The Fragility of Cooperation: A False Feedback Study of a Sequential Iterated Prisoner's Dilemma

John Monterosso Ph.D.[τ], George Ainslie, M.D.*,
Pamela Toppi Mullen, P.A.-C*, Barbara Gault, Ph.D*.

Final draft of an article published in
Journal of Economic Psychology 23, 437-448, 2002

[τ] University of Pennsylvania Department of Psychiatry
*Coatesville Veterans Affairs Medical Center Department of Psychiatry

---

Address correspondence to either:

George Ainslie, M.D.
Department of Psychiatry
Coatesville VA Medical Center
Coatesville, PA 19320
email: ainslie.george@coatesville.va.gov

John Monterosso, PhD
Department of Psychiatry
University of Pennsylvania
3900 Chestnut St.
Philadelphia, PA 19104
email: jmont@psych.upenn.edu

ABSTRACT

We examined the mutability of naturally occurring mutual cooperation and mutual defection. Forty-five pairs of subjects participated in an extended iterated prisoner's dilemma (median duration 1,807 trials) using a monetary payoff matrix. When stable cooperation or defection emerged, false feedback was provided indicating to each subject that his partner was choosing contrary to previously stable play. This was followed by recovery trials in which false feedback indicated to each subject that his partner had resumed making the previously stable choice.

While stable cooperation occurred more frequently than stable defection, it was considerably more vulnerable to the false feedback manipulation. This was true both in terms of the extent to which choice changed in response to false feedback ($p=.006$) and in terms of the extent to which the disruption persisted ($p<.001$). While the effect of four false feedback cooperations was undone by a single recovery false feedback defection, the effect of even a single false feedback defection was still apparent after 7 false feedback recovery cooperations. These results are discussed in relation to the analogy between interpersonal bargaining and intertemporal bargaining within individuals.

INTRODUCTION

The contingencies present in social dilemmas are such that rational individuals pursuing their self-interest will have collectively unfavorable outcomes (Hardin, 1968). However, when the payoff matrices of social dilemmas are applied to a sufficiently small group of players (especially 2), and play is open and indefinitely repeated, the conflict between self and collective interests is not inevitable. The repeated 2-person social dilemma known as the "iterated prisoner's dilemma" (IPD) provides each participant with the ability to create contingencies for her partner[1] that favor the partner's cooperation (Rapoport, Chammah, & Orwant, 1965). For instance, an individual using the straightforward strategy of tit-for-tat (Axelrod, 1980; Axelrod, 1984; Axelrod & Dion, 1988) simply does unto her partner whatever it was that was most recently done unto her. In so doing, she adds the prospect of her subsequent cooperation to her partner's payoff for cooperation. This brings the interest of the individual into alignment with that of the group (Rapoport et al., 1965).

Empirical studies of 2-person IPDs have examined the influence of a variety of factors on choice, including the specific pay-offs used, the effect of requiring sequential versus simultaneous moves, the effect of communication between participants, and the effect of specific participant characteristics (for review see Komorita & Parks, 1999). However, we know of no experiments that introduce ostensible moves in a controlled fashion and study their effect on established patterns of play. Such a manipulation could provide direct evidence regarding the effect of particular play on the evolution of bargaining patterns over time.

Simultaneous Vs Sequential Iterated Prisoner's Dilemma

In the standard IPD game, participants make their choices simultaneously on each round. Decisions may thus be influenced not only by expectations regarding one's opponent's future choices, but also by expectations regarding one's opponent's choice in the current round (Watabe, Terai, Hayashi, & Yamagishi, 1996)– an uncertainty that can no longer be influenced. In the *sequential* version of the IPD (Deutsch, 1960; Erev & Rapoport, 1990; Oskamp, 1974), partners do not choose simultaneously but instead take turns choosing between two options which are relevant to both individuals (e.g., 8 cents for me versus 5 cents for each of us). The sequential variant of the IPD is the better analog to most naturally occurring IPDs, (Boyd, 1988). For instance, neighbors calling on each other for help over time can be viewed as interacting in an IPD with decisions made sequentially rather than simultaneously. While any simultaneous prisoner's dilemma matrix can be transformed into a sequential matrix and vice versa, equivalent matrices across these two variants have evoked different choices. Depending on the particular payoff matrix used, in some cases the sequential IPD has yielded *more* cooperation than the simultaneous IPD, and in some cases the sequential IPD variant has yielded *less* cooperation than the simultaneous IPD (Oskamp, 1974).

Cooperation Over Time

Despite the potential for participants in an iterated prisoner's dilemma to establish contingencies that encourage their partner's cooperation, mutually cooperative relationships are by no means assured. Although it is not meaningful to generalize as to

[1] We use the term "partners" throughout to designate individuals involved in the same prisoner's dilemma. We do not intend to imply that such individuals are united or positively disposed to each other.

precisely how high cooperation rates are since this varies greatly with the particular payoff matrix (Murnighan & King, 1992; Rapoport et al., 1965) and myriad other factors (Nydegger, 1980; Surbey & McNally, 1997; Swensson, 1967), it is notable that across settings, overall performance rarely approaches perfect cooperation. Furthermore, a tendency for participants to become more cooperative with greater exposure to the contingencies is neither overwhelming nor consistent. In one study that reported a significant change in cooperation rate over a period of 50 trials, cooperation rate was found to increase across the first two blocks of 10 trials, and then to remain stable (Oskamp, 1974). In another set of studies that looked at repeated play over hundreds of trials, cooperation rates were reported to decrease initially, and then rise subsequently (Rapoport et al., 1965). Still another study looking at an iterated prisoner's dilemma reported a steady decline in cooperation over the course of subjects' interactions (Swensson, 1967).

Some of the variability of average long term course may come from the tendency of individual pairs to gravitate toward either consistent cooperation (CC) or consistent defection (DD). That is to say, the occurrence of rounds of asymmetrical play in which one player cooperates and the other defects (CD) decline and are replaced by CC or DD. Based on an extensive set of studies spanning several variants of the IPD, Rapoport and colleagues reported a, "steady decline of unilateral states, i.e., the increasing predominance of CC and DD states" (Rapoport et al., 1965). This empirical finding is not surprising. Unlike the ordinary social dilemma, the IPD has no dominant strategy. Optimality always depends on the strategy of one's partner. Early in play, participants cannot be certain about their partner's strategy and so their moves may be inconsistent, as they test their partner, or as they correct their conception of their partner. If a partner's strategy is static and can be discerned, a dominant move becomes apparent. For instance, if one learns her partner is playing tit-for-tat, she should cooperate consistently from that point on. Stability may arise as one's partner's strategy is revealed, and when it does, it is likely to arise either in the form of mutual cooperation, or in the form of mutual defection.

The properties of naturally arising stable behavior in an IPD context have not been explored in a controlled fashion. In this report, we ask the question: When a relationship of stable cooperation or stable defection emerges, how mutable is it? Using a sequential IPD we consider this question in two ways: 1) When stable cooperation or stable defection between partners is artificially disrupted by the insertion of false feedback regarding one's partner's behavior, how readily is the stable choice abandoned? and 2) Once stable cooperation or stable defection is abandoned, how readily will subjects return to it when they are provided with false feedback indicating that their partner has returned to it? Although prior research suggests that stable relationships of mutual cooperation emerge somewhat more frequently than do stable relationships of mutual defection (Rapoport et al., 1965), there have been no reports of how stability is affected by changes (real or apparent) in one player's behavior.

METHOD

Subjects

A "convenience sample" of 90 subjects was recruited from two elective substance abuse rehabilitation programs at the Coatesville Veterans Affairs Medical Center. While

functionally similar, the two programs operate as closed systems in separate buildings within the hospital complex-- Fraternizing between programs is prohibited. Each subject pair consisted of one subject from each of the two programs, so that we could insure there was no communication between subject pairs during the course of their week of participation.

All subjects were male veterans between the ages of 26 and 58 (mean age of 41). The breakdown of subjects' self-identified race was as follows: 73.1% African-American, 25.7% Caucasian, and 1.1% Native American. Reported drugs of abuse were alcohol (22.9%), cocaine (18.9%), alcohol and cocaine (50%), heroin and cocaine (2.2%), and all three (6.8%). Subjects were otherwise healthy. Recruitment was conducted through announcements made at inpatient group meetings.

Procedure

During the experiment, each subject of a pair sat at a computer terminal on different floors of a hospital facility. Subjects were read the following instructions at the beginning of their participation in the experiment:

> You are one of two players participating in the Bargaining Game. You will play five games with someone who is in another room. There are at least 100 turns per game. The object of the game is to get as many points as you can. The points you earn will be converted into canteen books [hospital store scrip].
>
> On every turn you will choose between two options: Option #1 is 100 points for yourself. Option #2 is 70 points for yourself, and 70 points for your partner. These options are the same for the other player. When he chooses #1, you don't get anything. When he chooses #2, you get 70 points. Choosing #1 is called "going it alone". Choosing #2 is called "cooperating". Once you've earned points they're yours to keep. A running tally of your points is displayed after every turn.
>
> Remember, the object of the game is to get as many points as you can. The points will be converted into canteen books. You will get $1 in canteen books for every 600 points, with your final total rounded up to the nearest dollar. First you'll play a practice game and then the real game will start.

One subject was randomly chosen to make the first move, after which subjects alternated for the remainder of the experiment. When it was a subject's turn, the following choice would appear on the screen:

"**Option 1**: 100 points for you."

"**Option 2**: 70 points for you, 70 points for him."

Subjects responded by striking the '1' or '2' key. Subjects were informed of their partner's choice on each of his turns.

The same pair participated in as many as four days of play over one week of participation. Each session was divided into five blocks that served primarily to break up the monotony of the task. One practice game preceded the five blocks on the first day of participation for each pair. In order to avoid endgame effects, blocks were terminated according to a pseudorandom schedule, which approximated the distribution of endpoints expected given a P=.05 chance of termination on each turn after trial 100.

Subjects were told how much they earned at the end of each session, but did not receive their money until they completed all four sessions or ended their participation in the experiment. Payment was given in scrip exchangeable for a wide-range of goods at the hospital store.

Stable play was defined operationally as 10 consecutive choices of defection by the pair, or 10 consecutive choices of cooperation by the pair. Whenever this criterion was reached, each subject received false feedback for 8 trials. The false feedback consisted first of 0,1,2,3, or 4 *contrary* false feedback trials for each member of the pair, followed by *recovery* false feedback for the remainder of the 8 trials for each member of the pair. For example, a pair that had cooperated consistently for 10 trials (5 per subject) might be provided with 6 trials (3 per subject) in which they were informed that their partner had defected – regardless of whether this was true. This would be followed by 10 trials (5 per subject) in which they were each informed that their partner had resumed cooperating -- again, regardless of whether this was true.

RESULTS

Subject Participation

Because of scheduling conflicts and attrition, not all of the 45 pairs of subjects completed all four days of the experiment. Eight subjects (17.7%) participated in 2 days of play, 9 subjects (20.0%) participated in 3 days of play, and the remaining 28 subjects (62.2%) participated in the entire 4 days of the play. The median total number of trials across pairs was 1807, with the lower quartile participating in 1610 trials and the upper quartile participating in 2166 trials of play.

Overall Cooperation Rates

The median rate of cooperation among the 90 subjects was 65.0%, with a cooperation rate of 46.8% at the lower quartile, and 76.3% at the upper quartile. Excluding those trials in which subjects received false feedback, the rates of cooperation were somewhat higher, with a median rate of 68.5%, and a cooperation rate of 48.1% at the lower quartile and 78.0% at the upper quartile. Cooperation rates between the two subjects in each pair were highly correlated ($r=.84$, $p<.001$) over the course of each pair's interactions. Furthermore, cooperation rates within subject pairs were moderately stable over time, with cooperation rates on day 1 correlating significantly with cooperation rates on day 4 ($r=.58$, $p=.005$). Based on a repeated measure ANOVA over the first 3 days of play (so as to maximize the number of subjects who could be included in the analysis) no main effect of time was observed ($F(2,37) = 1.15$, $p = .32$). The rate of cooperation among those pairs that participated in all four days of play did not differ significantly from that of pairs that participated in less than four days of play ($t=1.2$, $df=42$, $p=.22$).

The median rate of cooperation after an observed defection, computed by subject, was 35.5%, and after an observed cooperation was 80.1%. For each subject we computed the difference score between cooperation rate following cooperation of one's partner, and cooperation rate following defection of one's partner. This "tit-for-tat-tendency" was slightly correlated between the 2 subjects that made up each pair ($r=.23$, $p=.03$).

As would be expected given the relatively high overall rate of cooperation, the criterion for stable cooperation (10 consecutive cooperations) was more frequently met by subjects than was the criterion for stable defection (10 consecutive defections). The average total occurrences of stable cooperation among subject pairs was 33.4 (the median

was 34 and the interquartile range was 18.5 and 49). The average total occurrence of stable defection among subject pairs was 12.7 (the median was 8 and the interquartile range was 1 and 16.5). The total occurrence of stable cooperation and stable defection tended to be inversely related within subject pairs ($r=-.60$, $p<.001$). However, 22 of the 45 subject pairs had at least 6 occurrences of both types of stable play across the course of their participation.

The total occurrence of stable cooperation periods was positively associated with overall cooperation rate ($r=.77$, $p<.001$). In a regression analysis controlling for overall cooperation rate, the frequency of stable cooperation occurrences was also related to each pair's mean tit-for-tat-tendency ($t=2.7$, $df=42$, $p=.009$). The frequency of stable defection was negatively associated with overall cooperation rate ($r=.89$, $p<.001$). In a regression analysis, again controlling for overall cooperation rate, there was a trend suggesting more frequent stable defection among pairs with a higher tit-for-tat-tendency ($t=1.8$, $df=42$, $p=.07$).

False Feedback Contrary Moves

Cooperation rates following stable defection increased in response to false feedback of cooperation. Cooperation rates rose from 4.0% to 27.7% after 1 cooperation, to 33.9% after 2 cooperations, to 36.6% after 3 cooperations, and to 39.1% after 4 cooperations (See Figure 1). Conversely, cooperation rates following stable cooperation decreased precipitously in response to false feedback of defections. After a single false feedback of defection, cooperation rates dropped from 98.0% to 53.2%. Continued false feedback of defections produced further deterioration of cooperation – to 42.0% after the second defection, 37.0% after the third, and 39.4% after the fourth.

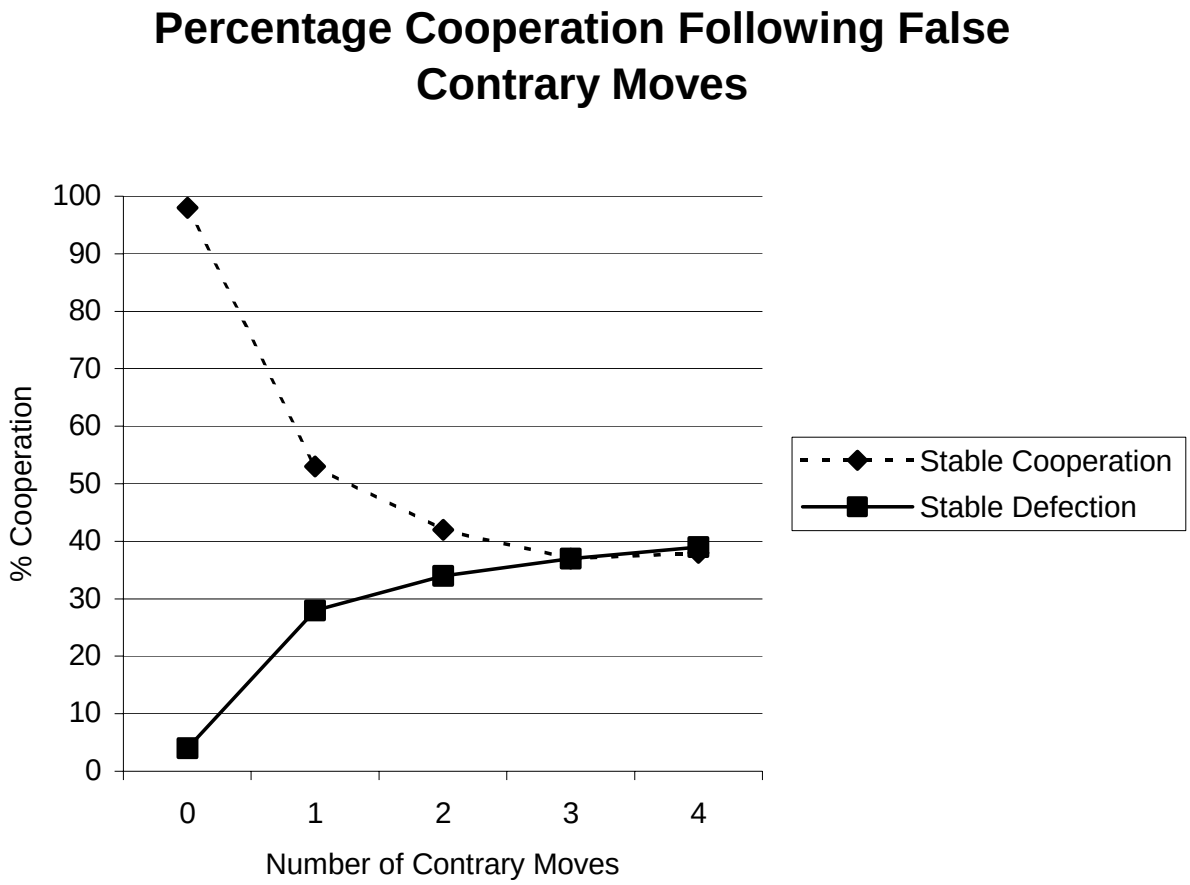

Figure 1- The above figure illustrates the affect of false feedback trials on mean cooperation rates. On these trials, pairs engaged in either stable cooperation or defection are informed that their partner had made the move contrary to actual stable play. While contrary moves effect both stable cooperation and stable defection, their effect on stable cooperation is greater.

In order to compare the rates of disruption that contrary moves caused to stable defection with that caused to stable cooperation, the absolute change in mean cooperation was computed for each pair across instances of 0 through 4 false feedback contrary moves. A repeated measure ANOVA was conducted on the resulting values, with type of prior stable play (cooperation or defection) included as a between subject independent variable and the number of false feedback contrary moves included as a within-subject independent variable. Because of the requirements of the repeated measure ANOVA, only pairs exposed to the range of false contrary moves (0-4) could be included in the analysis. As is suggested by Figure 1, false contrary moves led to significant disruption of stable play, and did so according to an approximately quadratic function, with the

greatest change occurring after the first false contrary move, $F(4,63)$[2]$=98.3$, $p<.001$). In terms of its effect on play, there was a significant interaction between the occurrence of contrary false feedback, and whether prior stable play was cooperation or defection. As is apparent in Figure 1, the disruption over time to stable cooperation was significantly greater than the disruption to stable defection, $F(4,63)=8.1$, $p=.006$.

False Feedback Recovery Moves

Figure 2 depicts the return to prior stable play with recovery false feedback of cooperation (return to cooperation) and recovery false feedback of defection (return to defection). As explained above, these recovery insertions occurred after inserted contrary moves. In a repeated measure ANOVA (by pair) with type of previously stable play as a between subject factor and the number of recovery moves as a within subject factor, a significant interaction was observed between the type of prior stable play and the number of recovery moves, $F(4,69)=5.4$, $p<.001$. As is apparent in Figure 2, return to previously stable defection was faster than return to stable cooperation. Considering separately subjects who had reached the criterion for stable defection, the increased cooperation rate stimulated by prior false feedback of cooperation was rapidly undone by recovery false feedback of defection. After a single such instance, there was only a trend across groups based on whether they had just been exposed to 0,1,2,3, or 4 false feedback cooperations, $F(4,21)=2.8$, $p=.06$. After a second recovery false feedback defection, no effect of previous (false) cooperation was apparent $F(4,21)=4.0$, $p=.39$

[2] Because type of stable play was a between subjects factor, a pair that was exposed to 0-4 false feedback contrary defections, and 0-4 false feedback contrary cooperations would contribute 2 degrees of freedom to the error term.

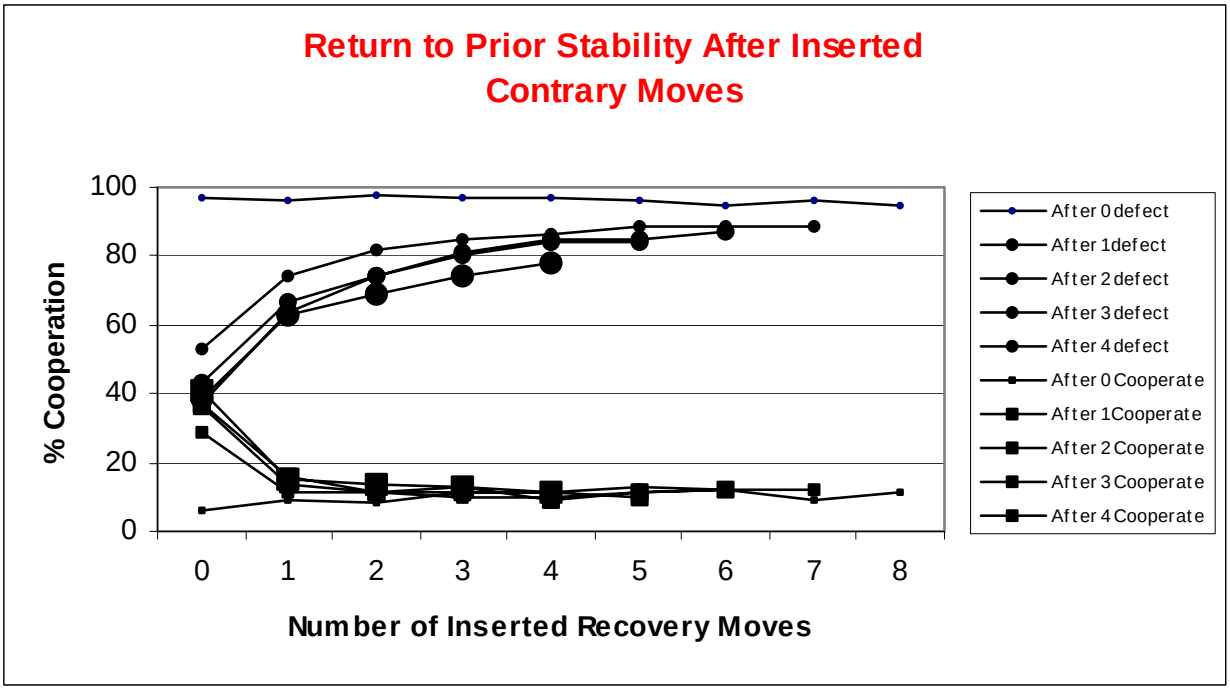

Figure 2- The above figure illustrates the effect of "recovery" moves occurring after varied amounts of false feedback contrary moves. Recovery moves were also false feedback, and informed subjects that their partner had made the move consistent with prior stable play. While cooperation rates tended to return to previous levels in both cases, the return to stable defection was more immediate than the return to stable cooperation.

By contrast, contrary false feedback defections had had an enduring effect on cooperation. After a single recovery false feedback cooperation, cooperation rates continued to differ dramatically based on how many previous false feedback defections had been presented $F(4,39) = 20.9$, $p<.001$. Furthermore, as is apparent in Figure 2, as far out as the maximum point of comparison, a significantly lower cooperation rate remained, based on whether or not a single contrary false feedback defection had occurred 7 moves earlier, despite false feedback of recovery cooperation during the intervening 7 moves $F(1,39) = 4.4$, $p=.04$.

DISCUSSION

Although cooperation was the more common response in the present study, naturally occurring mutual cooperation was, in two senses, more fragile than mutual defection. First, steady cooperation was more dramatically disrupted by the insertion of false feedback defection, as compared to the extent to which defection was disrupted by false feedback cooperation. Whereas cooperation rates dropped 44.8% after a single inserted false defection, defection rates only dropped by 23.7% in response to a single false feedback of cooperation. Second, and more dramatically, the deviation from stable cooperation resulting from false feedback of defection was far more long-lasting than was the deviation from stable defection in response to false feedback of cooperation. Whereas, in the latter case, a single instance of feedback that one's partner had returned to defection brought a near complete return to prior stable defection, the return to stable cooperation was not complete after as many as 7 trials in which feedback indicated that one's partner had returned to cooperative behavior.

These results can be considered in relation to the effects that "noise" can have in iterated prisoner's dilemmas. As has been pointed out, in naturalistic settings, the relationship between intent and behavior, and between behavior and feedback, is often far less clear than it is in experimental settings (Axelrod, 1997). This problem has been widely considered especially in modeling and analysis of the Tit-for-tat strategy (Axelrod, 1984; Bendor, 1993). If Tit-for-tat is playing itself, a single error or false feedback provided to both partners can turn mutual cooperation into permanent mutual defection. If the IPD is simultaneous, a single error or false feedback provided to one partner can transform stable cooperation into a perpetually "echoing" state of alternating unilateral cooperation (Downes, Rocke, & Siverson, 1986). Such states can only be corrected by either the occurrence of another error, or by behaviors that are not part of Tit-for-tat such as forgiveness or contrition (Wu & Axelrod, 1995). The results of this study suggest an additional basis for concern about the effect that imperfect feedback can have on mutual cooperation in naturally occurring IPD's. Here, recovery moves -- tantamount to acts of contrition which would be enough to restore cooperation among individuals playing Tit-for-tat-- are empirically shown to be highly imperfect in terms of restoring cooperation.

Also of interest, the results using the false feedback design demonstrated both the strengths and perils of the Tit-for-tat strategy. On the positive side, pairs that played more in accordance with Tit-for-tat tended to cooperate more. Furthermore, controlling for overall cooperation rates, pairs playing in accordance with Tit-for-tat tended to reach the criterion for stable cooperation rate more often ($p=.009$). This suggests that among pairs of players, Tit-for-tat was effective in promoting consistent cooperation. At the same time, when controlling for overall cooperation rate, there was a trend suggesting that pairs adopting more of a Tit-for-tat strategy also tended to reach the criterion for stable defection more frequently ($p=.07$). This empirical observation is consistent with analyses indicating pure Tit-for-tat's inherent vulnerability for becoming locked into relationships of mutual defection (Axelrod, 1984).

All subjects in this sample were male veterans enrolled in voluntary treatment for substance-dependence. We chose this population not out of a particular interest, but simply because they were available to us and their status as residential patients made it feasible to carryout long series of play without the risk that partners would communicate between sessions. It might be hypothesized that this group, perhaps given their history of addictive behavior, would be less inclined towards cooperative interactions. Alternatively, it might be supposed that because of their mutual affiliation within the group of veterans, or as addicts in voluntary recovery, they might be more inclined to cooperate with each other. We cannot rule out that some elements of the data we report do not generalize to other populations. It should be noted that our emphasis is not on the overall levels of cooperation, but rather the particular pattern of behavior that was observed (cooperation was the more common choice, but at the same time, the more easily disrupted choice). It seems unlikely that this pattern is specific to the particular population, though only similar designs in other populations would be conclusive on the matter.

The data reported in this study are of interest also with respect to a proposed analog between interpersonal prisoner's dilemma and *intertemporal* prisoner's dilemmas (Ainslie, 1975; Ainslie, 1992; Ainslie, 2001; Brown & Rachlin, 1999; Rachlin, 1997).

The analogy is as follows: Interests in the individual change systematically over time because rewards are discounted hyperbolically. An immediately more rewarding outcome (e.g., a rich desert) often is dominated by and then dominates an outcome that may be less rewarding in the short run but which is more valuable in the long run (e.g. an attractive figure). The individual thus is in a state of limited warfare with her future selves. While she prefers to eat cake today, she also prefers that her future selves will be abstemious. But of course, when tomorrow comes, that day's cake will again hold the temptation of immediate reward. Thus the payoff matrix the individual faces resembles a prisoner's dilemma: the most valued option is to defect (eat cake) while future selves cooperate (diet), the next best outcome is to cooperate while future selves cooperate, followed by the payoff for defection while the future defects as well, and followed last by the sucker's payoff in which the current cake is forgone while future selves indulge.

According to Ainslie (2001), in the resulting limited warfare among successive selves, implicit recognition of IPD-like properties may bring stability. One's own behavior in the face of present temptation provides an obvious probabilistic indication of how she is likely to respond to future temptations. And so there is reason for her to expect that if she "defects" now, future selves will defect as well. Although there are differences between the interpersonal and intertemporal IPDs (a current self cannot retaliate against a prior self so that the strategy of tit-for-tat is not available-- See Bratman 1999, pp. 45-50) both share the elemental contingency that a player can be expected to cooperate only insofar as she sees cooperation as necessary and effective at inducing the next player(s)' cooperation. (This logic is explored more fully in Ainslie, 2001, pp. 92-100.) If the rationale for behaving according to a general rule is the tacit perception that one's choices are precedents for cooperation or defection among successive selves, then the contingencies of an IPD apply.

The asymmetry observed in this study may be related to a long noted asymmetry in the intertemporal analog described above. Referring to struggles with temptation, the Roman physician Galen advised that to "remove the defilement of passions from his soul," the patient "must not relax his vigilance for a single hour (1963, p. 45)." The Victorian psychologist, Bain, noted that in struggles with temptation, "every gain on the wrong side undoes the effect of many conquests on the right (1886, p. 440)." In recent times, this observation has been empirically documented, primarily under the label "abstinence violation effect" (Marlatt & Gordon, 1980). According to this formulation, any perceived slip in an area of self-control has a high likelihood to result in a binge of failed self-restraint. This effect has been documented in such disparate areas as drinking among alcoholics in treatment (Collins & Lapp, 1991), smoking among individuals attempting to quit (Shiffman et al., 1997; Spanier, Shiffman, Maurer, Reynolds, & Quick, 1996), eating among dieters (Grilo & Shiffman, 1994; Johnson, Schlundt, Barclay, Carr-Nangle, & et al., 1995), and fantasies among pedophiles (Hudson, Ward, & France, 1992; Ward, Hudson, & Marshall, 1994). While not a perfect analog, the asymmetry observed in the present experiment provides some suggestive evidence that the IPD, in addition to its myriad familiar applications, may be useful as an experimental model for the limited intertemporal warfare operative in domains of temptation.

BIBLIOGRAPHY

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. Psychological Bulletin, 82, 463-496.

Ainslie, G. (1992). Picoeconomics: The strategic interaction of successive motivational states within the person. New York, NY, USA: Cambridge University Press.

Ainslie, G. (2001). Breakdown of Will: Cambridge University Press.

Axelrod, R. (1980). More effective choice in the Prisoner's Dilemma. Journal of Conflict Resolution, 24(3), 379-403.

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

Axelrod, R. (1997). The Complexity of Cooperation: Agent based models of competition and collaboration. Princeton: Princeton University Press.

Axelrod, R., & Dion, D. (1988). The further evolution of cooperation. Science, 242(4884), 1385-1390.

Bendor, J. (1993). Uncertainty and the evolution of cooperation. Journal of Conflict Resolution, 37(4), 709-734.

Boyd, R. (1988). Is the repeated prisoner's dilemma a good model of reciprocal altruism? Ethology & Sociobiology, 9(2-4), 211-222.

Brown, J., & Rachlin, H. (1999). Self-control and social cooperation. Behavioural Processes, 47(2), 65-72.

Collins, R. L., & Lapp, W. M. (1991). Restraint and attributions: Evidence of the abstinence violation effect in
alcohol consumption. Cognitive Therapy & Research, 15(1), 69-84.

Deutsch, M. (1960). The effect of motivational orientation upon threat and suspician. Human Relations, 13, 123-139.

Downes, G. W., Rocke, D. M., & Siverson, R. M. (1986). Arms races and cooperation. In K. Oye (Ed.), Cooperation Under Anarchy. Princeton: Priceton University Press.

Erev, I., & Rapoport, A. (1990). Provision of step-level public goods: The sequential contribution mechanism. Journal of Conflict Resolution, 34(3), 401-425.

Grilo, C. M., & Shiffman, S. (1994). Longitudinal investigation of the abstinence violation effect in binge eaters. Journal of Consulting & Clinical Psychology, 62(3), 611-619.

Hardin, G. (1968). The Tragedy of the Commons. Science, 162(5364), 1243-1248.

Hudson, S. M., Ward, T., & France, K. G. (1992). The abstinence violation effect in regressed and fixated child molesters. Annals of Sex Research, 5(4), 199-213.

Johnson, W. G., Schlundt, D. G., Barclay, D. R., Carr-Nangle, R. E., & et al. (1995). A naturalistic functional analysis of binge eating. Behavior Therapy, 26(1), 101-118.

Komorita, S. S., & Parks, C. D. (1999). Reciprocity and cooperation in social dilemmas: Review and future directions. In D. V. Budescu & I. Erev (Eds.), Games and human behavior: Essays in honor of Amnon Rapoport. (pp. 315-330). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., Publishers.

Marlatt, G. A., & Gordon, J. R. (1980). Determinants of relapse: Implications for the maintenance of behavior change. In P. O. Davidson & S. M. Davidson (Eds.),

Behavioral Medicine: Changing health lifestyles (pp. 410-452). New York: Guilford Press.

Murnighan, J. K., & King, T. R. (1992). The effects of leverage and payoffs on cooperative behavior in asymmetric dilemmas. In W. B. G. Liebrand & D. M. Messick (Eds.), Social dilemmas: Theoretical issues and research findings. International series in experimental social psychology. (pp. 163-182). Oxford, England UK: Pergamon Press, Inc.

Nydegger, R. V. (1980). The effects of information processing complexity and interpersonal cue availability on strategic play in a mixed-motive game. Journal of Personality, 48(1), 38-53.

Oskamp, S. (1974). Comparison of sequential and simultaneous responding, matrix, and strategy variables in a Prisoner's Dilemma game. Journal of Conflict Resolution, 18(1), 107-116.

Rachlin, H. (1997). Self and self-control. In J. G. Snodgrass & R. L. Thompson (Eds.), The self across psychology: Self-recognition, self-awareness, and the self concept. Annals of the New York Academy of Sciences, Vol. 818. (pp. 85-97). New York, NY, USA: New York Academy of Sciences.

Rapoport, A., Chammah, A., & Orwant, C. (1965). Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: The University of Michigan Press.

Shiffman, S., Hickcox, M., Paty, J. A., Gnys, M., Kassel, J. D., & Richards, T. J. (1997). The abstinence violation effect following smoking lapses and temptations. Cognitive Therapy & Research, 21(5), 497-523.

Spanier, C. A., Shiffman, S., Maurer, A., Reynolds, W., & Quick, D. (1996). Rebound following failure to quit smoking: The effects of attributions and self-efficacy. Experimental & Clinical Psychopharmacology, 4(2), 191-197.

Surbey, M. K., & McNally, J. J. (1997). Self-deception as a mediator of cooperation and defection in varying social contexts described in the iterated prisoner's dilemma. Evolution & Human Behavior, 18(6), 417-435.

Swensson, R. G. (1967). Cooperation in the Prisoner'S Dilemma Game: I. the Effects of Asymmetric Payoff Information and Explicit Communication. Behavioral Science, 12(4), 314-322.

Ward, T., Hudson, S. M., & Marshall, W. L. (1994). The abstinence violation effect in child molesters. Behaviour Research & Therapy, 32(4), 431-437.

Watabe, M., Terai, S., Hayashi, N., & Yamagishi, T. (1996). Cooperation in the one-shot Prisoner's Dilemma based on expectations of reciprocity. Japanese Journal of Experimental Social Psychology, 36(2), 183-196.

Wu, J., & Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. Journal of Conflict Resolution, 39(1), 183-189.