The Tragedy of the Coffeehouse

COSTLY RIDING AND HOW TO AVERT IT

LOU MARINOFF
Department of Philosophy
City College of New York

This study reports and analyzes the results of three trials of a one-shot collective game—a variety of the assurance game—conducted via the World Wide Web. In the first two trials, most players apparently mistook the assurance game for a prisoner's dilemma and consequently attained the worst possible outcome. After contemplating appropriate strategic considerations, players who participated in the third trial attained a better outcome but did so in a surprising way. This report accounts explicitly for the unexpected results. Implicitly, it also reveals some potential for social experimentation in cyberspace.

Philosopher Ron Barnette is the proprietor of an engaging Web site called Zeno's Coffeehouse, whose patrons he invites to solve problems, resolve paradoxes, and participate in virtual experiments.[1] The proliferation of communicative links in cyberspace has facilitated a recrudescence of experimental philosophy; this study gives an account of one such experiment in virtual conflict and its resolution. It concerns Zeno's Coffeehouse's first challenge and its two subsequent trials, to which I refer collectively as "Zeno's Coffeehouse problem." What I term the "tragedy of the coffeehouse" is the common result of trials 1 and 2, which yielded the worst possible payoff to all players. That the tragedy is avertible is shown by the results of trial 3, which did not yield the best possible payoff but which (I claim) yielded the best attainable one.

The game itself is uncomplicated; its rules, invariant. Each player simply selects either box A or box B, in light of three possible outcomes, with payoffs in virtual dollars:

(O1) $1,000 to each player, if and only if all players choose box A;
(O2) $100 to each player who chooses box B, if and only if at least one fourth of the players choose box B;
(O3) $0 to each player, otherwise.

1. The URL of Zeno's Coffeehouse is http://www.valdosta.peachnet.edu/~rbarnett/phi/

AUTHOR'S NOTE: I am grateful to all the participants at Zeno's Coffeehouse for making the experiments possible, to Ron Barnette for his professional encouragement and technical support, and to the referees for the *Journal of Conflict Resolution* for their constructive criticisms of this report.

JOURNAL OF CONFLICT RESOLUTION, Vol. 43 No. 4, August 1999 434-450

TABLE 1
Data of the Three Trials

Trial	Number of Players	Number of A-Boxers	Number of B-Boxers	% A-Boxers (± 1 SD)	% B-Boxers (± 1 SD)	Outcome
1	156	134	22	85.9 (± 3%)	14.1 (± 3%)	$0 to all players
2	247	198	49	80.2 (± 2.5%)	19.8 (± 2.5%)	$0 to all players
3	34	4	30	11.8 (± 11%)	88.2 (± 11%)	$0 to A-boxers, $100 to B-boxers

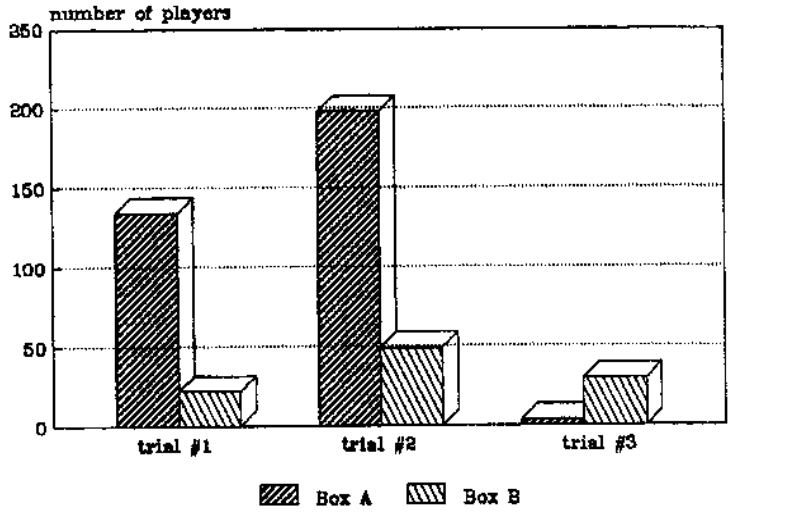

Figure 1: Zeno's Coffeehouse Problem: Histograms of Three Trials

In both trials 1 and 2, the players realized the worst possible outcome (O3). I then analyzed these trials and offered a prescription for future success. I predicted that rational players who read my analysis and prescription would avert the worst outcome if a third trial were held. These materials were posted at Zeno's Coffeehouse, and a third trial was subsequently held. Its results strongly confirm my prediction in one sense but also disconfirm it in another. This study elucidates that contrast. Figure 1 depicts a bar chart of the three trials; Table 1 shows their raw data, relative proportions, and associated outcomes.

ARGUMENTS FOR CHOOSING BOX B

Both probabilistic and causal arguments prescribe choosing box B. In my pretrial 3 analysis, I offered the following probabilistic argument only. We can interpret the

empirical frequency with which all players choose box A as an average probability with which any player chooses box A. Admittedly, this claim is contentious. Frequentists would certainly object that because nothing remotely like a limiting frequency has been attained empirically, the term "probability with which any player chooses box A" is without meaning (e.g., von Mises 1981). And although Bayesians hold that each player sustains some a priori personalist probability of choosing box A, they do not necessarily grant the premise that such probabilities are interpersonally averageable over a group. Nonetheless, the empirical frequency with which box A was selected in trial 1 (85.9%) remains consistent with the assumption that each player selected that box with an a priori average probability of .859.

If so, then the probability that all 156 players chose box A was $(.859)^{156}$, or about 5×10^{-11}. This denotes odds of 1 in 20 billion. And in trial 2, a still-smaller empirical frequency (.802) is exponentiated by an even larger number of players (247): hence, the probability that all players chose box A in this trial was $(.802)^{247}$, or 2×10^{-24}. These odds are equivalent to 1 in about 500 sextillion (1 in 5×10^{23}), which is the order of magnitude of Avogadro's number. So this probability corresponds to the chance of randomly picking one particular atom from a liter of an atomic gas at standard temperature and pressure. In either trial, the prudent (if retrodictive) inference is that the likelihood that all players chose box A was effectively zero. Hence, a vote for box A was a squandered vote.

For completeness: on the same assumption and in each trial, what were the probabilities that at least one fourth of the players chose box B? In trial 1, that probability was 2.2×10^{-4}. This is 4,400,000 times greater than the probability that all players chose box A. And in trial 2, it was 2.5×10^{-2}. This in turn is 1.3×10^{22} times greater than the probability that all players chose box A. These probabilities are required to maximize expected utilities. Because the payoff of (O1) is only one order of magnitude larger than the payoff of (O2), whereas the probability that (O1) obtained was six orders of magnitude smaller than the probability that (O2) obtained in trial 1—and 22 orders of magnitude smaller in trial 2—maximizing expected utilities overwhelmingly prescribes selecting box B.

In sum, although it is surely rational to prefer the best outcome (O1) to the second-best outcome (O2), it is surely irrational not to recognize that (O1) is probabilistically unattainable. In that light, it is rational to prefer a less rewarding but attainable outcome (O2) to a more rewarding but unattainable outcome (O1) because it is rational to prefer $100 to nothing. But in trials 1 and 2, a collective irrational preference for the unattainable albeit best outcome (O1) resulted in the actual attainment of the worst outcome (O3). On probabilistic grounds, box B is the only rational choice.

Ron Barnette posted a version of the foregoing argument at Zeno's Coffeehouse, along with my prediction that any rational player who understood it would choose box B in a subsequent trial. He then conducted trial 3 itself.

Before discussing the trial 3 results, I should like to advance a causal argument for choosing box B, intended for those (e.g., hard-line frequentists) who may discount the foregoing probabilistic argument. I now assume that each player's deliberation is shaped by some finite set of sufficient reasons for choosing one box or the other, somewhat analogous to a superposition of forces or vectors, whose "resultant" reason determines—whether decisively or reluctantly—his or her eventual decision.

Many players (evidently) posit a fallacious argument for choosing box A, which goes something like this: "Individually and collectively, all players would be better off choosing box A; therefore I should choose box A." That argument is unsound because some people may deem themselves "better off" not by dint of acquiring the largest monetary payoff, rather by preventing others from doing so. Conceivably, the thought of choosing box A might suppress some player's production of endorphins, but the thought of choosing box B might enhance that player's production thereof—in which case, that player might feel "better off" choosing box B, at least in the short hedonistic run. More precisely formulated, then, the naive argument for choosing box A runs, "Individually and collectively, if all players felt better off choosing box A, then I should choose box A." Unfortunately, because the rules of the game proscribe collusion, a player has no way of establishing the truth of the antecedent and thus finds no *modus ponens* to the consequent.

On closer consideration, it is rather the other prescription that obtains. A player may reason soundly that he or she cannot realize the best payoff unless everyone chooses box A and that not everyone chooses box A unless he or she chooses box A. But the player may not conclude conversely that by choosing box A, he or she will realize the best payoff because it simply does not follow that if he or she chooses box A, then everyone else will choose box A. Not all players necessarily deliberate the same way; moreover, it is obviously true that neither the player's deliberation nor his or her choice exerts any causal influence on the deliberations or choices of the other players. For all he or she knows, other players may flip coins or consult oracles to determine their choices. Because the player cannot assume the rationality of the other players—for there were no such criteria demanded for participation in the game—he or she must protect himself or herself against their potential irrationality (see von Neumann and Morgenstern 1944, 128ff.). Hence, box B is the only rational choice.

Thus, a player should choose box A if and only if he or she is certain that all other players will choose box A. Although a player who subscribes to causal determinism asserts that every other player will find sufficient reason for his or her respective choice, the putative mechanism of choice of each player is opaque to every other player. In consequence, no player can be certain that all other players will choose box A, and so every player should choose box B.

RELATIONS TO ASSURANCE GAMES AND WOLF'S DILEMMA

The prescriptions of both probabilistic and causal reasoning converge on box B. Why, then, in trials 1 and 2 did large majorities of players choose box A? As I indicated in my pretrial 3 analysis, some perhaps were motivated by the gambler's fallacy: that the wager with the largest payoff is best (regardless of relative odds). But I hypothesized that most who chose box A did so out of misguided cooperative predisposition, having mistaken the problem for a prisoner's dilemma (PD), which of course it is not. In the generic coffeehouse problem, assume N players altogether ($N \geq 4$), and let each

TABLE 2
Absence of Dominance in the Coffeehouse Problem

N − 1 Column Players → / *Row Player ↓*	*Box A,* M = 0	*Box A or Box B* 0 < M < (N/4) − 1	*Box A or Box B* M = (N/4) − 1	*Box A or Box B* M ≥ N/4
Box A	$1,000, $1,000	$0, $0	$0, $0	$0, $0 or $100
Box B	$0, $0	$0, $0	$100, $0 or $100	$100, $0 or $100

player view himself or herself as playing a game against $N-1$ others. Suppose M players (other than oneself) choose box B. Then every possible game state and its respective payoff is entailed by Table 2. Observe that the dominance principle does not apply in this game; there is no choice such that the row player is better off making it regardless of the other players' choices. Unfortunately, however, there is a Nash equilibrium in column 2—yielding nothing to everyone—which the players managed to discover in the first two empirical trials.

To anyone familiar with the definitive payoff structure of the PD, the difference in kind between it and the coffeehouse problem is obvious. So what kind of game is the latter? It appears to belong to the class of assurance games or is at least a variation on that theme. The assurance game resembles the PD in that it has a Pareto-optimal outcome, which obtains only when all players cooperate. But the assurance game differs saliently from the PD, in that defection is never unconditionally dominant over cooperation. In the generic assurance game, all players receive nothing if any player defects. Franzen (1995) offers the relay race as a social analogue; all runners must cooperate (i.e., strive individually) for the team to win. If any runner shirks, the team loses. Thus, the generic assurance game is not a social dilemma according to Harsanyi's (1977) definition because the absence of any incentive to defect leaves cooperation as the sole rational choice.

But now suppose we introduce such an incentive: this produces Wolf's dilemma. Wolf's original formulation (in Hofstadter 1985, 752-75) is as follows. Twenty players must decide, each in isolation from the others, whether to push a button. If no one pushes the button, each player receives $1,000. But if anyone pushes the button, then he or she receives $100, but those who refrain from pushing theirs receive nothing. This is indeed a social dilemma. We could find an unsalutary analogue in a corrupted relay team, whose racers are offered bribes for shirking. Setting moral implications aside for now, the decision-theoretic distinction between the generic assurance game and Wolf's dilemma is, as Franzen (1995) points out, that cooperation in the latter is no longer dominant. If any player in a Wolf's dilemma has good reason to believe that another player will defect, then his or her own defection becomes a rational choice (Franzen 1995). Thus, players may be impelled to a Nash equilibrium that is not Pareto optimal.

The empirical result of a Wolf's dilemma is conditioned chiefly by two factors: the relative temptation to defect and the absolute size of the group. For example, in Franzen's (1995) trials, with payoffs of 100 utiles to each player if all players cooperate versus 50 utiles guaranteed for individual defection, the chances of attaining Pareto optimality diminish rapidly as a function of group size: 32% for two persons, 15% for three, and 1% for five (Franzen 1995, 195). For groups larger than five, the chances rapidly approach zero. Note the similarity to the coffeehouse problem.

The structural difference between Wolf's dilemma and the coffeehouse problem is also clear. In Wolf's dilemma, the payoff for individual defection is unconditional—that is, independent of the number of other players (if any) who defect. But in the coffeehouse problem, it is conditional—that is, dependent on the attainment of a threshold proportion of defectors. The threshold at Zeno's Coffeehouse was set at 1/4 by Ron Barnette, who found empirically, in classroom settings, that this level of tension made the game interesting for participants.[2] In any case, it appears taxonomically defensible to classify Wolf's dilemma as an assurance game with unconditional temptation to defect and the coffeehouse problem as an assurance game with conditional temptation to defect. The generic assurance game, as we have seen, has no temptation to defect.

MISGUIDED MOTIVATION AND COSTLY RIDING

To reiterate: I hypothesized that most who chose box A in trials 1 and 2 did so out of misguided cooperative predisposition. Prisoner's dilemmas and free riding have been much discussed lately (e.g., Glance and Huberman 1994). Free riders attract moral censure (e.g., Pettit 1986), Pareto-optimal outcomes in such problems are attained through cooperation (e.g., Axelrod 1984), choosing box A appears superficially to be the analogue of cooperating (though it is not), and thus a well-intentioned majority chose box A.

Shubik and Wolf (1974) extend the PD matrix in such a way as to eliminate dominance and elucidate motivation. In the standard PD, a player must choose exclusively between cooperation and competition; there is no middle ground. But Shubik and Wolf introduce an additional stance, namely that of individualism. Their three-choice matrix allows players to distinguish between cooperative, individual, and competitive gain. Empirically, they discovered a range of conditions under which cooperation outweighs competition while individualism also outweighs cooperation.

This relates to the coffeehouse problem in the following way. As I hypothesized, most players chose box A over box B in the first two trials not only because they correctly perceived box A as a cooperative choice but also because they (mis)perceived box B as a competitive choice. "Better to cooperate than compete in a PD," they reasoned, unaware that this problem is not a PD. Here Shubik and Wolf's (1974) insight comes into play: had players interpreted box B as an individualistic choice instead of a competitive one, sufficient numbers could have made that choice and benefited from it (as they did in trial 3).

Consider the following three-box assurance game, which illustrates this point by amalgamating the PD with Wolf's dilemma and the coffeehouse problem. Each player chooses either box A, box B, or box C. If all players choose box A, then each receives a

2. Private communication, 1997.

large payoff. As usual, this is the Pareto-optimal outcome. If not all players choose A, then A-boxers get nothing. Any player who chooses box B receives a guaranteed but modest payoff. Any player who chooses box C receives a payoff proportional to the number of A-boxers and inversely proportional to the number of C-boxers. The potential range of box C's payoff would extend from the highest to the lowest in the game. If almost all players choose box A, whereas one (or just a few) choose C, then the payoff to each C-boxer would significantly exceed the Pareto-optimal payoff to each A-boxer. But if almost all players choose box C, whereas none (or one or just a few) choose A, then the payoff to C-boxers would be significantly less than the modest payoff guaranteed to B-boxers. This three-box assurance game clearly embodies Shubik and Wolf's (1974) distinction: box A is the cooperative choice; box B, the individualistic choice; box C, the competitive choice.

Although the foregoing model disambiguates players' motives, the coffeehouse problem presents a Wolf's dilemma in prisoner's clothing. Coffeehouse players who reflexively misconstrue their choice as between cooperation and competition in a PD choose box A; players who more thoughtfully construe their choice as between individualism and cooperation in a non-PD choose box B. The important lesson is that appropriate construal of a game is a prerequisite for effective play. Viewed in this light, the coffeehouse problem is problematic only insofar as it inadvertently ambiguates competitive and individualistic motives.

Recall my hypothesis that the relative few who chose box B in trials 1 and 2 of the coffeehouse problem realized full well the extreme improbability of attaining the best possible outcome (O1) and so settled rationally on the best attainable one (O2). The would-be "cooperative" majority actually compelled the worst possible outcome (O3). So our problem—in which a rational minority is undermined by a well-intentioned but misguided majority, resulting in the worst outcome for all—is deservedly called the "tragedy of the coffeehouse."[3] Accordingly, players who chose box A in this context are clearly "costly riders."[4]

STATISTICAL SIGNIFICANCE OF TRIAL 3

I predicted that if a third trial were held and most players were rational and consulted my analysis of trials 1 and 2, then at least one fourth (and very possibly many more) would select box B. But I also cited an important caveat—namely, Howard's (1971) "existentialist axiom" of metagame theory. It states that people are apparently free to falsify any prediction about their voluntary behavior made known to them in advance. The results of trial 3 confirm my forecast in a decidedly unexpected way, which lends more credence to Howard's axiom than to my powers of prognostication.

Recall that trials 1 and 2 yielded identical outcomes: nothing to everyone. Nor do these trials differ in any statistically significant way. The percentage of A-boxers in

3. This nomenclature, of course, follows from the "tragedy of the commons" (see, e.g., Hardin 1973).
4. This in contradistinction to free riders (see, e.g., Pettit 1986).

trial 1 is 85.9% ± 4.5%, giving a range of 81.4% to 90.4% across a confidence interval of three standard deviations.[5] In trial 2, the percentage of A-boxers is 80.2% ± 3.75%, giving an associated range of 76.5% to 84% across a like interval. The overlap of these ranges denotes their positive correlation. And note that even the minimum value of trial 2's statistical range (76.5%) lies above the critical 75% threshold, thus continuing to yield an outcome of nothing to everyone. Statistically, the population of trial 2 learned essentially nothing from the lesson of trial 1. It is this lack of statistical improvement that prompted me to wade in with my analysis and—with Ron Barnette's philosophical and logistical support—to instigate a third trial.

Trial 3's outcome is radically different: 30 of 34 players (88.2%) chose box B, thus gaining 100 virtual dollars each. The 4 players who chose box A gained nothing. The statistics are also telling: the range of B-boxers across a confidence interval of two standard deviations is 88.2% ± 11%, or 77.2% to 99.2%. At the lower end of this range, the percentage of B-boxers soars far above the critical 25% threshold, but at the upper end, A-boxers are all but extinct. On this statistical view, it appears that my prediction was vindicated.

But these proportions tell only part of the story. The three trials see a sizable increase followed by a drastic decline in successive populations of players: 156, 247, and 34, respectively. This pattern is mirrored (but with a more precipitous drop) in successive populations of A-boxers: 134, 198, and 4. But that symmetry is broken by successive populations of B-boxers: 22, 49, and 30. And the broken symmetry is statistically significant. For the three trials, the mean overall population is 146, with a standard deviation from the mean (i.e., a root mean square) of 87. The dispersion about this mean is so large that both trial 2 and trial 3 populations fail to lie within two standard deviations of it. The mean number of A-boxers for the three trials is 112, with a standard deviation from the mean of 81. Again, the actual numbers of A-boxers in trials 2 and 3 are dispersed beyond two standard deviations from their mean. Finally, the mean number of B-boxers for the three trials is 34, with a standard deviation of 11. Although the number of B-boxers in trial 2 lies just beyond a standard deviation from the mean (49, 34 ± 11, respectively), the number of B-boxers in trial 3 lies comfortably within one standard deviation of it (30, 34 ± 11). In fact, the number of B-boxers in trial 3 lies closer to its respective mean than in any other trial, whereas both the overall population in trial 3 and the number of A-boxers in trial 3 lie further from their respective means than in any other trial.

These statistics imply a rather interesting circumstance—namely, that although the number of A-boxers fluctuated considerably from trial to trial, the number of B-boxers remained relatively constant. And on this statistical view, the B-boxers prevailed overwhelmingly in trial 3, not because significant numbers of former A-boxers saw the light and switched to B (as I argued they should and predicted they would), but rather because significant numbers of former A-boxers simply abstained from voting in trial

5. A standard deviation is calculated as $[f(1-f)/n]^{1/2}$, where f is the observed frequency and n the population size. A range of two standard deviations produces a confidence interval of 95%; of three standard deviations, a confidence interval of 99%.

3 and abandoned the field wholesale to the Bs. This more-or-less stable group of B-boxers, which had been overwhelmed by A-boxers in trials 1 and 2, now found numerical supremacy by default. As Figure 1 depicts, the opposition did not switch; it vanished.

Assuming that my pretrial 3 analysis had the dual effects of reconfirming to former B-boxers the reasonableness and defensibility of their choice and confronting former A-boxers with the unreasonableness and indefensibility of theirs, the question is the following: why did the former A-boxers abstain en masse rather than switching to and reaping the virtual rewards of box B?

WHAT SOME PARTICIPANTS SAID

I will offer an answer to this question in due course. But first it is instructive to examine some revealing ratiocinations submitted by participants in trial 3. The first group of quotations was volunteered by players who obviously read and understood my analysis of trials 1 and 2:

> "In the light of all the convincing theorizing that has been going on about this challenge, I think that the probability of all participants choosing A is closer than ever to nil. I will therefore obediently join the statistical herd and vote for box B."
> "Box B all the way. I chose this before, btw, for similar reasons to those espoused by the prof."
> "And even though I didn't bother reading that big long explanation, I just assumed that common sense would dictate that not everyone would pick the same thing, thereby making box A a rather stupid choice."
> "OK, I agree with Marinoff's analysis. I'll choose B. . . . This is a fascinating twist to the problem and if Marinoff is correct (as I believe he is) should result in a dramatic shift in behavior. If so then this will show the power of education in that without the analysis, a repeat of the former tragedies would have been likely."
> "Look, there are always a few a******s around—people who just want to be contrary. If we vote A, they'll vote B, and everyone gets screwed (no one gets any money). If we vote B, they'll vote A, and the only ones who get screwed will be them (we all get $100, and they get nothing). That means the best bet for the rest of us (non-a******s) is to vote B. Don't let the antisocial people cheat you out of $100. . . . A is for A******s. Vote B."
> "I am choosing Box B, and I am only thirteen."

A blessing on the 13-year-old: wisdom is not the province solely of the aged. In addition to the evident majority of B-recidivists and wise teenagers, a precious few former A-boxers saw the light and switched to B:

> "I just voted B and wanted you to know that your analysis . . . of my motives on the earlier vote was right on the mark. Both the maximum payoff and the desire to be a good guy motivated my choice. Knowing that being a good guy is consistent with B and that A is a highly improbable payoff have changed my vote. This time it's truly *enlightened* self-interest."
> "In light of the new information, Box B is my choice. Let's hope the results of this one will bear fruit."

> "I choose Box B, but I doubt that the (no doubt) triumph of B means very much, since the (excellent) analysis serves as an opinion leader undermining the cooperative instinct which supports Box A (to put it another way, it raises the probability of not choosing A so high that even the most altruistic can't ignore it)."

Note again the unfortunate habit (conditioned social reflex?) of associating box A with cooperative (and therefore ostensibly "good") behavior and box B with defective (and therefore ostensibly "bad") behavior. The comment lauds my analysis while missing its point. Then again, there are always some who do the right thing for the wrong reasons:

> "I choose Box B—a real no brainer as Marinoff has already stated HE is going to choose Box B and hence not everyone is choosing Box A. . . . However, if Marinoff is not playing the game, I would go with Box A."
> "Disagreeing with Marinoff's analysis, the only rational choice on a third try is A. Unless, of course, as is the case, we know that Marinoff himself will vote for B, guaranteeing that A cannot be the outcome. His analysis, then, whether right or wrong, changes the situation in a fundamental way. Hence I must vote for B."

Nowhere in my analysis did I state that I would participate in the third trial and vote for B. I merely tried to persuade others that B is the only rational choice in any hypothetical trial. Meanwhile, the next two quotes show respectively that a select few usually manage to do either the wrong thing at the wrong time or the wrong thing for the wrong reasons:

> "I chose B the first time. I chose B the second time. I choose A now because I think the morons who ruined my gain twice do not deserve a third chance. So another (O3) would actually make me feel better (be a bigger gain) than a virtual 100 bucks."
> "If Marinoff is not playing the game, I would go with Box A. Reason: If anyone does not choose Box A, then all who did would be made losers by the one who did not choose it. Via the golden rule—a classical [sic] rule for decision making—I choose Box A because that is what I would wish everyone to do for me."

The penultimate quotation epitomizes creative self-destructiveness. A desire for revenge has ironically backfired. The desire itself seems fuelled by an assumption that former A-boxers are "morons." They have elsewhere been called "good guys," "antisocial," "altruists," and "a******s." Exactly what social or moral epithets they merit—if any—remain to be evaluated.

The last quotation brings us full circle, for it reasserts the fundamental fallacy that my pretrial 3 analysis intended to expose. As alternatives to the golden rule, there are any number of "classic" decision rules that one could invoke, such as Kant's (1898) categorical imperative (do unto others only that which you could will that everyone do), Hobbes's (1957) contingent contractarianism (do unto others that which your prudence dictates they will do unto you), Spinoza's (1958) free-ranging egoism (do unto others whatever suits you), Machiavelli's (1988) preemptive survivalism (do unto others before they do unto you), *Old Testament* (1960) retributive sanction (do unto others as they have done unto you), Buddha's version of karma (do unto others as you would do unto yourself, for so you shall; *Dhammapada* 1980), and Talmudic pragmatism (do

unto others as they do unto themselves). The task of decision theory is not blindly to apply some oft-cited principle for the sake of its application; rather, it is to ask meta-theoretical questions about what kinds of principles are applicable—or inapplicable—in given situations.

INTRINSIC AND EXTRINSIC MORAL CONTENT

In the wake of trial 3 and with a view to accounting more fully for its results, two critical questions need answering. First, do the choices in Zeno's Coffeehouse problem possess intrinsic (i.e., first-order) moral content? Second, does the game itself possess extrinsic (i.e., second-order) moral content? I shall argue that although the answer to the first question is negative, most A-boxers supposed it to be affirmative. That is, they ascribed moral content to the choices even though no such ascription is entailed by the rules. Moreover, I hypothesize that although A-boxers ascribed no extrinsic moral content to the game in trials 1 and 2, they pejoratively ascribed such content prior to trial 3, which caused them overwhelmingly to abstain rather than to reselect box A or switch to box B.

To appreciate the ostensive distinction between extrinsic versus intrinsic moral content, consider for example the game of roulette. Roulette can possess extrinsic moral content. It belongs to the class of gambling games and, as such, can be proscribed—or, for that matter, enjoined—by any number of religious or secular moralities. In consequence, would-be roulette players first consult their respective moral codes to determine whether playing that game is deemed morally permissible, morally impermissible, or morally neutral by their respective ethical lights.

But roulette possesses no intrinsic moral content because the colors and the numbers that constitute the players' possible choices neither denote nor connote moral rightness or wrongness. Even the number 13, which some deem unlucky, is not therefore morally unsavory. And although colors play an undeniable role in denoting and connoting things susceptible to moral judgments—such as red-light districts and black markets—those same colors on the roulette wheel are normally innocent of such associations. Thus, roulette possesses no intrinsic moral content. In consequence, roulette players consult no theories of ethics to abet the placement of their wagers.

Does the generic PD possess any extrinsic moral content? Apparently not: hundreds of social scientific experiments have been conducted with PDs, involving thousands of voluntary subjects with diverse religious, ethical, and cultural orientations. There has come to light neither any ethos endorsing participation on the grounds that the PD is a virtuous game nor any ethos proscribing participation on the grounds that the PD is a vicious game. Thus, the generic PD nominally possesses no extrinsic moral content.[6]

6. But many scenarios modeled as PDs can indeed possess extrinsic moral content. For example, the cold war nuclear arms race can be viewed as a two-player PD and can be held to have been a "good" or a "bad" thing.

Does the PD possess intrinsic moral content? I claim that although in specific instances it could, generically it does not. In the generic PD, two suspects are held incommunicado by the authorities. Each must choose between keeping silent or informing on the other. If both keep silent, both are released. If one informs on the other, one receives freedom and a monetary reward; the other gets a long jail term. If each informs on the other, both receive short jail terms. Thus, the so-called "cooperative" choice actually means cooperating with the other suspect and not with the authorities. If one suspect knows the other to be guilty and "cooperates," is he not then withholding evidence, obstructing justice, and thereby behaving immorally as well as unlawfully? Then again, if one suspect knows the other to be guilty and "defects," is he not then giving useful testimony, serving justice, and thereby behaving morally as well as lawfully? If so, then cooperation is morally wrong, whereas defection is morally right; cooperators are bad, and defectors are good. This generic counterexample surely repudiates the egregious assumption that unqualified cooperation is always "good" in a PD. Now suppose that the suspects are freedom-loving activists held in the clutches of a totalitarian secret police. That one suspect knows the other to be guilty or innocent (say of "subversive political activity") may now be irrelevant to his or her deliberation; given the greater evil against which he or she fights, unqualified cooperation (i.e., refusal to give or to fabricate evidence against his or her fellow prisoner) might after all be the morally defensible choice. My point is that the generic PD is ethically underdetermined and therefore morally incoherent.

We can introduce intrinsic moral content into a noncooperative game if we permit collusion between or among the players. For example, suppose the suspects in a two-player PD are allowed to collude. Not unreasonably, they both agree that cooperation is their best joint strategy. Each affirms that he or she promises to cooperate if the other so promises. But when they are reseparated and each suspect again faces an individual choice, each finds—as Rapoport and Chammah (1965) have ably pointed out—that the dilemma persists. Instead of deciding whether to cooperate or defect, each player must now decide whether to keep his or her promise and, depending on what kind of ethical theory he or she brings to bear, the decision may or may not be conditioned by whether he or she thinks the other will keep his or her promise too. In any case, such a decision is clearly fraught with moral content; conflicting ethical theories prescribe different moral choices (e.g., see Marinoff 1994). Similarly, if one had allowed collusion in Zeno's Coffeehouse problem, then on the assumption that all players would have promised to choose box A, their subsequent choices would have possessed intrinsic moral content as well.

To explain what caused the trial 3 results in Zeno's Coffeehouse problem—and, at the same time, what motivated the results of trials 1 and 2—I offer the following hypothesis. I suppose that no players initially attributed extrinsic moral content to the game itself but that most players who chose box A in the first two trials attributed intrinsic moral content to the possible choices. As I supposed prior to trial 3, most players who chose box A in the first two trials did so out of misplaced cooperative predisposition. They associated with box A a morally "good" choice; with box B, a morally "bad" choice. They preferred making the ostensibly "good" choice and receiving a virtual payoff of nothing to making the ostensibly "bad" choice and receiving a virtual

TABLE 3
A-Boxer's Attribution of Intrinsic Moral Content

N − 1 Column Players → / *Row Player ↓*	*Box A,* M = 0	*Box A or Box B,* 0 < M < *(N/4) − 1*	*Box A or Box B,* M ≥ *N/4 − 1*
Box A	Moral comfort + $1,000	Moral comfort + $0	Moral comfort + $0
Box B	Moral discomfort + $0	Moral discomfort + $0	Moral discomfort + $100

payoff of $100. In contrast, those who chose box B attributed no intrinsic moral content to the possible choices; they merely sought to obtain the best attainable outcome and so deemed a virtual payoff of $100 preferable to a virtual payoff of nothing. Specifically, the former group seems to have smuggled into the matrix additional payoffs of moral comfort (attained, they supposed, by choosing box A) and moral discomfort (attained, they supposed, by choosing box B). The associated payoff matrix is depicted in Table 3. But the rules of Zeno's Coffeehouse problem—by which it, like any game, is defined (see von Neumann and Morgenstern 1944)—do not entail payoffs of moral comfort or discomfort. Albeit implicitly, innocently, or inadvertently, the A-boxers reified a matrix containing extraneous payoffs. In consequence, on this hypothesis, those who chose box A in the first two trials implicitly, innocently, or inadvertently violated the rules: in effect, they were playing a different game.

So what happened in trial 3? The data suggest (and the players' comments corroborate) that roughly the same group of B-boxers from trials 1 and 2 consistently choose box B again. If anything, my analysis and prediction reconfirmed for them the rationality and therefore the game-theoretic "correctness" of their choice. But what happened to all the A-boxers? In raw numbers, their constituency plummeted from 134 and 198 (in trials 1 and 2, respectively) to only 4 in trial 3—and this without significant reversions to box B. Why?

I hypothesize that A-boxers in trials 1 and 2 smuggled into the game an (unwarranted) ethical commitment to choose box A, from which they reified and expressed preference for a payoff of moral comfort at any cost. In other words, they did not attribute extrinsic moral content to the game but did attribute intrinsic moral content to the choices. Prior to trials 1 and 2, no one informed them (and they obviously did not conclude for themselves) that in the context of this game, such ethical commitment is superfluous. I further suppose that for most A-boxers, this ethical commitment was grounded in their passions and not in their reason. In its most unexamined formulation, such grounding asserts that cooperation in a PD is altruistic, defection is egoistic, and moreover that altruism is good and egoism is bad. It follows validly but unsoundly for A-boxers that cooperation is good, and—perhaps perversely—the costlier the cooperativeness, the better. My pretrial 3 analysis made it clear that Zeno's Coffeehouse problem is not a PD, that choosing box A is not analogous to cooperating (and is thus not "good" but simply irrational), and that choosing box B is not analogous to defecting (and is thus not "bad" but simply rational). But that which is transparent to reason remains opaque to passion.

TABLE 4
The A-Boxer's Dilemma

N − 1 Column Players → / *Row Player ↓*	*Box A,* M = 0	*Box A or Box B,* 0 < M < *(N/4) − 1*	*Box A or Box B,* M ≥ *N/4 − 1*
Box A	Moral comfort + rational discomfort + $1,000	Moral comfort + rational discomfort + $0	Moral comfort + rational discomfort + $0
Box B	Moral discomfort + rational comfort + $0	Moral discomfort + rational comfort + $0	Moral discomfort + rational comfort + $100

TABLE 5
The A-Boxer's Avoidance

A-Boxer	*Outcomes*
Participate in trial 3: choose either A or B	Appear either irrational or immoral
Abstain from trial 3: choose neither A nor B	Appear neither irrational nor immoral

Perhaps more contentiously, I further suppose that A-boxers' passions are saturated and animated by prevailing politicized social theories, which assert—self-servingly but egregiously—that societal evolution owes more to cooperation than to competition. If my hypothesis is correct, then trial 3 placed former A-boxers on the horns of a dilemma: if they reselected box A, they would—in light of my reasoned pretrial analysis—appear profoundly irrational. If they selected box B, they would—in light of their impassioned belief in the primacy of cooperation—appear profoundly immoral. Their putative dilemma is illustrated in Table 4.

Unwilling to impale themselves on either horn, I suppose that they made a second-order choice not to choose at all. That is, their hypothetical first-order dilemma obliged them to attribute extrinsic (second-order) moral content to the game itself. Not wishing to appear irrational (by choosing box A) or immoral (by choosing box B), they deemed abstention preferable to participation. This is illustrated in Table 5. That completes my account of the results. Because decision theory is underdetermined by data, I invite other theoreticians to posit more plausible explanations.

CONCLUSIONS

One B-boxer expressed the view that people's behaviors differ significantly in experiments with virtual payoffs versus experiments with tangible ones:

> "Box B—Yeah, Yeah, you've convinced me in the terms of this test, but in real life, with real money, would anyone vote for anything but Box A?"

This is an interesting and important question, bound to elicit as much difference of opinion as did the experiment from which it emerged. I would have framed the rhetoric just the other way around: with real money, would anyone vote for anything but box B?

Over the course of several years and in several institutions, I have played the following game with many and varied undergraduate philosophy classes: choose either to gain x dollars unconditionally or to gain and risk x dollars for a potential 10:1 payoff on the toss of a fair coin. In the latter case, if "heads" comes up, then one forfeits x dollars; if "tails," then one wins $10x$ dollars. (Not surprisingly, no student ever attributed either extrinsic or intrinsic moral content to this game.) Although the mathematical expectation of the second choice is always fivefold greater than that of the first for any x, it is my uniform experience that the resultant proportion of choice is conditioned strictly by the hypothetical amount of money (i.e., by its value to the players) that one substitutes for x. Typically, all students are willing to risk \$1 to gain \$10; most are willing to risk \$10 to gain \$100; fewer are willing to risk \$100 to gain \$1,000; very few indeed are willing to risk \$1,000 to gain \$10,000. If one continues to increase x, then one eventually reaches some "ceiling" amount (or least upper bound) such that no one is willing to risk it. My experience is with virtual payoffs only.

Now I offer a minor and a major conjecture. The minor conjecture is that the same kind of result would obtain if the game were played with a group of paupers, billionaires, or a random assortment of players of varying wealth. The only difference would be in the degree of the result—that is, in the values of x that would yield given proportions of choice—which in turn depends on the relative evenness or disparity of wealth within the group. The ceiling depends ultimately on the absolute wealth of the group's wealthiest member. I repeat that these differences are of degree only, not of kind.

The major conjecture is that if the same experiment were performed with real money, players in any group would behave either identically or more conservatively (relative again to their respective means). I expect that the values of x that would yield given proportions of choice would either remain fairly constant or would diminish slightly. For example, I expect that some students willing to risk 10 virtual dollars for 100 virtual dollars would not be willing risk 10 real dollars for 100 real dollars; moreover, all students unwilling to risk 10 virtual dollars for 100 real dollars would be unwilling to risk 10 real dollars for 100 real dollars. In other words, I expect that the substitution of real payoffs for virtual ones would have the effect of making the players more prudent, that is, more apt to take what they are guaranteed to get. Although this experiment should be performed, it is unlikely to attract necessary or sufficient funding.

The aforementioned game is neither cooperative nor competitive; rather, it is "acooperative": its payoff to a given player is not a function of the other players' choices. But what of noncooperative games in general and Zeno's Coffeehouse problem in particular, in which the payoff to a given player is a function of the other players' choices? I submit that if the transition from virtual to real payoffs in an acooperative game increases the prudence of the individual players (which *ex hypothesi* it does), then a similar transition in a noncooperative game would even more markedly increase their individual prudence, provided that they do not ascribe intrinsic moral content to their choices. Imagine playing Monopoly with real money and real estate (or simply imagine living in the United States): do the players (i.e., property owners, home owners, hotel owners, landlords, bankers, etc.) exercise more prudence in the virtual game or in the real one? This question must remain unanswered in the case of Zeno's Coffeehouse problem until experiments with real payoffs are funded.

But in the absence of such experiments, I offer one last conjecture. If trials 1 and 2 of Zeno's Coffeehouse problem were reconducted using real payoffs and a new population, I expect that similar results would obtain: no one would win anything. Given a similar pretrial 3 analysis, most A-boxers would once again experience their peculiar dilemma, would thus ascribe extrinsic moral content to the game, and so would consider abstaining. But given real payoffs, I conjecture that most would temper their potential abstinence either with naked avarice, a sudden reluctance to squander real opportunity, or a desire to reassert a vain principle at a palpable cost to all: hence, most would participate in trial 3. The bigger the payoffs, the greater the participation. And what would they elect to do? I submit that although some A-boxers would see the amoral light and switch to B, most would continue irrationally to imbue their choices in this game with intrinsic moral content and so impale themselves on that horn of the dilemma that offers them moral comfort but deprives everyone else of that which is readily attainable.

On the least charitable interpretation, choosing box A is neither cooperative nor altruistic; rather, it is counterproductive and selfish. However it may appear at first blush, Zeno's Coffeehouse problem does not pit duty-minded Kantian deontologists against consequentialist maximizers of expected utility—for A-boxers are not Kantians. The A-boxers' maxim is not "choose such that you could will your choice to become universal law"; rather, it is "choose such that if everyone chooses like you, then all will gain the biggest payoff, and such that if not everyone chooses like you, then at least you will feel moral comfort." The A-boxer maxim is profoundly un-Kantian because it is concerned solely with consequences; a player in this game cannot but choose in light of desired or anticipated outcomes.

In sum, I submit that the tragedy of the coffeehouse is manufactured from a moral hallucination, itself derived from a self-gratifying but otherwise dysfunctional ethos that equates the good with the unattainable. Beneath the sound and fury of their cooperative and altruistic professions, A-boxers resemble nothing if not would-be social insects. Although hymenopterans (i.e., ants, bees, and wasps) indeed behave cooperatively and perhaps even altruistically (see Hamilton 1964), such behavior itself is predetermined and compelled by genetic programming. Social insects are therefore amoral, for they cannot behave otherwise. And notwithstanding its cooperativeness and ostended altruism, the body politic of the social insect is utterly totalitarian. I have argued that unrepentant A-boxers have erroneously deemed their choice cooperative and altruistic and have fallaciously identified it as morally good. Given political power, I suspect they would coerce all players into choosing as they do. In that case, everyone would indeed realize the biggest payoff but at the greatest political cost: all would find themselves at the furthest conceivable remove from individual liberty, social cooperativeness, and moral goodness alike.

One invites the tragedy of the coffeehouse by allowing costly riders to cast unreasoned votes, which results in an unexamined choice between box A and box B. One averts the tragedy of the coffeehouse in one of two ways: either by totalitarian coercion

of the electorate, with box A alone on the ballot (which entails worse tragedies than it prevents), or by reasoned—if unpopular—persuasion of the electorate, which results perforce in an examined choice between voting for box B and casting no vote at all. The decision-theoretic implication for democratic politics is clear: elections turn out better for all when the costly riders abstain.

REFERENCES

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Dhammapada. 1980. Translated by H. Kaviratna. Pasadena, CA: Theosophical University Press.

Franzen, A. 1995. Group size and one-shot collective action. *Rationality and Society* 7:183-200.

Glance, N., and B. Huberman. 1994. The dynamics of social dilemmas. *Scientific American* 275 (3): 76-81.

Hamilton, W. 1964. The genetical theory of social behavior I & II. *Journal of Theoretical Biology* 7:1-16, 17-62.

Hardin, G. 1973. The tragedy of the commons. In *Heredity and society*, edited by A. Baer, 226-39. New York: Macmillan.

Harsanyi, J. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge, UK: Cambridge University Press.

Hobbes, T. 1957. *Leviathan*. Oxford, UK: Basil Blackwell.

Hofstadter, D. 1985. *Metamagical themas*. New York: Basic Books.

Howard, N. 1971. *Paradoxes of rationality: Theory of metagames and political behavior*. Cambridge: MIT Press.

Kant, I. 1898. *Fundamental principles of the metaphysics of morals*. Translated by Thomas Abbott. London: Longmans, Green, and Co.

Machiavelli, N. 1988. *The prince*. Cambridge, UK: Cambridge University Press.

Marinoff, L. 1994. Hobbes, Spinoza, Kant, highway robbery and game theory. *Australasian Journal of Philosophy* 72:445-62.

Old Testament (according to the masoretic texts). 1960. Philadelphia, PA: Jewish Publication Society of America. *Exodus* 21:22-25.

Pettit, P. 1986. Free riding and foul dealing. *Journal of Philosophy* 83:361-79.

Rapoport, A., and A. Chammah. 1965. *Prisoner's dilemma*. Ann Arbor: University of Michigan Press.

Shubik, M., and G. Wolf. 1974. Solution concepts and psychological motivation in prisoner's dilemma games. *Decision Sciences* 5:153-63.

Spinoza, B. 1958. *The political works*. Translated by R. Wernham. Oxford, UK: Oxford University Press.

von Mises, R. 1981. *Probability, statistics and truth*. New York: Dover.

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. New York: John Wiley.