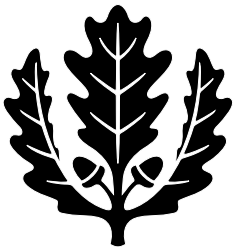


University of Connecticut

Department of Economics Working Paper Series

Lying, Integrity, and Cooperation

Lanse P. Minkler
University of Connecticut

Thomas J. Miceli
University of Connecticut

Working Paper 2002-36

August 2002

341 Mansfield Road, Unit 1063
Storrs, CT 06269–1063
Phone: (860) 486–3022
Fax: (860) 486–4463
http://www.econ.uconn.edu/

Abstract

While talk is cheap to some, it is expensive to others for whom moral considerations come into play. We employ a simple two-stage modified prisoner's dilemma game where integrity is endowed on a continuum to analyze when agents will lie in random economic interactions. If there is sufficient integrity in the population, all agents make a promise in the first stage to cooperate in the second. Some agents always lie, some always tell the truth, and some behave conditionally. Enhanced cooperation is a byproduct of integrity.

Journal of Economic Literature Classification: A13, D00

Lying, Integrity, and Cooperation

1. Introduction

Some people do not lie, even if they could get away with it. This seemingly innocuous fact poses important economic implications. Most fundamentally, it means that economic interactions do not always need to be mediated by binding contracts, which are sometimes difficult, if not impossible, to write, monitor and enforce. It means that sometimes sellers, workers, managers, and collaborators of all sorts can be counted on to perform even when nothing seems to assure that they will. For many economists such behavior is puzzling, especially if it is postulated that the only reason economic agents keep their word is to further their material interests, either directly or through reputation. In this account verbal representations are nothing but "cheap talk" – they are not credible. The problem is that casual empiricism and experimental evidence do not support this view. We address this apparent anomaly by providing an explanation for truth-telling that hinges on the existence of integrity; that is, integrity sometimes provides a reason not to lie, even when material incentives suggest the contrary. Verbal representations and promises invoke integrity.

The experimental literature is replete with cases of people cooperating in opposition to their material incentives.[1] Typically experimenters construct social dilemmas like a prisoner's dilemma or voluntary contribution to a public good where, based on instrumental rationality, the dominant strategy is to defect. But consistently great numbers do not defect, even where there is anonymity and the game is played only once.[2] In his meta-analysis of 37 different studies consisting of 130 distinct social dilemma experiments, Sally (1995, p. 62) calculates a mean cooperation rate of 47.4% for the entire pooled sample. Perhaps even more surprising is the result that cooperation rates tend to jump dramatically when subjects are first allowed to talk with one another, and then leave the experiment in isolation.[3] Sally (1995) estimates (non-credible) promises to cooperate elicited by experimenters increase cooperation by 12-30%, depending on the regression model.

---

[1] Why would people cooperate in social dilemmas (experimental or otherwise) in contraposition to their material interests and dominant strategies? Explanations include: reinterpreting the notion of rationality (Gauthier 1986; criticized by Heap and Varoufakis 1995, p.164), the existence of evolutionarily predisposed cooperators (Frank 1987), social learning in threat type games (Witt 1986), high costs to calculating population characteristics and/or optimal strategies (Guttman 1996), the existence of a "genetic bully" who punishes defectors (Sethi 1996), fairness considerations (Rabin 1993), "lumpy" moral preferences (Dowell, Goldfarb, and Griffith 1998), and preferences for inequity aversion (Fehr and Schmidt 1999).

[2] See, for instance, Marwell and Ames (1981); Schneider and Pommerehne (1981); Caporael, Dawes, Orbell, and van de Kragt (1989); Davis and Holt (1993); Frey and Bohnet (1995); and Ledyard (1995).

[3] After reviewing the available evidence, John Ledyard (1995), in his chapter on public goods in the *Handbook of Experimental Economics*, concludes that the evidence suggests that pre-play communication counts as a "strong effect" which increases cooperation. For instance, Dawes, McTavish, and Shaklee (1977) find that payoffs increased from 31% to 71% when *relevant* communication (i.e., representations on what subjects agree to) was allowed in one-shot, public good type experimental settings. Isaac and Walker (1988) find contribution rates of over 80% with communication in one-shot settings, and over 90% with communication in repeated games (Isaac and Walker 1991).

For economists promise-making shouldn't matter because representations against one's own material interests (e.g., "I will invest in the public good if you do the same") lack credibility. Part of the issue involves context. Cheap talk may indeed be meaningless in economic interactions amongst oligopolists, for instance. A low cost potential entrant should not worry about the blustering of a high cost incumbent when making an entry decision. Such a context may be one of mutual deceit, where it is known and acceptable to lie. But a problem arises in trying to import credibility as necessary to assure compliance in *all* economic interactions. We postulate the *sole* reliance on credibility is responsible for the divergence between theory and evidence. Compliant behavior in social dilemmas becomes easier to understand when its source is located in *either* credibility or integrity. When people talk to one another and make certain representations their integrity becomes a factor, even when credibility is not, and hence they may honor their agreements in the face of countervailing material incentives.

Accordingly, we ask two fundamental questions about lying. First, who would choose to make a promise? And second, how can we explain the simultaneous existence of those who always lie, sometimes lie, and never lie? Our explanation hinges on both the existence of integrity, differentially endowed, and also the expected cost of dealing with a low integrity partner. Hence, our explanation attributes both material and moral interests to economic agents.[1]

The rest of the paper is organized as follows. First, we review the philosophical literature on lying in order to gain insights on when it is morally acceptable to lie, and what might count as a reason not to lie. Many, if not most people entertain ethical considerations before making fundamental decisions, so it is important to understand the contributions of those who have thought deeply about the issue. Then, to develop the notion that integrity serves as a reason not to lie, we construct a simple game theoretic model to analyze when people will lie in random economic interactions. There are two stages to the game. In the first, agents must decide whether or not to make a promise to cooperate in the second stage. In the second, they decide to cooperate or defect. We assume that the level of integrity (from very low to very high) is uniformly distributed throughout a population. It turns out that there is a critical amount of integrity such that either nobody makes first period promises or everyone makes first period promises. In the latter case, those who tell the truth do so either because they possess sufficient integrity, or, if they are endowed with less integrity, because they believe there is sufficient chance that they will interact with a high integrity partner. Still, the chance exists that they will be lied to and suffer some material loss. Mutual promise-making is rational in our model and increases expected utility.

---

[1] Others who have attempted to add a moral component into economic behavior include Sen (1978), Etzioni (1986), Dowell, Goldfarb, and Griffith (1998), and Minkler (1999). For arguments why moral considerations should be included into economic analyses, see Griffith and Goldfarb (1991), and Hausman and McPherson (1993). Goldfarb and Griffith (1991) discuss the promise and problems of trying to describe moral acts as stemming from rules, constraints and preferences. Our account most resembles a preference one.

1.1 Philosophy of Lying

According to Bok (1995) a minimalist interpretation of morality acknowledges that all groups must work out basic values of (1) positive duties of care and reciprocity, (2) negative injunctions concerning lying, deceit and betrayal, and (3) norms for what counts as just. Lying is "any intentional deceptive message in the form of a statement" (Bok, 1978, p.15) and thus must be accounted for in any moral system.[1] Note that both intent and deception are necessary. A promise-breaker may not initially intend to break a promise, but end up doing just that if circumstances change. If a promisor intends to break the promise at the time of the promise, perhaps out of strategic motives, he is also a liar. Lying is the (un)ethical twin of violence because both coerce people to act against their will. It also breaks down societal trust by calling into question the reliability of promisors. Trust is necessary for human interaction; certainly it aids economic interactions because without it enforceable agreements and their prohibitive costs become universally required.

Does one ever have license to lie? According to the pre-eminent deontological philosopher Immanuel Kant, and some theologians like St. Augustine, the answer is unequivocally no. Most philosophers and theologians are not so unconditional. For instance Utilitarians argue that if by the act of lying one could bring about positive net consequences then one should do so (how one would be able to measure and assess a lie's full consequences is less certain). More concretely, one could ask certain questions. Should one lie to children or to the dying to comfort them (assuming a lie could do so)? Should one lie to liars? Should one lie to prevent some harms? And, should one lie to a spouse in order to make them feel good, especially if they recognize it as such and engage in mutual deception themselves? Bok (1978) addresses all of these questions and suggests that the answers depend on their justifications. She argues that justifications must be able to withstand public scrutiny by reasonable persons. On this account, lying to liars in order to exact revenge, as one would do in fairness or tit-for-tat strategies in game theory, would not be justified.[2] Lying to prevent harm may be. And lying in cases of mutual deceit for reasons of spousal affection or in games of fun like poker are indeed justified.

For our purposes, Bok's analysis aids in (1) differentiating between not lying for reasons of credibility versus integrity, (2) explaining why lying to liars is not generally justified even if strong psychological impulses render it attractive, (3) justifying why a lie may be permissible to prevent harm, and (4) delineating circumstances like mutual deceit in which lying is morally permissible.

---

[1] Interestingly, even though philosophers talk a lot about “Truth,” very few have explicitly developed a full analysis of lying. Sissela Bok appears to be the most prominent exception.

[2] For instance, Rabin (1993, p. 1281) writes: "If someone is being nice to you, fairness dictates that you be nice to him. If somebody is being mean to you, fairness allows-- and vindictiveness dictates-- that you be mean to him." This passage illustrates how certain conceptions of fairness may not be particularly moral.

The first insight on integrity versus credibility drives the theoretical analysis here. As noted, most economists focus solely on the latter. For instance, people may nurture their credibility in order to build up a reputational asset they can draw on to increase their material well-being. In contrast, someone keeping a promise for the reason of integrity does so out of moral commitment and in order to preserve their own sense of identity. Reputation may be a by-product, but it was not the cause of the behavior.

Since integrity is not commonly used in economics, it may be helpful to briefly elaborate. Integrity is often defined as adherence to honesty, which, in some ways, is sufficient for our purposes. Philosophers, in contrast, believe the idea to be more complex. For them integrity requires commitment to moral principles (like honesty), and also that these commitments when taken together provide an important source of one's own identity (see, for example, Williams in Smart and Williams 1973; McFall 1987). These commitments must be coherent, which means consistency between one's commitments and actions, and between the commitments themselves. If integrity is thought of as identity conferring commitments, then, as McFall argues, if one violates their own commitments they also render their personal identity incoherent, leading to a personal loss of the most fundamental kind. So, for some, lying would change the way a person views him or herself, and not for the better. This loss will be a fundamental aspect of our analysis.

The second point, on lying to liars, is important for our analysis because we assume those of *sufficient* integrity will not lie, even to known liars. Most economists investigating the conditions for cooperation assume that, at best, cooperators will cooperate with cooperators but will defect with defectors. In the context of lying that means that if player A knew their partner B lied last period, A would lie in the subsequent period to punish B –- the fairness or tit-for-tat strategies.[1] While our model generates this kind of behavior too, it does so for different reasons. Additionally, we also show that some keep their word regardless of their partner's integrity or behavior. We justify this behavior on the grounds that, first, there is a moral injunction against lying for the reason of revenge, and some people surely recognize this, and, secondly, because any lying lays inroads to one's own integrity. As Bok (1978, p. 127) suggests, to act otherwise makes one's own behavior the prisoner of others' character defects.

Still, lying may be justified on the grounds of preventing harm to oneself or others (point 3). For instance, a smuggler of Jews during WWII could appeal to Bok's public scrutiny by reasonable persons to justify lying to Nazis. Most dilemmas are not so extreme, and other options exist. Sometimes there is an opportunity to teach truth-telling. At others, it may be best to avoid interactions with suspected or known liars. For that reason, we include the possibility of exit in our model.

---

[1] Some evidence suggests that reciprocity alone is insufficient to induce cooperation. See Bohnet and Frey (1999).

Before introducing our model we briefly comment on the origin of integrity, since its existence is crucial for the analysis. Its origin could be located in evolutionary processes, cultural, emotional or norm guided behavior, or the philosophical reflection of a free will, or any combination.[1] We are ultimately agnostic on that question, but insist that wherever integrity comes from it does exist and that the reflective person recognizes that lying can diminish it, and as such it constitutes an entirely separable reason than credibility not to lie.

2. Model

We consider a two-stage model designed to capture the willingness of individuals to make and honor promises to cooperate in economic interactions. Promises could be either explicit or implicit, but in either case they are clearly understood by both parties. The context could be a public goods one, where each agrees to contribute but where non-contribution is the dominant strategy. Or it could represent a buyer-seller interaction where each party could individually benefit by substituting less costly actions for the ones agreed to (e.g., sellers provide low instead of high quality, and buyers pay late or an amount less than agreed). Or the context could be an employment one, where, before an employment relationship is consummated, a worker states that she will work hard, and an employer claims to provide a good working environment. In each case both parties understand that their representations are not binding (i.e., there is no material penalty to breaking them). Those whose representations are followed by cooperation are not liars. In contrast, those who defect are liars (e.g., if the worker shirks, or if the employer foregoes the expense of a good working environment).

Often, there is no moral injunction against defecting. For a variety of good reasons, some will choose not to finance a public good. Certainly most economists presume that workers will shirk without adequate controls (e.g., monitors, incentive contracts); the only judgment rendered on this behavior is that it's rational. A moral judgment can be rendered, however, in cases where someone has said they will contribute to a public good or will provide the agreed upon effort and

---

[1] While most scholars have ignored integrity, they have considered related factors. For a good evolutionary account of cooperation, see Caporael, Dawes, Orbell, and van de Kragt (1989). Boyd and Richerson (1990) augment the evolutionary account to include cultural influences. Elster (1989,1996) argues that emotions both sustain and trigger social norms, including ones that promote cooperation. He thinks it premature to locate the origins of emotional dispositions in evolutionary processes. We have yet to discover if emotions such as anger confer positive net benefits [Elster (1998)]. To the extent that preferences provide reasons to cooperate, some preferences might be endogenously influenced by social and economic institutions [see Bowles (1998]. Interestingly, Adam Smith (1976) recognized the role of religion in providing a reason to honor one's agreements:

> The idea that, however we may escape the observation of man or be placed above the reach of human punishment, yet we are always acting under the eye, and exposed to the punishment of God, ... is a motive capable of restraining the most headstrong passions, with those at least, who by constant reflection, have rendered it familiar to them. (p. 281)

then fails to do so. We assume that the moral requirements of keeping one's word are clear and that the context is not one of mutual deceit. Defecting is associated with lying, and a person may choose not to defect if lying diminishes his or her own integrity. So, by assuming that agents represent that they will cooperate, we add moral considerations to the usual prisoner's dilemma type interaction.

In the first stage of our model, agents randomly pair with one another and choose to *promise* to cooperate in stage two, or to remain *silent* (i.e., they do not make a promise). In the second stage, agents play a variation of the one-shot prisoner's dilemma game to be described below. Whether or not a player makes a promise in stage one is public knowledge before the second stage begins. Players have the option to *exit* a relationship prior to playing the stage-two game after observing whether or not their partner has made a promise to cooperate. If either party exits, both receive a fixed payoff and the game ends.[1] The exit option puts a lower bound on payoffs the players will accept in stage two and gives those who intend to keep their promise the option of not having to couple with a probable liar. The possibility of exit allows for other (moral) ways to deal with liars and captures the idea that most economic relationships are voluntary in the sense that agents can exit "at will" (absent a contractual obligation).

We assume that agent i's utility, U_i, includes wealth, W_i, and some disutility associated with loss of integrity from lying, x_i. That is

$$U_i = \begin{cases} W_i & \text{if cooperate} \\ W_i - x_i & \text{if defect.} \end{cases} \qquad (1)$$

x_i is the internal cost to diminishing one's own integrity from lying. Recall that credibility as a reason not to lie depends on responses of others. Integrity, in contrast, does not, so in our model the existence of x is not contingent on the behavior of one's partner: defecting players suffer a cost of x even if they play with a defector. This formulation draws on the notion that, in general, lying for revenge is not morally justified, and that otherwise lying to liars makes one a prisoner to other's character defects.

The payoffs from a standard prisoner's dilemma are given in Matrix 1, where the first entry in each cell is Player II's payoff and the second entry is Player I's payoff. In order to make the game a prisoner's dilemma, we assume that

---

[1] Technically, the exit option adds an intermediate stage to the game. Specifically, in stage one players either "promise" or "remain silent"; in stage two they either "exit" or "play"; and in stage three they play the modified prisoner's dilemma. The outcome of the overall game is unaffected by treating the exit stage informally.

$$c > a > b > d. \quad (2)$$

We further assume that $2a>c+d$, which ensures that mutual cooperation is the jointly optimal outcome (i.e., that it yields a higher total payoff than the outcome where one player defects and the other cooperates). Absent considerations of integrity, the equilibrium of this game is mutual defection, which yields each player a payoff of b.

		Player I: Cooperate	Player I: Defect
Player II	Cooperate	a,a	d,c
	Defect	c,d	b,b

Matrix 1.

Assume that x, the cost of lying, varies randomly across players. Specifically, suppose that x is uniformly distributed on the interval $[0,\theta]$, where $\theta>0$. Also, define the indicator function, y, to reflect an agent's stage one strategy:

$$y_j = \begin{cases} 1, & \text{if agent j promised to cooperate in stage 1,} \\ 0, & \text{if agent j was silent in stage 1.} \end{cases} \quad (3)$$

The payoffs from the modified stage-two game are given in Matrix 2.

	Player I: Cooperate	Player I: Defect

	Cooperate	a, a	d, c-$y_I x_I$
Player II			
	Defect	c-$y_{II} x_{II}$, d	b-$y_{II} x_{II}$, b-$y_I x_I$

Matrix 2.

As noted above, players know their partner's y from stage one, but they do not observe their partner's x (though they know its distribution). Note that if y=1 for at least one player in a given pair (i.e., at least one player promised to cooperate in stage one), the game is no longer a prisoner's dilemma and mutual defection is not necessarily the equilibrium.[1]

We seek a subgame perfect equilibrium of the overall game. We therefore begin by deriving the outcome of the stage-two subgames, and then consider the optimal stage-one strategies. Note that there are three possible pairings of players at the end of stage one: (P,P), (S,S), and (P,S), where P=promised to cooperate and S=remained silent (the pairings (P,S) and (S,P) are treated as indistinguishable). If neither player exits after observing his or her partner's stage-one strategy, the pair plays a one-shot, simultaneous-move version of the game in Matrix 2.

Before examining the stage-two subgames, consider the exit option. One plausible value for the payoff from exiting is the equilibrium payoff from the unmodified prisoner's dilemma. In that case, players would never remain in a partnership promising an expected return less than b. This establishes the outcome of the standard prisoner's dilemma as the next-best option of players, given the at-will nature of stage-one relationships.

Now consider the pairings in each stage-two subgame. The first pairing of two silent players, (S,S), implies that $y_I=y_{II}=0$. In that case, Matrix 2 reduces to Matrix 1, regardless of the player's x-values. Thus, both players choose the dominant strategy of defection and each receives a payoff of b. In this case, players are indifferent between playing the stage-two game and exiting prior to stage two. When indifferent, we assume they play the game. (One interpretation of exit is therefore that players revise their stage-one strategies so that both choose to be silent.)

Consider next the pairing of two players who promised to cooperate, (P,P). In this subgame, $y_I=y_{II}=1$, so the x's matter. Players know their own x, but not that of their partner. Thus, each player calculates the probability, p, that a randomly chosen partner will cooperate, given that both players have promised to do so. Given p, the expected payoff from cooperating is

$$U^c=pa+(1-p)d=p(a-d)+d \qquad (4)$$

while the expected payoff from defecting for a type-x player is

$$U^d=pc+(1-p)b-x=p(c-b)+b-x. \quad (5)$$

A type-x player cooperates if $U^c \geq U^d$, or if

$$x \geq p(c-a)+(1-p)(b-d) \equiv x^*(p) \quad (6)$$

where (b-d) and (c-a) are both positive from (2).

Note that (6) implies that players with $x<(c-a)$ will defect even if their partner is expected to cooperate (tell the truth) with certainty (p=1), while players with $x \geq (b-d)$ will cooperate even if their partner is expected to defect (lie) with certainty (p=0). If we assume that $(c-a) \leq (b-d)$, then the preceding players form two distinct groups. Specifically, players with $x<(c-a)$ defect for all p -- we call them "pure defectors," and those with $x>(b-d)$ cooperate for all p-- we call them "pure cooperators."[1] (Note that pure defectors always exist since $(c-a)>0$, while pure cooperators exist if $\theta>(b-d)$; we return to the latter below.) If (b-d) is strictly greater than (c-a), then there exists an intermediate group of players (those with $(c-a)<x<(b-d)$) whose strategy depends on the equilibrium value of p, which is yet to be determined. We refer to these as "conditional players." Figure 1 graphs $x^*(p)$ in (x,p) space for the case where b-d>c-a. Players above the $x^*(p)$ line cooperate and receive a payoff of U^c, while those below the line defect and receive a payoff of U^d.

The third subgame involves one player who promised to cooperate and one who remained silent, (P,S). If player I is the promisor, then $y_I=1$ and $y_{II}=0$ in Matrix 2. Consider first player II's strategy. For an arbitrary p, the condition for this player to defect is given by

$$p(c-a)+(1-p)(b-d)>0 \quad (7)$$

Since this condition holds for all p, defection is a dominant strategy for the silent player.

Now consider the optimal strategy of the promisor. Note first that, given (7), he rationally calculates that p=0 for his silent partner. From (4) and (5), the promisor will therefore cooperate if d>b-x, or if x>b-d. Thus, as shown above, only pure cooperators will cooperate against a known defector and will receive a payoff of d. But this player can do even better by exiting prior to the start of stage 2 because in that case he can receive a payoff of b>d. The

---

[1] Poundstone (1992: pp. 222-226) makes a similar point.

possibility of exit thus ensures that a pure cooperator (someone with a high cost of lying) will never be compelled to be a "sucker" for a known defector. In effect, the promisor switches his stage-one strategy from P to S—i.e., he retracts his promise to cooperate. Thus a (P,S) pairing will never proceed to the stage two subgame since the promisor will always opt out.

To this point, we have shown that only (S,S) and (P,P) pairs will (potentially) proceed to the stage two subgame. That is, players who remain silent will only be able to pair with other silent players, while promisors will only be willing to pair with other promisors. The next question is what strategy players will rationally adopt in stage one, given the feasible pairings and the payoffs from the stage two subgames.

Consider first a player who, if he promises to cooperate in stage one, will find it optimal to defect in stage two (i.e., $U^d>U^c$ given p). This type of player will prefer to remain silent in stage one and pair with another silent player if $b>U^d$, or if

$$x>p(c-b)\equiv \hat{x}(p) \qquad (8)$$

where c-b>0 from (1). The locus $\hat{x}(p)$ is graphed as the positively sloped line in Figure 1. Players above the line (those with a high cost of lying) prefer to remain silent in stage one and defect in stage two, while those below the line (those with a low cost of lying) prefer to promise to cooperate in stage one and defect in stage two.

Consider next a player who in stage two will honor his promise in stage one to cooperate (one for whom $U^c>U^d$ given p). Such a player prefers to remain silent in stage one and pair with another silent player if $b>U^c$, or if

$$p<\frac{b-d}{a-d}<1 \qquad (9)$$

As shown in Figure 1, this critical point occurs as the intersection of x*(p) and $\hat{x}(p)$. To the right of this locus, players prefer to promise to cooperate in stage one and honor the promise in stage two, while to the left they prefer to remain silent in stage one and defect in stage two. Note

---

[1] If (c-a)>(b-d) then players with (b-d)<x<(c-a) will defect if p=1 and cooperate if p=0. We do not consider this case.

that the choice here depends not on a player's type, but on the equilibrium probability of cooperation.

The preceding shows that *no players* will cooperate in stage two if $p<(b-d)/(a-d)$ because none find it optimal to promise to cooperate in stage one. Thus, there cannot exist an equilibrium of the overall game with a strictly positive p which is also less than (b-d)/(a-d). One possible equilibrium therefore involves $p^*=0$, with all players choosing to be silent in stage one and defecting in stage two. This, of course, is just the outcome of the standard prisoner's dilemma. The payoff to all players in this equilibrium is b. This case can be interpreted as one where the amount of integrity in the population is insufficient to sustain cooperation because no player finds it individually rational to promise to cooperate in stage one.

The question is whether there is an equilibrium with some cooperation. As we saw above, any such equilibrium requires that $p>(b-d)/(a-d)$ so that stage-two cooperators are willing to make promises in stage one. Also, the equilibrium p must equal the true fraction of cooperators in stage two so that expectations are fulfilled. From (6), this requires that

$$p=Pr(x\geq x^*(p)), \tag{10}$$

assuming that in this equilibrium all players promise to cooperate in stage one (which we will show to be the case). Since x is uniformly distributed on $[0,\theta]$, (10) becomes

$$p=\frac{\theta - x^*(p)}{\theta} \tag{11}$$

Substituting for $x^*(p)$ from (5) and solving for p yields

$$p^* = \frac{\theta - (b-d)}{\theta - (b-d) + (c-a)} \tag{12}$$

which is strictly positive if $\theta>(b-d)$, and also less than one.

Using (12), we can write the condition that $p^*>(b-d)/(a-d)$ as

$$\theta \geq \frac{(b-d)(c-b)}{(a-b)} \qquad (13)$$

where the right-hand side is positive by (2). Equation (13) sets a lower bound on θ for an equilibrium with cooperation to exist. Further, since (c-b)/(a-b)>1, (13) says that there must be a minimum number of "pure cooperators" in the population (i.e., those with b-d<x<θ) in order to sustain an equilibrium with cooperation. And since from (12) p* increases in θ, the rate of cooperation increases as more pure cooperators are added to the population.

When (13) holds, it is optimal for *all* players to make a promise to cooperate in stage one, and then to play the stage-two subgame with a randomly chosen partner rather than to exit. In the subgame, a fraction p* of players cooperate (those with x≥x*(p*)), while a fraction 1-p* defect (those with x<x*(p*)). Both types of players receive higher expected payoffs in this equilibrium compared to the one-shot prisoner's dilemma with mutual defection. Thus, a sufficient amount of integrity results in a Pareto-improvement by promoting a critical amount of cooperation.

A numerical example illustrates the preceding equilibrium. Let c=6, a=5, b=2, and d=0. The right-hand side of (13) thus equals 8/3. In order to satisfy (13), let θ=3. It then follows from (12) that p*=.5 (which exceeds (b-d)/(a-d)=2/5, as required), and from (6) that x*(p*)=1.5. In this numerical example, pure defectors are those with x<1, pure cooperators are those with x>2, and conditional players are those with 1<x<2. In the latter group, those with 1<x<1.5 defect, and those with 1.5<x<2 cooperate. Thus, players with x≥1.5 cooperate in the stage-two subgame and receive a payoff of $U^c(p^*)=2.5$, while players with x<1.5 defect and receive a payoff of $U^d(p^*)=4-x$. Figure 2 graphs these payoffs as a function of x, where the darkened segments show the relevant payoffs. As noted above, the payoff for all players exceeds b=2, the payoff from mutual defection in the standard prisoner's dilemma.

To summarize, we have shown that when sufficient integrity (θ) exists, all players will choose to make promises in the first stage to cooperate in second. When there is a cost to lying, agents will actively seek to engage in transactions where others similarly make "non-binding" promises to cooperate. Mutual promise-making not only increases cooperation rates and expected payoffs, it is rational. The introduction of a continuum of integrity means that three different modes

of purposeful behavior emerge. Those adopting the first mode, pure defectors, are concerned only with their own material payoffs. These players lie in the first stage and then defect in the second. This is the behavior typically assumed by economists for *all* economic agents and is the purported source of cheap talk. In contrast, members of the second group, pure cooperators, never lie and are never distrustful. Their talk is golden, not cheap. They may end up playing with a liar and defector, but in our model the exit option means that pure cooperators do not have to lie and do not have to play the sucker. The third group, conditional players, are something of a mixture between the first two. Its members may cooperate with cooperators because they possess some integrity, but their preferences are sufficiently material that they eschew being suckers. They would lie and then defect with known defectors (or exit) rather than turning the other cheek. Therefore, in our model if someone tells the truth the reason they do so is either because they have high integrity, or, if they have lower integrity, because they believe that there is sufficient probability that they will interact with someone with enough integrity who will likewise keep their word. Both material and moral motivations are included and behavior is not the result of any imposed strategy, it arises endogenously.

Once moral motivations are incorporated into traditional analyses, a lot of real economic behavior becomes easier to explain. Our model provides an explanation for the results found in the experimental literature: when people say they will contribute to a public good they often do so, even in opposition to their material interests, precisely because to do otherwise would be to lie. For some, the moral injunction against lying is great. For others, it is sufficient to prevent lying to those believed to be truth-tellers. Talk may be cheap for oligopolists engaging in mutual deceit, but it is not for many workers, employers, buyers, and sellers.

2.1 Relationship to the Literature

Our model differs from most because it provides a role for integrity and non-credible promises. In the contexts we have chosen, integrity could only be relevant if agents have made promises to one another, which obviously necessitates communication. That means that our analysis does not explain cooperation in games without communication and without promise-making. Still, it is instructive to briefly report further differences to the literature on cooperation in which communication is not central. Our results differ from Frank (1987), Witt (1986), Guttman (1996), and Sethi (1996) because cooperative behavior emerges in a one shot-game in our model. Guttman and Sethi find cooperation because of external costs of interaction, we find it because of the "internal" costs associated with ones own integrity debasement. In Frank's model agents can increase the probability of interacting with cooperators by incurring detection costs. In our analysis promise-making performs a similar function. Our model is similar to Rabin's (1993) in that he also

includes an adjustment to material payoffs (a "kindness function") which potentially changes the players' strategies (and hence, the equilibrium of the game). However, his adjustment factor is based on beliefs about other's intent, whereas our adjustment is *internal* to each player. As a result, players in our model can be single-minded cooperators or defectors (as in Frank), or conditional players (though the two types "shade" into one another), whereas Rabin only considers the latter. Finally, Fehr and Schmidt (1999) posit a preference for inequity aversion. Cooperation in public goods dilemmas can be sustained if either there is a sufficient number of inequity averse players, or if punishment is allowed, there is a sufficient number of players with a high degree of inequity aversion who are (credibly) willing to punish defectors. Interestingly, their model generates behavior resembling our conditional players in that some are willing to cooperate if *all* others do as well (p.841). Of course, in our model conditional behavior occurs even when it is known there are a significant number of defectors in the population. And, in our model, cooperation stems from promise-making and integrity, which is logically as well as operationally distinct from preferences over equity.[1]

A few analyses exist which consider communication and promises, either implicitly or explicitly. Dowell, Goldfarb, and Griffith (1998) include honesty as a preference, but as a lumpy one – meaning that it is discontinuous and cannot be traded off against other (material) goods at the margin, to explain moral behavior in economic interactions. The idea is that an act is either moral or not. While not modeled explicitly, presumably, communication is necessary for at least some honest behavior (but not others, like the example of drug dealing used by the authors). While our analysis is in some ways similar because high integrity can be thought of as a lumpy moral preference, our use of a continuum of integrity is more general because some agents, conditional player, *do* trade off moral and material considerations at the margin. Not only do we include those who do and don't (pure defectors and cooperators) make those tradeoffs, we also model their random interactions in order to find out how they will behave. Caporael, Dawes, Orbell and van de Kragt (1989) suggest that *sociality* could explain increased cooperation in experimental games. Sociality is an evolutionary mechanism which selects for cooperative behavior in order to assist group functioning, and is triggered by communication. The authors do not provide a formal model of sociality, but, in any case, we are unable to speculate about its relationship to the explanation offered in this paper. Finally, Orbell, van de Kragt, and Dawes (1991) explicitly consider promise-making but suggest that it only confers moral obligation (and hence can alter behavior) when others promise also.[2] As is hopefully clear by now, the existence of integrity in our model means that

---

[1] Our analysis is similar to Casson's (1991) effort in the important way that it also introduces a non-material, internal cost to cheating. Contextually, Casson focuses on a leader's ability to instill the preference of trustworthiness in his followers.

[2] In further contrast to our integrity based account, Rawls (1971) argues that promise-keeping is obligatory if one voluntarily accepts the benefits of the social practice of promise-keeping. To do otherwise is to free-ride by

most have to consider the cost of promise-making, meaning a moral obligation occurs quite independently of the behavior of others.

3. Conclusion

Talk is not always cheap; to believe otherwise is to ignore not only everyday experience, but also the experimental evidence. Accordingly, our analysis has sought to more fully capture real human behavior. While some people do not possess abundant integrity, others do. The analysis offered here shows that a fuller characterization of behavior need not lead to trivial results. We did not impose a preference for cooperation. Instead, the notion of integrity was employed to show why mutual promise-making is rational, and to generate three types of (equilibrium) behavior.

By modeling integrity in a one-shot game we assure that any resulting cooperation does not come from reputation or punishment effects associated with a repeated game. That truth-telling against one's material interests can occur in a one-time interaction is the main point of the analysis. Still, it would be instructive to extend the model by adding additional periods to see if and how equilibrium behavior changes. While a proper treatment of that issue is beyond the scope of this paper, we report the results of a simple extension of the model that involved adding a second period of randomly interacting agents, each of whom employ Bayesian updating (rather than backward induction) based on the observed first period behavior of their (new) partner.[1]

There are three possible second period pairings: two cooperators (both told the truth in the first period), two defectors (both lied), and one cooperator and one defector. It turns out that first period cooperators will continue to cooperate with each other and first period defectors will continue to defect. In both cases each agent can infer the range of integrity of their partner and calculate the probability of future behavior with certainty. The interesting case is the third one, where two different first period behavioral types are paired. The first period cooperator must calculate the probability that a first period defector will now cooperate with a cooperator, and a first period defector must calculate the probability that a first period cooperator will now cooperate with a defector. It turns out that if there is the possibility of exit it is rational for the first period cooperator to exit; a first period defector never gets to pair with a first period cooperator. If, however, the possibility of exit is removed, for instance because agents are locked into a second period contract, then the story changes. Some first period cooperators continue to cooperate (those with high integrity), but some now defect. That is, they *switch* their behavior based on their own

---

accepting the benefits of others' compliance, without bearing any of the cost of compliance. Scanlon (1990) questions an account based on social practice, suggesting instead that a promisor's obligation derives from the expectation created in another.

[1] Conlisk (1996) surveys the experimental literature (as well as other evidence) and provides significant support for adaptive expectations and bounded rationality.

integrity and the updated estimated probability that their partner will defect. Moreover, while some first period defectors still defect, some now cooperate in the second period with first period cooperators. We find this last result most interesting. If it is found that switching behavior continues to hold in more general repeated frameworks, it would mean that the addition of integrity implies that the overall cooperation rate in an economy increases with the ability of agents to interact with known cooperators/truth-tellers. Furthermore, the existence of integrity means that even former liars could reform in time.

References

Bohnet, I., and B. Frey, 1999. "The Sound of Silence in Prisoner's Dilemma and Dictator Games," Journal of Economic Behavior and Organization, 38, pp. 43-57.

Bok, S., 1978. Lying: Moral Choice in Public and Private Life, New York: Pantheon.

Bok, S., 1995. Common Values, Columbia: University of Missouri Press.

Bowles, S., 1998. "Edogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions," Journal of Economic Literature, 36, pp. 75-111.

Boyd, R., and P. Richerson, 1990. "Culture and Cooperation," in J. Mansbridge (ed.), Beyond Self Interest, Chicago: University of Chicago Press.

Caporael, L., R. Dawes, J. Orbel, and A. van de Kragt, 1989. "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives," Behavioral and Brain Sciences, 4, pp. 683-98.

Casson, M., 1991. Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance, Oxford: Clarendon Press.

Conlisk, J., 1996. "Why Bounded Rationality," Journal of Economic Literature, 34, pp.669-700.

Davis, D., and C. Holt, 1993. Experimental Economics, Princeton: Princeton University Press.

Dawes, R., J. McTavish, and H. Shaklee, 1977. "Behavior, Communication, and Assumptions about Other People's Behavior in a Commons Dilemma Situation," Journal of Personality and Social Psychology, 35, pp. 1-11.

Dowell, R., R. Goldfarb, and W. Griffith, 1998. "Economic Man as a Moral Individual," Economic Inquiry, 36, pp. 645-53.

Elster, J., 1989. "Social Norms and Economic Theory," Journal of Economic Perspectives, 3, pp. 99-117.

Elster, J., 1996. "Rationality and the Emotions," Economic Journal, 106, pp. 1386-97.
Elster, J., 1998. "Emotions and Economic Theory," Journal of Economic Literature, 36, pp. 47-74.

Etzioni, A., 1986. "The Case for a Multiple-Utility Conception," Economics and Philosophy, 2, pp. 159-183.

Fehr, E., and K. Schmidt, 1999. "A Theory of Fairness, Competition, and Cooperation," Quarterly Journal of Economics, 114, pp.817-868.

Frank, R., 1987. "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One With a Conscience?," American Economic Review, 77, pp. 593-604.

Frey, B., and I. Bohnet, 1995. "Institutions Affect Fairness: Experimental Investigations," Journal of Institutional and Theoretical Economics, 151(2), pp. 286-303.

Gauthier, D., 1986. Morals by Agreement, Oxford: Clarendon.

Griffith, W. and R. Goldfarb, 1991. "Amending the Economist's "Rational Egoist" Model to Include Moral Values and Norms, Part 1: The Problem," in K. Koford and J. Miller (eds) Social Norms and Economic Institutions, Ann Arbor: University of Michigan Press.

Goldfarb, R. and W. Griffith, 1991. "Amending the Economist's "Rational Egoist" Model to Include Moral Values and Norms, Part 2: Alternative Solutions," in K. Koford and J. Miller (eds) Social Norms and Economic Institutions, Ann Arbor: University of Michigan Press.

Guttman,., 1996. "Rational Actors, Tit-For-Tat Types, and the Evolution of Cooperation," Journal of Economic Behavior and Organization, 29, pp.27-56.

Hausman, D. and M. McPherson, 1993. "Taking Ethics Seriously: Economics and Contemporary Moral Philosophy," Journal of Economic Literature, 31. pp. 671-731.

Heap, S., and Y. Varoufakis. Game Theory, London: Routledge.

Isaac, R., and J. Walker, 1988. "Communication and Free Riding Behavior: The Voluntary Contribution Mechanism," Economic Inquiry, 26, pp. 585-608.

Isaac, R., and J. Walker, 1991. "Costly Communication: An experiment in a Nested Public Goods Problem," in T. Palfrey (ed) Laboratory Research in Political Economy, Ann Arbor: University of Michigan Press.

Ledyard, J., 1995. "Public Goods: A Survey of Experimental Research," in Kagel, J. and A. Roth (eds), Handbook of Experimental Economics, Princeton: Princeton University Press.

Marwell, G. and R. Ames, 1981. "Economists Free Ride, Does Anyone Else?," Journal of Public Economics, 15, pp. 295-310.

Minkler, L., 1999. "The Problem with Utility: Toward a Non-Consequentialist/ Utility Theory Synthesis," Review of Social Economy, 62, pp.4-24.

Orbell, J., A. van de Kragt, and R. Dawes, 1991. "Covenants Without the Sword: The Role of Promises in Social Dielmma Circumstances," in K. Koford and J. Miller, Eds., Social Norms and Economic Institutions, Ann Arbor: University of Michigan Press.

Poundstone, W., 1992. Prisoner's Dilemma, New York: Doubleday.

Rabin, M., 1993. "Incorporating Fairness into Game Theory and Economics," American Economic Review, 83, pp. 1281-1302.

Rawls, J., 1971. A Theory of Justice, Cambridge: University of Harvard Press.

Sally, D., 1993. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958-1992," Rationality and Society, 7, pp. 58-92.

Scanlon, T.M., 1990. "Promises and Practices," Philosophy and Public Affairs, 19, 199-226.

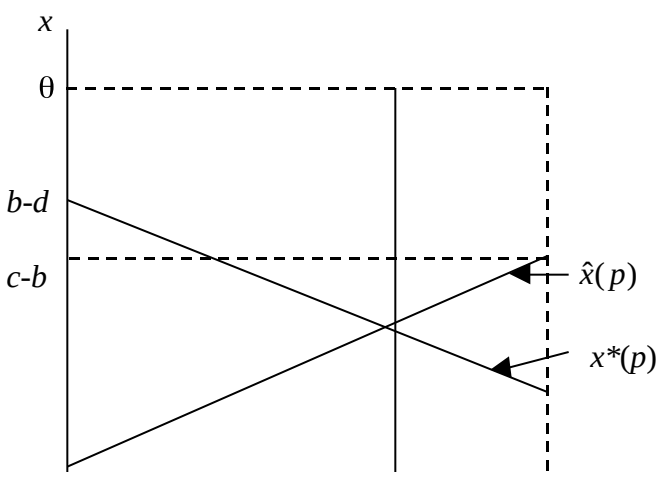
Schneider, F. and W. Pommerehne, 1981. "Free-Riding and Collective Action: An Experiment in Public Microeconomics," Quarterly Journal of Economics, 96, pp. 689-704.

Sen, A., 1978. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," in H. Haris (ed.) Scientific Models and Men, London: Oxford University Press.

Sethi, R., 1996. "Evolutionary Stability and Social Norms," Journal of Economic Behavior and Organization, 29, pp. 113-140.

Smith, A., 1976. The Theory of Moral Sentiments, Liberty Classics.

Witt, U., 1986. "Evolution and Stability of Cooperation without Enforceable Contracts," KYKLOS, 39, pp. 245-66.


x
θ
b-d
c-b
x̂(p)
x*(p)

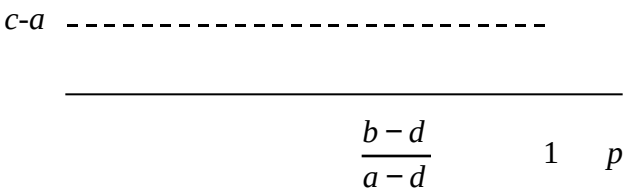


Figure 1.

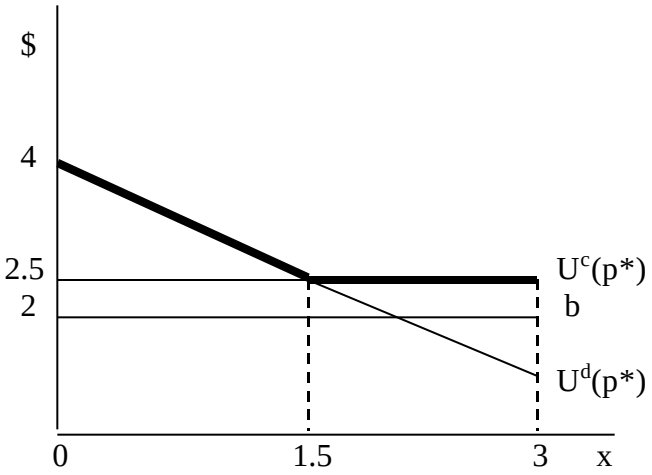


Figure 2.