bobbs-Merrill.

Arnold, J.E.M. and J.Gabriel Campbell. 1986. "Collective Management of Hill Forests in Nepal," in *Proceedings of the Conference on Common Property Managment*. National Research Council ed. Washington: National Academy Press, pp. 425-55.

Arrow, Kenneth J. 1971. "Political and Economic Evaluation of Social Effects and Externalities," in *Frontiers of Quantitative Economics*. M. D. Intriligator ed. Amsterdam: North Holland, pp. 3-23.

Arrow, Kenneth J. 1972. "Gifts and Exchanges." *Philosophy and Public Affairs*, 1:4, pp. 343-62.

Axelrod, Robert and William D. Hamilton. 1981. "The Evolution of Cooperation." *Science*, 211, pp. 1390-96.

Baland, J.M. and J.P. Platteau. 1996. *Halting Degradation of Natural Resources - Is there a Role for Rural Communities?* Oxford: Clarendon Press.

Barr, Abbigail. 2001. "Social dilemmas, shame-based sanctions, and shamelessness: experimental results from rural Zimbabwe." Centre for the Study of African Economies Working Paper WPS/2001.11: Oxford University.

Ben-Porath, Yoram. 1980. "The F-Connection: Families, Friends, and Firms and the Organization of Exchange." *Population and Development Review*, 6:1, pp. 1-30.

Bentham, Jeremy. 1970 [1789]. *An Introduction to the Principles of Morals and Legislation*: Athlone Press.

Bergstrom, Ted. 2002. "Evolution of social behavior: individual and group selection." *Journal of Economic Perspectives*, 16:2, pp. 67-88.

Bliss, Christopher J. 1972. "Review of R.M. Titmuss, The Gift Relationship: from human blood to social policy." *Journal of Public Economics*, 1, pp. 162-65.

Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior & Human Decision Processes*, 63:2, pp. 131-44.

Bohannan, Paul. 1959. "The Impact of Money on an African Subtistance Economy." *Journal of Economic History*, 19:4, pp. 491-503.

Bohnet, Iris and Yael Baytelman. 2005. "Institutions and Trust: Implications for Preferences, Beliefs and Behavior."

Bohnet, Iris, Bruno Frey, and Steffen Huck. 2001. "More Order with Less Law: On Contractual Enforcement, Trust, and Crowding." *American Political Science Review*, 95:1, pp. 131-44.

Bolton, Gary E. and Elena Katok. 1998. "An experimental test of the crowding out hypothesis: The nature of beneficent behavior." *Journal of Economic Behavior & Organization*, 37, pp. 315-31.

Bowles, Samuel. 1989. "Mandeville's Mistake: Markets and the Evolution of Cooperation." *Presented to the September Seminar, London.*

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature*, 36:1, pp. 75-111.

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Bowles, Samuel, Jeffrey Carpenter, and Herbert Gintis. 2001. "Mutual Monitoring in Teams: The Importance of Shame and Punishment." *University of Massachusetts.*

Bowles, Samuel and Yong-Jin Park. 2005. "Inequality, Emulation, and Work Hours: Was Veblen Right?" *Economic Journal*, (forthcoming).

Breer, Paul E. and Edwin A. Locke. 1965. *Task experience as a source of attitudes*. Homewood, Ill.,: Dorsey Press.

Brown, Martin, Armin Falk, and Ernst Fehr. 2004. "Relational Contracts and the Nature of Market Interactions." *Econometrica*.

Burke, Edmund. 1955 [1790]. *Reflections on the Revolution in France*. Chicago,: Gateway Editions; distributed by H. Regnery Co.

Camerer, Colin. 2003. *Behavioral Game Theory: Experimental Studies of Strategic Interaction*. Princeton: Princeton University Press.

Camerer, Colin and Ernst Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists," in *Cooperation, Punishment and Self-Interest: Experimental and Ethnographic Evidence from 15 Small Scale societies*. Joe Henrich, Samuel Bowles, Robert Boyd, Colin Camerer, Ernst Fehr and Herbert Gintis eds. Oxford: Oxford University Press.

Cameron, J, K Banko, and W. David Pierce. 2001. "Pervasive negative effects of rewards on intrinsic motivation: The myth continues." *Behavior Analyst, Special Issue*, 24:1, pp. 1-44.

Cardenas, Juan Camilo, John K. Stranlund, and Cleve E. Willis. 2000. "Local Environmental Control and Institutional Crowding-out." *World Development*, 28:10, pp. 1719-33.

Cooter, Robert. 1998. "Expressive Law and Economics." *Journal of Legal Studies*, 27, pp. 585-608.

de Tocqueville, Alexis. 1958 [1830]. *Democracy in America, Volume II*. New York NY: Vintage.

Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin*, 125:6, pp. 627-68.

Eisenberger, R and J Cameron. 1996. "Detrimental effects of reward: reality or myth." *American Psychologist*, 51, pp. 1153-66.

Falk, Armin and Michael Kosfeld. 2005. "Distrust: the hidden cost of incentives." *University of Bonn*.

Fehr, Ernst and Urs Fischbacher. 2004. "Third party punishment and social norms." Institute for Empirical Research in Economics, University of Zurich.

Fehr, Ernst and Simon Gachter. 2000. "Cooperation and Punishment in Public Goods Games." *American Economic Review*, 90:4, pp. 980-94.

Fehr, Ernst, Simon Gachter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65:4, pp. 833-60.

Fehr, Ernst and Simon Gaechter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14:3, pp. 159-81.

Fehr, Ernst and Simon Gaechter. 2002. "Do Incentive Contracts Crowd Out Voluntary Cooperation?" University of Zurich.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2001. "Fairness, Incentives and

Contractual Incompleteness." *CESifo and CEPR*.

Fehr, Ernst and John List. 2004. "The hidden costs and returns of incentives: Trust and trustworthiness among CEOs." *Journal of The European Economic Association*, 2:5, pp. 743-71.

Fehr, Ernst and Bettina Rockenbach. 2003. "Detrimental effects of sanctions on human altruism." *Nature*, 422:13 March, pp. 137-40.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114:3, pp. 817-68.

Fischbacher, Uris, Christina Fong, and Ernst Fehr. 2005. "Fairness, errors, and the power of competition."

Frey, B. and R Jegen. 2003. "Motivation Crowding Theory: A Survey of Empriical Evidence." *Journal of Economic Surveys*, 15:5, pp. 589 - 611.

Frey, Bruno S. 1997. "A Constitution for Knaves Crowds Out Civic Virtues." *Economic Journal*, 107`:443, pp. 1043-53.

Frey, Bruno S. and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *American Economic Review*, 87:4, pp. 746-55.

Frohlich, Norman and Joe A. Oppenheimer. 1995. "The Incompatibility of Incentive Compatible Devices and Ethical Behavior: Some Experimental Results and Insights." *Public Choice Studies*, 25, pp. 24-51.

Fudenberg, Drew and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica*, 54:3, pp. 533-54.

Gaechter, Simon and Armin Falk. 2002. "Reputation or Reciprocity? Consequences for Labour Relation." *Scandinavian Journal of Economics*, 104:1, pp. 1 - 26.

Gaechter, Simon, Esther Kessler, and Manfred Konigstein. 2004. "Performance Incentives and the Dynamics of Voluntary Cooperation."

Galbiati, Roberto and Pietro Vertova. 2005. "Law and Behavior in Social Dilemmas." *University of Siena*.

Garcia-Barrios, Raul and Luis Garcia-Barrios. 1990. "Environmental and Technological Degradation in Peasant Agriculture: A Consequence of Development in Mexico." *World Development*, 18:11, pp. 1569-85.

Gauthier, David. 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Gneezy, Uri. 2003. "The W effect of incentives." University of Chicago Graduate School of Business.

Gneezy, Uri and Aldo Rustichini. 2000a. "A Fine is a Price." *Journal of Legal Studies*, 29:1, pp. 1-17.

Gneezy, Uri and Aldo Rustichini. 2000b. "Pay enough or don't pay at all." *Quarterly Journal of Economics*, 115:2, pp. 791-810.

Grafen, Alan. 1979. "The Hawk-Dove Game Played between Relatives." *Animal Behavior*, 27:3, pp. 905-07.

Greif, Avner. 1994. "Cultural Beliefs and the Organization of Society: An Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, 102:5, pp. 912-50.

Greif, Avner. 2002. "Institutions & Impersonal Exchange: From Communal to Individual Responsibility." *Journal of Institutional and Theoretical Economics*, 158:1, pp. 168-204.

Gurven, Michael, Kim Hill, and Hillard Kaplan. 2002. "From forest to reservation: Transitions in food sharing among the Ache of Paraguay." *Journal of Anthropological Research*, 58, pp. 93-120.

Harrison, Ross. 1983. *Bentham*. London: Routledge & Kegan Paul.

Hauser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2004. "When punishment fails: Research on sanctions, intentions, and non cooperation." *George Mason University*.

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, (in press).

Henrich, Joe, Robert Boyd, Samuel Bowles, Ernst Fehr, and Herbert Gintis eds. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence in 15 Small-scale Societies*. Oxford: Oxford University Press.

Hirschman, Albert O. 1985. "Against parsimony:three ways of complicating some categories of economic discourse." *Economics and Philosophy*, 1:1, pp. 7-21.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7:3, pp. 346-80.

Hume, David. 1964. *David Hume, The Philosophical Works*. Darmstadt: Scientia Verlag Aalen.

Hurwicz, Leonid. 1975. "The Design of Mechanisms for Resource Allocation," in *Frontiers of Quantitative Economics II*. M. D. Intrilligator and D. A. Kendrick eds. Amsterdam: North Holland Press.

Irlenbusch, Bernd and Dirk Sliwka. 2004. "Incentives, decision frames, and motivation crowding out: an experimental investigation." *London School of Economics*.

Jodha, N.S. 1990. "Rural Common Property Resources: Contributions and Crisis." *Economic and Political Weekly*:June 30, pp. A65-A78.

Kahan, Dan M. 1997. "Social Influence, Social Meaning, and Deterrence." *Virginia Law Review* (Virginia Law Review), 83:2, pp. 349- 95.

Kahneman, Daniel, Jack L. Knetch, and Richard H. Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review*, 76:4, pp. 728-41.

Kahneman, Daniel and Amos Tversky. 2000. *Choices, Values and Frames*. Princeton: Princeton University Press.

Kreps, David M. 1990. "Corporate Culture and Economic Theory," in *Perspectives on Positive Political Economy*. James Alt and Kenneth Shepsle ed. Cambridge, UK: Cambridge University Press, pp. 90-143.

Laffont, Jean Jacques. 2000. *Incentives and Political Economy*. Oxford: Oxford University Press.

Laffont, Jean Jacques and Mohamed Salah Matoussi. 1995. "Moral Hazard, Financial Constraints, and Share Cropping in El Oulja." *Review of Economic Studies*, 62:3, pp. 381-99.

Lanjouw, Peter and Nicholas Stern eds. 1998. *Economic Development in Palanpur Over Five Decades*. Delhi: Oxford University Press.

Layard, Richard. 1980. "Human Satisfactions and Public Policy." *Economic Journal*, 90:360, pp. 737-50.

Lazear, Edward. 2000. "Performance Pay and Productivity." *American Economic Review*, 90:5, pp. 1346 - 61.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1:3, pp. 593-622.

Lipsey, R. and K. Lancaster. 1956-1957. "The General Theory of the Second Best." *Review of Economic Studies*, 24:1, pp. 11-32.

Mallon, Florencia E. 1983. *The Defense of Community in Peru's Central Highlands: Peasant Struggle and Capitalist Transition 1860 - 1940*. Princeton: Princeton University Press.

Mandeville, Bernard. 1924. *The Fable of the Bees, or Private Vices, Publick Benefits*. Oxford: Clarendon Press.

Marx, Karl. 1959 [1847]. "The Poverty of Philosophy," in *Basic Writings on Politics and Philosophy*. Lewis Feuer ed. Garden City: Doubleday.

Maskin, Eric. 1985. "The Theory of Implementation in Nash Equilibrium: A Survey," in *Social Goals and Social Organization; Essays in Memory of Elisha Pazner*. Leonid Hurwicz, David Schmeidler and Hugo Sonnenschein eds. Cambridge: Cambridge University Press, pp. 173-341.

Milgrom, Paul R. and John Roberts. 1990. "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities." *Econometrica*, 58:6, pp. 1255-77.

Mill, John Stuart. 1867[1848]. *Principles of Political Economy with Some of Their Applications*. London: Longmass, Green, Reader, and Diver.

Naidu, Suresh and Samuel Bowles. 2004. "Institutional equilibrium selection by intentional idiosyncratic play." Santa Fe Institute.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Ostrom, Elinor. 2000. "Crowding out Citizenship." *Scandinavian Political Studies*, 23:1, pp. 3-16.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review*, 86:2, pp. 404-17.

Platteau, Jean-Philippe. 1994. "Behind the Market Stage Where Real Societies Exist - Part II: The Role of Moral Norms." *Journal of Development Studies*, 30:4, pp. 753-817.

Pommerehne, W.W. and Hannelore Weck-Hannermann. 1996. "Tax rates, tax administration and income tax evasion in Switzerland." *Public Choice*, 88:1-2, pp. 161-70.

Robertson, William. 1769. *View of the Progress of Society in Europe*.

Rustrom, E.Elisabet. 2002. "Sparing the Rod Does not Spoil the Child: An Experimental Study of Incentive Effects." *Moore School of Business, University of SouthCarolina*.

Schmitz, Hubert. 1999. "From ascribed to earned trust in exporting clusters." *Journal of International Economics*, 48, pp. 138-50.

Schotter, Andrew, Avi Weiss, and Inigo Zapater. 1996. "Fairness and Survival in Ultimatum and Dictatorship Games." *Journal of Economic Behavior and Organization*, 31:1, pp. 37-56.

Sengupta, Nirmal. 2001. *A new institutional theory of production: an application*. New Delhi: Sage.

Sethi, Rajiv and E. Somanathan. 1996. "The Evolution of Social Norms in Common Property Resource Use." *American Economic Review*, 86:4, pp. 766-88.

Shapiro, Carl. 1983. "Premiums for High Quality Products as Returns to Reputations." *Quarterly Journal of Economics*, 98:4, pp. 659- 79.

Smith, Adam. 1937 [1776]. *An inquiry into the nature and causes of the wealth of nations*. New York: Random House.

Smith, Adam. 1976 [1759]. *Theory of Moral Sentiments*. Oxford: Clarendon Press.

Somanathan, E. 1991. "Deforestation, Property Rights and Incentives in Central Himalaya." *Economic and Political Weekly*, Vol. XXVI: 37-46.

Taylor, Michael. 1976. *Anarchy and Cooperation*. London: John Wiley and Sons.

Tenbrunsel, Ann and David M. Messick. 1999. "Sanctioning systems, decision frames and cooperation." *Administrative Science Quarterly*, 44, pp. 684-707.

Thompson, J.T., David Feeny, and R.J. Oakerson. 1986. "Institutional Dynamics: The Evolution and Dissolution of Common Property Resource Management." *Conference on Common Property Resource Management*: 391-424. U.S. National Academy of Science Press: Washington.

Titmuss, Richard M. 1971. *The Gift Relationship: From Human Blood to Social Policy*. New York: Pantheon Books.

Tyran, Jean-Robert and Lars Feld. 2004. "Achieving Compliance when Legal Sanctions are Non-deterrent."

Upton, William Edward III. 1974. "Altruism, attribution, and intrinsic motivation in the recruitment of blood donors." *Dissertation Abstracts International*, 34:12, pp. 6260-B.

Warr, P. 1982. "Pareto optimal redistribution and private charity." *Journal of Public Economics*, 19:131-138.

Woodruff, Christopher. 1998. "Contract enforcement and trade libealization in Mexico's footwear industry." *World Development*, 26:6, pp. 979-91.

Yamagishi, Toshio. 1988. "The Provision of a Sanctioning System in the United States and Japan." *Social Psychology Quarterly* (Social Psychology Quarterly), 51:3, pp. 265-71.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.