

HARRIS SCHOOL WORKING PAPER
SERIES 06.05I

**CHAPTER 7: ANOMALIES IN EXPERIMENTAL
ECONOMICS ***

Howard Margolis

**This is a draft chapter from Cognition and Extended Rational Choice
(forthcoming Routledge 2007).*

FIGURE 1

-
- 1a. competition (markets)
 - 1b. competition (games)
 - 2a. cooperation (weak free-riding temptation)
 - 2b. cooperation (strong): coordination
-

Response contexts that might guide an individual's intuition in a social interaction.

Market competition (1a) is a situation where each agent is concerned only about own payoff. Games competition (1b) is the quite different context in which a person is motivated by how well he is doing relative to other players. In a market context, 10 for me is better than 9 for me, no matter how that interacts with your payoff. But in a games context, 10 for me and 10 for you is only a tie: I prefer 1 for me, 0 for you, which is a win. In either context, specifically NSNX effects are not involved: an agent is concerned only with own-payoff, either absolutely (1a) or relative to other agents (1b), as will occur in many real situations even if motivation is NSNX.

And for the 2nd pair:

Weak cooperation (2a) is where an agent sees the context as an opportunity for mutually advantageous choices, but subject to free-rider concerns, so that even an agent who would want to cooperate might not, since he does not want to be exploited should too many others free-ride. Strong (but not always easy) cooperation (2b) is where the player sees the problem as coordinating choices to get a good result with free-riding either not even an available option or where it seems so unlikely in the context that it is not a complication for the chooser (as in a transaction with a known party and a well-established pattern of cooperation). Even for context 2b, sometimes cooperation might still be risky, as when communication is difficult and finding a mutually preferred choice seems unlikely. But the difficulty is not due to a working conflict of interest across players.

I will use the label *frame* to refer the subjective perception, in distinction from *context*, which refers to the objective situation. The heart of the discussion here will concern the possibility that the *frame* governing a choice may not match the context even when in hindsight there is a puzzle about how the chooser could have missed that.

Mixed cases (e.g., a context where a common interest in coordination but conflicting interests in the coordination point) are common. A buyer and seller have a common interest in making a deal but a conflicting interest in the price to agree on. But for the situations discussed in this chapter, the mixed cases can be set aside, though they will obviously be prominent in applications

Figure 1 shows how a person might understand a social context, or the judgment a person might reach about how someone else is understanding that context. When I write about social perceptions and misperceptions in this chapter, I mean the subjective framing of social context within the taxonomy of figure 1. On the evidence of Chapter 6, it makes sense to consider the possibility that social intuitions, not just individual intuitions, might sometimes be governed by what turn out to be adverse defaults. Parallel to what was seen in the context of responses to logically simple but cognitively difficult puzzles, misunderstandings as well as accurate understandings of social situations might occur. That would open the door to difficulties in reaching or sustaining cooperation which are invisible to the parties involved, as in Chapter 6 neglect defaults invisible to choosers can sometimes yield choices that the chooser herself, on reflection, comes to see as transparently unsound.

In the individual choice context, choosers whose only motivation is to get a simple logical puzzle right get it wrong. In the social context, both motivations and situations are far more complicated. NSNX choosers would want to be "neither selfish nor exploited". But choice within a natural setting, not a set-piece puzzle, usually can't be reduced to simple logic, even in a simple context like choosing among alternatives on a routine trip to the grocery store. But the situations that concern us here are where choices affect other people as well as the chooser, and where how others are choosing, or might be choosing, can change what our chooser would want to do.

We are interested, in particular, in cases where parties to a social interaction in which everyone could gain from cooperation, and in which everyone would prefer cooperation, might nevertheless act in ways that lead to failures of cooperation. The source of the difficulty might lie in misperceptions that, parallel to the neglect defaults of Chapter 6, could be entirely out of sight of the parties involved. And what I want to show in this chapter is evidence that adverse defaulting parallel to what was seen in Chapter 6, but here among social contexts in the taxonomy of figure 1, indeed can be a source of serious difficulties, and where the cognitive difficulties here turn out to be usually linked to the neglect defaulting which was the concern in Chapter 6.

Even without this possibility of covert difficulties, cooperation might be harmed by noisy communications, unreliable feedback, lack of commitment mechanisms, and so on. I might misunderstand your actions, or your intentions even if there is no misunderstanding the social context. All of these possibilities have been much discussed. But if covert difficulties are involved I might even misunderstand my own actions. It is that more surprising possibility that is the special concern of this chapter.

For the puzzles of Chapter 6 almost everyone eventually comes to see the usual responses as illusory. We then have very good reason to look for some covert process that could account for how people who are certainly not stupid are prompted to firm intuitions that are certainly not smart. I want to find a counterpart to that stark situation, where responses from intelligent subjects seem to make no sense given the situation they are actually in. In particular, we are interested in experiments where we could expect subjects to be able to do well by cooperating but instead they do badly. And to make the problem clearly visible we need to start with a special subset of such experiments in which managing cooperation should be transparently easy. .

Cooperation experiments usually involve games in which if players could communicate freely and make binding agreements cooperation would be indeed be easy. But cooperation is then made hard (since otherwise the results are likely to be uninteresting). Typically, players have never met, and are not allowed to communicate. Each will receive her payoff privately, so that a player need not be concerned with being confronted by others in the game, and everyone knows that. No side-payments or binding agreements of any kind are allowed.

Despite all this, in a context where cooperative choice would benefit everyone, a NSNX agent should still respond to some measure of other-regarding concern. Given the conditions that is likely to be cautious and partial, and perhaps even invisibly small. But what we see should be choices that could look reasonable to an agent concerned with being "neither selfish nor exploited". In the taxonomy of figure 1, the context is clearly that of weak cooperation (2a). Since NSNX actors respond to social values as well as to self-interest, if NSNX is right there is something to be explained if that is entirely missing.

Accounting for such puzzles will come up very prominently in later chapters. But for the argument of this chapter, I want to set aside any commitment to social motivation. We want to identify a special category of experiment which provides puzzling results that do not depend on whether or how far subjects respond to NSNX motivation or any other notion of other-regarding motivation. This requires experiments where self-interest coincides with, rather than competes with, any tendency to social motivation. Then whatever motivation is plausibly in play, a cognitive puzzle needs to be resolved.

The qualifier "plausible" is needed since some sort of rationale can be concocted for any choice whatever. Perhaps the players are really stupid, or masochists, or sadists, or neurotic, or just trying to confuse the experimenter. Perhaps they think it would be nice to deliberately keep their payoffs down to save the experimenter money. Such things might happen, conceivably even the last, but not often enough to be plausible explanations when a large fraction of responses have to be explained that way. If we can find experiments in which self-interest, or group-interest or any compromise in between would all, on any plausible account, yield the same result, but the result observed is far from that, then we will have what we need. We can use such results to explore whether, and if so how, adverse defaulting might extend beyond the individual choices considered in Chapter 6 to the social contexts that are our real concern. And, perhaps surprisingly, it is not hard to find experiment of the special sort we are looking for.

In a standard Public Goods experiment, players choose how much of an endowment to contribute to a common pool. The tokens in the pool are then multiplied and divided among the players equally. In a standard game, the multiplier is > 1 but less than the number of players, so what an individual gets back from his own contribution is less than he gave, but the amount shared across the group is more. So the group does best when all contribute fully, but each individual can profit by free-riding on the contributions of others. But starting with Andreoni (1995), a line of Public Goods experiments have been manipulated to eliminate any rational basis for contributing. The robust result has been that what Andreoni labeled "confusion" plays a large role in player choices, since in the degenerate variants contributions remain about half of what they are in an otherwise identical actual Public Goods game. Andreoni's alteration was to add a paragraph to his instructions which explained that tokens earned in the pool were used

only to rank players. Actual cash payoff then depended solely on rank. This converts what looks like an experiment about cooperation into a strictly zero-sum game. With payment strictly by rank, contributions can only lower a player's rank relative to others in his group.

Since payoff depends only on rank there is no point at all to concern about how many tokens are earned. So unless you are the sort of person (assuming there actually is that sort of person) who likes to lose in order to make his opponents better off, it makes no sense whatever to contribute. In terms of the taxonomy of figure 1, Andreoni changed the context of weak cooperation (2a) to strong (zero-sum) competition (1a). A player can only improve her own payoff by reducing other players payoffs. And players should not feel uncomfortable about that. After all, that is the way games are usually played.

Later experiments (Houser and Kurzban 2002, Ferraro & Vossler 2005) were even more stark, since there were not even other players. Rather, subjects were explicitly told, with quite elaborate precautions to make sure this was understood, that others in their group did not actually exist. Responses from others in their group would be generated by a computer which mechanically reports pseudo-contributions. A player would receive his own payoff from the pool (from his own contribution plus the pseudo-contributions). But there were no other earnings, no other actual players, and no connection between the mechanical pseudo-contributions and the actual contributions by the human subject.. If the return was half a token per token in the pool, the only consequence to a player of contributing a token was just to throw away half a token from his payoff.

In these games with "robot" partners, the more a player contributes, the less he earns and no social value of any kind exists to qualify that stark situation. A post-game survey confirmed that players indeed understood that robot earnings did not actually exist. Nevertheless, as in Andreoni's game, contributions in this degenerate game were around half of what they were in a parallel actual Public Goods game. Again these excess contributions were interpreted as due to confusion.

But the responses to the puzzles of Chapter 6 could also in some sense be ascribed to "confusion", which I put in scare quotes since for the puzzles no one is likely to doubt that something more systematic than mere confusion is involved. In the puzzles, subjects

have confident, widely shared intuitions about the right answer, and those confident, widely-shared intuitions are demonstrably wrong. For it is simple to set up physical situations where choices for the 3-cards or Monty Hall problems can be run through many iterations in a few minutes. So we can easily generate data sufficient to satisfy any reasonable demand for a statistically significant demonstration. A skeptic does not need to grasp the logic. He can see what in fact happens. But if there is something systematically misleading is going on in a degenerate Public Goods game it will be harder to see. There is no *particular* inappropriate response that attracts predominant but unsound support. Given n tokens, there are $n+1$ choices available (from 0 to n), and usually with multiple rounds of play. So while non-zero responses in the degenerate games are always unsound, player choices vary across rounds and individual choices vary within each round, so things indeed look confused, as if subjects just try out various choices to see what happens, or just blindly imitate what they see from other players.

Yet if the odd results of Public Goods games with a degenerate twist indeed revealed that choices in Public Goods games were in large part *merely* confused, some robust characteristics of Public Goods data would be very mysterious. One that will play a role in some data we will consider in Chapter 8 is the “restart” effect.. The games typically run 10 rounds. And it was Andreoni who found that contrary to clearing up confusion if the players were then invited to play the game over, they would behave much the same way as in the original 10 rounds. This has proved to be a very robust feature of Public Goods games. On the “confusion” hypothesis, we would have to say (among other implausible things) there is a very reliable renewal of “confusion” if players are given the opportunity to play again, this time immediately following 10 rounds of experience. But if what accounts for choices that make no sense when the game is degenerate is not confusion, then what is it?

The degenerate Public Goods experiments provide a starting point for considering that. But we are especially interested in this study in cooperation, so we especially want data in which the “confusion” or whatever else it at work, yields failures of cooperation, not excesses of pointless attempts to cooperate. And as explained earlier, these must come in contexts where even a strictly self-interested player should find it sensible to

cooperate. It turns out these special requirements are not so hard to meet. Here are three examples.

1. *The Convertible prisoner's dilemma*

Neugebauer (2003) recruited students to play a version of the Prisoner's Dilemma with an option that made it easy to almost guarantee that both players cooperate. So the key factor for the prisoner's dilemma is missing here. The dilemma is ordinarily that players have no way to commit themselves to cooperation, hence easily end up jointly defecting, to their mutual disadvantage. In Neugebauer's game players are offered a very simple way to escape the dilemma. Yet overwhelmingly, they failed to take advantage of that, with very bad consequences for their payoffs.

Neugebauer's PD payoff matrix (fig. 2) was slightly disguised relative to the usual display. And the choices were neutrally labeled "A" and "B", not Defect and Cooperate. Players within a group of 8 (four groups in all) were randomly and anonymously matched for 50 rounds. At the start of each round a player lets his partner for that round know whether he accepts a 20 point penalty if he chooses A. So a player who accepts the penalty but then defects (chooses A) can never end up with a positive payoff, whatever his partner does. For a player who intends to cooperate, accepting the penalty for defecting is just a no-cost way of signaling his partner that indeed he intends to cooperate. He will never incur the penalty. Then, knowing whether the other player has accepted the penalty, each player decides whether to pass this round or play it. And if both decide to play, each decides whether to play A or B.

FIG. 2 HERE

The only strategy that makes sense in this game is (1) Accept the penalty for defecting. (2) Continue only with another player who also has accepted the penalty. (3) Cooperate. Call this Strategy X. This is plainly socially best, maximizing group payoff, but also plainly best simply for self-interest. For wouldn't a person have to be stupid to enter this game against a player who refused to commit himself to playing B? Refusing to commit is likely to mean you either won't get to play this round (which yields 0 payoff) or will get to play it only against someone who has refused to commit, hence presumably intends to play A, giving you a negative payoff. Risk-aversion would make

Strategy X even more attractive. Forward induction, taking account of the fact that you are playing 50 rounds with only 7 possible partners would make it still more attractive. But neither refinement is needed to make blindness to the advantage of accepting a penalty that will never be incurred puzzling for subjects smart enough to be university students.

This is so even though it does not formally dominate a competitive rather than cooperative alternative (call it Strategy Y). Strategy Y would not commit, offer to play only if partner does commit, then defect. When this works it earns 20 rather than the 10 available from mutual cooperation. But it is hardly surprising that it did not work very often.

If players chose Strategy X all the time, everyone would have earned 500 British pence in Neugebauer's game. But most players earned less than 150. Not one of the 32 players was 100% consistent in following Strategy X. Only four players were as much as 90% consistent with Strategy X. But these four earned on average nearly three times as much as the far larger number (13) who, despite plenty of opportunity to learn over 50 rounds, followed Strategy X less than 50% of the time.

Why was such a large majority of these players (English university students) so incompetent?

2. Inverted Public Goods games

A few "Public Goods" experiments have reversed the usual incentive to free-ride by making the return to each player from contribution to the pool larger than the contribution. So these games are degenerate, like the Andreoni and other games described earlier, but in the opposite direction. Instead of removing any rational reason to contribute, in these games it is any incentive to free-ride that is removed. In the inverted games if you give a token it is not just the group return that exceeds 1, but each player, including the donor, gets more than 1 token in return. In the standard Public Goods game other-regarding motivation is required to make giving generously to the pool seem reasonable. In the inverted game pure self-interest is enough. It pays to send all your tokens to the pool.

Figure 3 here

This inverted experiment has been run in both Japan (Saijo and Nakamura, 1995) and Canada (Brunton et al, 2001). In both experiments contributions were not much larger in the inverted game than in a standard game with the same subjects. Figure 3 shows the result of the Japanese trials in which players played 10 rounds with (on the right) a return to each player per token of 7/10 token, then 10 rounds with the super-return of 10/7 (about 1.43). The reverse order is on the left. Closed circles are for the 7/10 return (free-rider incentive), open circles are for the inverted 10/7 (easy-rider) incentive. You can see that the average contribution was about .5 in all rounds for the “easy rider” game (where the more a player gives the more he earns), and also about half in most rounds of the actual free-rider game (where the less a player gives the more he earns). Offered a risk-free opportunity to choose a higher payoff, these players were not moved very far./1

Saijo and Nakamura labeled the behavior "spite". Giving a token nets .43 tokens to the donor, but it gives 1.43 tokens to the other players. So a player focused on how well he is doing relative to other players might hold back on giving. That damages the player's own payoff but damages other players' payoffs even more. So "spite" has become an item in the repertoire of candidates for amending the basic utility function to account for other-regarding behavior, now allowing other-regarding choices to include wanting to make others worse off, or at least worse off relative to the donor.

But when the experiment was later repeated with Canadian students, a preliminary test was run to identify spiteful subjects. The Canadians offered their players a clear opportunity to exhibit spite, as by letting them choose in the preliminary game between the equivalent of a payoff of [10,5] and [10,15], where the first number is what they get and the second number is the payoff that an anonymously matched partner will get. Very few players made the spiteful choice of [10,5]. And the players most prone to spiteful choice in the Public Goods game were not the players who made especially spiteful choices in the preliminary game.

Indeed even within the main Japanese and Canadian trials, choices are incoherent if what motivated holding back on donations was spite. It is not that some players are spiteful, giving 0, while most maximize their payoff, yielding the average results in figure 3. Rather, almost everyone is spiteful. Even without the negative evidence of the

Canadian preliminary test, would it be plausible that almost all players are spiteful? And if inclined to be spiteful, shouldn't they be markedly more inclined to be spiteful when that is reinforced by an increase in their own payoff (when 'spite' is also profitable)? But response to the inverted incentive was so inconsequential that some players actually gave more when giving was costly than when it was profitable. Or shouldn't spiteful players give least in the final round, when there is no chance that others will retaliate? But there is no decline in late rounds. On the other hand, perhaps players are unwilling to be cooperative if others aren't, even if it costs them to do that. But how could that account for withholding contributions in round 1, when there is as yet neither evidence nor any reason to expect that others will be less than fully cooperative?

So as with Neugebauer's game, the question arises: what makes these students choose so perversely? And again the issue does not turn on how far, or if at all, they are inclined to compromise self-interest by other-regarding concerns. Whatever their motivation, the choices appear to be make no sense.

3. *The Minimum game.*

The table shows the payoffs for the "Minimum" game studied by Van Huyck et al (1990) using rather large groups (around 15), but more recently and repeatedly replicated with smaller groups. When there are only two players they are generally able to coordinate on their mutual best outcome (choose 7). But the situation deteriorates as more players are added, and for groups of six or more, the results become terrible.

Figure 4 HERE

Players must each choose a number between "1" and "7". The lowest number picked within a round determines the column in the table that governs payoffs in that round. A player's own choice (possibly itself the minimum) then determines which of the payoffs within that column he will get. Hence, as you can see in figure 4, those who chose the low number get the highest payoff available for that round; and everyone else is penalized by a dime for each notch her choice is above what turns out to be the minimum.

But the lower the minimum number chosen, the lower the payoff to everyone, including the player who chose what turned out to be the minimum.. In the extreme case, if someone chooses "1" in each round, he guarantees himself a payoff of $8 \times 70¢ = \$5.60$

over the 8 rounds. Anyone who fails to match this earns less. But if everyone picks '7' in each round, each player gets almost double (\$10.40). If players could communicate they could quickly agree to pick '7', and this would be self-enforcing since each individual profits from complying. It is best for the group, and in a way that a player cannot improve on by being anything less than fully cooperative. But the players are not allowed to communicate. And they do very badly.

The very robust result of Minimum game experiments is that groups quickly converge to coordination on their worst outcome ('1'). This is a puzzling result from the NSNX + cognition perspective here, but elsewhere has been seen as easily-explained. And indeed, players are correctly foreseeing that others will choose a number less than 7, so that choosing a number less than 7 indeed increases their payoff. A player who reports '7' even in round 1 always loses some payoff. So how could it be a puzzle that players anticipate that and act accordingly?

But I will argue that indeed a cognitive illusion is needed to account for what happens in this game. That will, however, not come until Chapter 10. Here we are only going to be concerned with a degenerate version of the game, where everyone will agree that the players must be vulnerable to some cognitive illusion, since most choices make no sense at all.

The data come from a game framed as a variation on the standard Public Goods game which will be introduced in a more detailed way in Chapter 8. In the variant (Fatas et al 2006), players choose a number of tokens to offer to a common pool, with the balance kept for a player's private account. This looks different from Van Huyck's game in which players choose a number guided by the payoff table in figure 4. But although the games at a glance look different in fact the incentives are essentially identical. In the variant, a pool is generated by doubling the minimum contribution within a group, then multiplying that by the size of the group. The pool that results is then divided equally across the group, adding to whatever each player kept in his private account. This creates the same incentive to avoid giving more than the minimum that drives the Van Huyck results.

But in the degenerate version (which is what concerns us here) only the smallest offer is taken. Any higher offers are reduced to match that smallest offer. It is impossible to waste a token. A marginal token is either needed to increase the minimum, or if not it goes back into the choosing player's private account. So the game is degenerate in the stark sense that the best choice in terms of either self-interest or group-interest is trivially obvious (offer all your tokens). No one could possibly lose by offering to give all his tokens, nor gain any advantage over other players by offering less. What of any possible interest could be learned from such a trivial game? But it turns out to yield a strange and striking result.

In this Spanish experiment players were all university students studying economics. They were given instruction in how the contingent offer arrangement works. They have then passed a test to show they understood how it works. So players who have demonstrated they know enough to make it logically trivial to see that offering all their tokens is the only sensible choice now get to choose. This might seem a pretty easy task even for players who were not university students studying economics.

But in round 1 of the game only 5 of 24 players chose to assure they would take all the payoff they could get. Since 5 of 24 players choosing the maximum level of cooperation would not be at all unusual even in either a Public Goods game or in the standard Minimum game, it is not clear that *any* of the 24 players responded to the game they were playing.

FIG 5 HERE

In one of the six groups of 4, over 20 rounds no one *ever* chose 50 (figure 5), and in another of the groups no one offered a full donation earlier than round 18. In only one of six groups did everyone catch onto how to make the most money until very close to the 20th round. So we are not looking at an occasional lapse of attention or some other sort of odd choice that can be dismissed as just the inevitable quirks in experimental data. This is highly systematic but quite transparently inane behavior, governing the great majority of choices. Outside the laboratory, we see many stupid things, but not *this* stupid.

The cascade conjecture

So we now have a sample of results which meet the criterion set out earlier. In each of these three games, choices that would maximize gain for everyone also maximize strictly self-interested gain for the chooser. There is no need to decide how to model other-regarding choice, and indeed no need to decide whether to allow for any other-regarding motivation at all. It makes no difference at all. On very simple reasoning, in each game there is only one sensible line of play, and in each of these games the players mostly miss it.

Here is a NSNX account of how that might happen, which extends the defaulting discussion of Chapter 6 to the context of social choice.

In Chapter 6, adverse neglect defaulting appeared as an inappropriate response in the unfamiliar and impoverished environment of a puzzle. The default still made sense as what might be entrenched as what to do when cues at hand do not confidently point to what to do. But as a response to a simple puzzle at hand, where there was no logical ambiguity, it did not make sense. Nevertheless, in the unfamiliar and impoverished environment of a puzzle, verbal cues that logically pointed away from that default were not strong enough to overcome it.

So consider a parallel to those effects in the more complex context of social interactions. I want to specify how defaulting could be expected to work in terms of the basic contexts set out in Figure 1 at the start of this chapter, which a reader should review before proceeding. The sketch that follows is only a "just-so" story, tentative in its details. But it gives us enough to consider what we might make of the sampling of strange results we now have at hand.

Figure 6 shows what I will call the cascade.

FIG. 6 HERE

Consider an encounter with another agent, who might or might not turn out to be cooperative. Over Darwinian time, some default response would evolve for ambiguous encounters—what to do when at first it is not clear what to do. And the default response to an unfamiliar encounter could hardly be anything but caution. At the first level of the cascade the default would be the competitive branch. And within that branch, it is the zero-sum games-competitive frame (1a) that would be at least a tentative default, since there certainly appears to be much less chance of bad (for life in the jungle, possibly

fatal) consequences in treating a context as at least possibly zero-sum until you see evidence for a more benign sense of the situation.

But although caution would favor starting from the zero-sum frame, that should be in some tentative way. It would be dangerous to be too easily moved to dovish behavior. But to be uncontrollably hawkish would also not be good. An agent would be alert for cues that would move away from the zero-sum frame (1a) over to the payoff-maximizing frame (1b), where it feels prudent to go about your business, rather than be constantly on alert for a fight or flee choice.

More positive cues that the situation is one that engages social motivation would prompt a shift across to the cooperative branch. The cautious default there would be the risky cooperation frame (2a), where intuition would be guided by "neither selfish nor exploited" propensity to favor what is socially useful, but alert to free-rider concerns. But if a social context appears to involve no free-rider risk, that would prompt a move to the strong cooperation (coordination) frame (2b). If cues are sufficiently favorable (for example, you have had successful dealings with this individual in the past, or there is an opportunity for social gain with no risk of harm or exploitation to self), a NSNX agent could be expected to go immediately to the strong cooperation (coordination) frame (2b).

Facing an opportunity for cooperation subject to free-rider temptation (you are not sure that even if you do your part of the deal, he will do his) a NSNX agent would be in the weak cooperation frame (2a). She must decide whether or not to chance complying (or chance not complying). But as an agent in the zero-sum frame would be alert for cues that warrant shifting across to payoff-maximizing (if it is safe to be there, it better to be there), so an agent in the risky cooperation frame (2a) would be alert for cues that make it prudent to be in the coordination frame (2b). The argument of Chapter 1 which underpins the NSNX balance between social and self-interested motivation would favor alertness to the possibility of a context where cooperation need not be compromised by concern that that others have an interest that conflict with cooperation.

On the competitive branch, NSNX is irrelevant. Motivation beyond self-interest could be in play, but the motivating group-interest would not involve the people you are interacting with. On the cooperative branch, the "neither selfish nor exploited" concerns with free-riders that characterize NSNX become prominent (making the weak

cooperation frame 2a the default) but so does concern with social consequences (making the coordination frame 2b a possibility).

So consider how these cascade effects, possibly interacting with the neglect defaulting effects of Chapter 6, might given an account of the puzzling results of the three examples of challenging results now at hand: the convertible Prisoner's Dilemma, the inverted Public Goods game, and the “money back guarantee” version of the Minimum game.

The convertible Prisoner's Dilemma

Since even pigeons will learn something over 50 rounds, it is scarcely surprising that human subjects did better as play proceeded through the 50 rounds of Neugebauer's game. Nevertheless, as noticed earlier, across all 50 rounds only 4 of 32 players followed the only sensible strategy in this game as much as 90% of the time (accept the penalty, play only if partner has also accepted the penalty, then cooperate.), compared to 13 who followed it less than half the time. The four quick learners averaged triple the payoff of the 13 slow learners.

But we saw examples in Chapter 6 where subjects sometimes neglect what seems too obvious to miss. In the three-cards problem no subject could be so stupid as to find it hard to understand that if the side showing is red, the chance that the card you picked is one with a red side has increased. But in the Fox & Levav “sides” variant, more than a third of their Duke University subjects missed that. Information is staring them in the face, but they neglect it. Since outright neglect can be seen, it cannot be too surprising that incremental cognitive “distance” that is logically trivial can have large effects, as with presentation of a base-rate as “34 out of 100” rather than .34. Relative to these examples, Neugebauer's English university students faced a considerably more difficult task.

People often do not easily recognize the perimeter of their own country if it is shown upside down. Here the payoff matrix (figure 2) is upside-down relative to the way the prisoner's dilemma is always presented in classrooms and textbooks. The “cooperate” and “defect” choices were neutrally labeled “A” and “B”. And unless a person is comfortable with the game theory notation, seeing incentives in the form of a payoff

matrix may be intrinsically cryptic. Combining all these effects, what to the experimenter is transparent might easily be translucent bordering on opaque to a typical subject. On the record, apparently that was so.

Initially, at least, players mostly missed the cooperative structure of the game, which made hard to do anything but badly by playing competitively, but easy to get a big payoff by following cooperative Strategy X. But then, having failed to recognize the context as cooperative, on the cascade conjecture a player would by default start from the competitive branch of the cascade. Within that competitive branch, the default would be the zero-sum strongly competitive frame (1a), looking to "win" the round, where the scare quotes are needed because there is no payoff for winning a round. In a variant of the blundering behavior of economists in the "opportunity cost" example at the end of Chapter 6, subjects would treat the game as too complicated to figure out. With two players, each potentially facing three decisions per round, players missing the cooperative structure built into the game (which makes Strategy X the single best line of play) would see a complicated set of many possibilities. For as many rounds as that persisted, the way to do well would seem to turn on somehow outwitting or at least outlucking another player seen as an opponent, not on finding how to cooperate with a potential partner. Players seem to experiment with whatever happened to come to mind.

That makes no logical sense, given the incentives actually at hand. But it serves well enough as a description of what we see most players doing. Especially in the earlier rounds it was even common for a player to commit himself to cooperation, then play (rather than pass) the round against another player who declined to commit himself. This happened 82 times, and a reader will not be surprised, but apparently players were, that overwhelmingly (70 times) this yielded the severely negative "sucker" payoff.

The most telling situation is that of a player who accepts the penalty for defecting, seeming to assure cooperation, but who then defects against a partner who also has accepted the penalty. Call that a betrayal. To a player in the zero-sum frame (1a) this would not at all feel like a betrayal, but rather like a poker player managing a successful bluff. He "wins" the round, netting nothing from a payoff of 20 dissipated by the 20 point penalty, but also losing nothing while the other player loses 10. But on the cascade conjecture it is not hard to slip across frames within a branch. Given a strong and direct

cue (betrayal reduces payoff from 10 to 0), at the moment of choice even players who indeed started from the zero-sum frame would seem unlikely to remain caught on that default. Then a player who was not seeing his choice at all as a choice to cooperate, but merely doing what makes sense for himself, would make the payoff-maximizing choice (from frame 1b). That happens to be the cooperative choice but with the penalty in force it need not be motivated by any inclination beyond strict self-interest.

So although even players whose opening choice is to accept the penalty might be starting from the zero-sum frame, at the moment of choice they are likely to be nudged over to payoff maximizing (frame 1b). Their choices, in this round, are the same that a player starting from the cooperative branch would make. And a player who declines the penalty and then either does not have a cooperate/defect choice (the round is not played out) or has a choice and defects, also might or might not be starting from the zero-sum frame. There is no way to tell. If he declines the penalty but then cooperates anyway, then he cannot have started from the zero-sum frame. This occasionally happened, but far too rarely to conflict with the possibility that frequently players did start from the zero-sum frame. So can we see evidence in the data that tests the cascade conjecture that starting from the zero-sum frame, though illogical, should nevertheless be common? That turns out to be possible because, in addition to the game with a 20-point penalty, Neugebauer also ran an experiment with a 10 point penalty, which allows us to draw some comparative inferences.

For a player who accepts the penalty facing a partner who also accepted the penalty, but with the penalty reduced to 10, either choice ("A" or "B") gives an identical short-run payoff (+10). But for a zero-sum player choosing "A" (defect) yields both the 10 tokens and a win [me +10, him -10], while "B" yields only a tie [me +10, him +10]. So in the 10-point game, we might be able to actually see players who apparently did start from the zero-sum frame, because they record final choices that make no sense other than if they are still in that zero-sum frame. There might not be many such choices. Even in the 10-point game, forward induction would alert a player that there is likely to be a payoff cost. But (on the cascade conjectures) we have reason to look for betrayals in the 10-point game.

And a second inference might discriminate intrinsic zero-sum behavior (players explicitly prefer reducing payoff to the other player even when yields no gain for themselves) from tactical zero-sum players (who are responding intuitively to a misperception of what kind of game they are in). On the NSNX + cognition argument we should expect to see the latter not the former. But if intrinsic motivation is zero-sum, then since betrayal must reduce the prospects of cooperation from the betrayed player in a future round, betrayal ought to be more restrained in round 1 (where the forward induction motive against betrayal is strongest) relative to round 50, where that motive has disappeared. On the other hand, if zero-sum responses are prompted by adverse defaulting (on reflection a player would regret what she had done), the opposite should hold. As players come to know the game better, vulnerability to that adverse default should decline. It becomes more likely in later rounds, and in particular by round 50, that a player has come to understand the game she is in, and she would no longer be starting from the zero-sum frame. So for intrinsic zero-sum choice, betrayal is more likely in round 50 than in round 1, and the reverse for defaulting zero-sum choice.

Fifty rounds of play among 32 players generate $32 \times 50 = 1600$ individual choices (800 pairwise interactions). For a player who saw the actual structure of the game, deciding whether to accept the penalty is easy. There is no need to guess, since accepting is best every time (Strategy X). But declining the penalty was sufficiently common in the 20-point game that at least one of the partners refused the penalty more than half the time.

In the 20-point game, among the 800 interactions there were only 345 that were played out between players who had both accepted the penalty. A very large majority of those 345 both-commit rounds produced the cooperative result. That is not surprising given the strong direct payoff cue, the subtler forward induction cue early, and the accumulation of experience later. Among the 690 individual opportunities for betrayal in the 20-point game, there were only 21 actual betrayals, of which 11 came from just 2 of the 32 players.

But in the 10-point game, though the number of accept-the-penalty interactions is almost the same (345 in the 20-point game, 354 in the 10-point game), the payoff nudge to break free of a zero-sum default is weak. Although there remained the forward

induction concern to cue a shift to the payoff maximizing frame (1b), betrayals were now almost as common as mutual cooperation. Players mostly did make the cooperative choice (563/708) when they had that choice to make. But since it takes two cooperative choices to avoid a betrayal, that left 145 betrayals among 354 pair-wise opportunities (vs. 21/345 with the 20-point penalty). Total earnings fell to half what they were in the 20-point version. So the first inference clearly holds. Betrayals indeed are conspicuously more common in the 10-point game.

And as the second conjecture anticipates, players accept the penalty more often as they gain experience, and betray less. In the strongest comparison, in the first round 15 (of 32) players accepted the penalty, which produced four accept-the-penalty matches, three of which yielded betrayals. Among the 8 opportunities to betray, there were 4 betrayals (one a dual betrayal involving both players in a match). In the final round, 25 of the 32 players accepted the penalty, yielding 18 opportunities for betrayal, of which 3/18 occurred, in contrast to 4/8 in round 1. All this is as would be expected on the cascade conjecture that betrayals are likely to be seen in the 10-point game, but due to adverse defaulting rather than intrinsic zero-sum preferences.

This evidence for adverse defaulting to the zero-sum frame is only suggestive. Whatever might prompt betrayals -- perhaps an intrinsic taste for spiteful behavior not adverse defaulting -- would prompt more of it when betrayal is cheaper. The contrast between betrayals in round 1 and in round 50 favors the adverse defaulting explanation, but is also not clear-cut. That might have been mostly because players who naively declined the penalty would have cooperated in round 1 if they had accepted the penalty. But by round 50 many more are following Strategy X, getting to play, hence show up among the 25 acceptors in round 50 instead of being hidden among the 17 not-acceptors in round 1. The detailed data do not really support that. But it is enough for the NSNX + cognition view to claim only that the test is consistent with the conjecture. The accumulation of other evidence to come strongly reinforces the cascade interpretation.

Some of that is in the balance of this chapter. More will come later, in particular in Chapter 9, where we will see data that incidentally, but emphatically tests for a propensity to actual zero-sum preferences. If players betray because they are often actually spiteful then we should see plenty of spiteful behavior when it costs nothing at

all to be spiteful and when in addition the players face provocation that would give reason to exercise spite. But in the data in Chapter 9, we will see very marked reluctance to impose costs on another player even when provoked by nasty behavior from the other player, and even though no cost to self would be involved. The puzzle in Chapter 9 will be to account for the relative absence of nasty responses even in a context where nasty responses would seem normatively appropriate.

The Inverted Public Goods game

How to interpret what looks like spiteful choice is even more prominent in the inverted Prisoner's Dilemma game, where that is the only straightforward explanation and has been the standard explanation. On a "spite" reading, the Japanese and Canadian players in these games are not motivated exclusively by spite. They do not single-mindedly value doing better than fellow students over cooperation that increases payoffs for everyone. Rather, spite is tempered by greed, so on average they contribute about half their tokens, and withhold the other half. But the data (figure 3) show very nearly as much cooperation when spite must compete with greed as when spite would be reinforced by greed. The open-circles data in figure 3 (for the inverted trials) show no hint of either the holdup of contributions in round 10 or the high contributions in round 1 that would make sense if motivation was spite qualified by greed. Several players actually contributed more when contributing reduced their payoff (the 7/10 return) than when it increased their payoffs (the 10/7 return).

All that suggests that spite and greed are not really insightful categories for understanding what generates these results. Rather, the data show a pattern of choices consistent with NSNX motivation *distorted* by an adverse default in the cascade. In this case (in contrast to that of the convertible prisoner's dilemma), an adverse default on the cooperative rather than competitive branch. Players respond as if they recognize the context as cooperative, which indeed fits the logic of the game they are in. But on the logic of the game they should escape the *risky* cooperation default (2a). There is zero free-rider risk. The more you contribute the more you profit. Responses, however, fit the default anyway. They reveal "neither selfish nor exploited" concern about being

exploited if cooperating more than others, and they do so starting in round 1, before there is any reason to suppose that would happen.

But if adverse defaulting accounts for what is happening, we want to be able to say why the cooperative character of the game is transparent to the players in this inverted Public Goods context while the also cooperative character of the convertible Prisoner's Dilemma game is missed by the players. And we want to say why, having seen the context as cooperative, players then fail to see that cooperation is easy (they can just think about coordination), since there is no free-rider risk in this game. But with the notions we now have in hand, neither is hard to do.

In the convertible Prisoner's Dilemma, a player has to tease out the implications of the payoff matrix, adjusted for the implications of the opportunity to accept or reject the penalty provision, and then for the opportunity to accept or reject actual play of this round. None of this seems at all difficult, but in the aggregate it also not trivial on the scale of the difficulties that yield terrible performance in the puzzles of Chapter 6. In contrast, the Public Goods game does not have that logically modest but perhaps cognitively substantial additional layering of difficulty. Relative to the difference (again) between a base rate presented as 34 out of 100 vs. the same base rate presented as .34, the incremental difference in complexity between the inverted Public Goods game and Neugebauer's presentation of the convertible Prisoner's Dilemma is not inconsequential.

But players come to the game with normal human experience of cooperation. The game looks like a situation every player has encountered often, where contributions to a group effort plainly benefit everyone. But usually some members of the group might be tempted to shirk. And what players miss in the inverted game is the unusual feature which makes the immediate return from a player's own contribution so large that shirking would be stupid. A person does best, even in terms of narrow self-interest, to contribute everything he can. So although social experience would prompt a person to expect a free-rider issue to be present in this interaction with anonymous others, here it isn't. Intuitions need to be adjusted to catch that. But over and over in Chapter 6 we saw that choosers are frequently caught by a neglect default that leaves them cognitively blind to such details.

The patterns of contribution in figure 3 for the inverted game look almost indistinguishable from the patterns for a game which in fact does entail a free-rider risk. Somehow players act as if they did not notice that for the inverted payoff rounds, returning 10/7 token per token contributed instead of 7/10. The quantitative information, which logically shifts the game into the coordination frame (2b), where there is no free-rider concern, is not cognitively effective enough to overcome the default tendency to respond in a common pool situation as if there were a free-rider risk. But we have more than once had occasion to notice parallel insensitivity to large quantitative shifts in other contexts.

What appears to happen here is that players generally escape the default first branching of the cascade, but then are caught on the weak cooperation frame on the cooperative branch. They respond to the familiar context of a group endeavor by favoring the cooperative over the default competitive branch. But they then sufficiently neglect an atypical quantitative detail to be caught by the risky cooperation default on that cooperative branch. There "neither selfish nor exploited" intuitions prompt players to be wary of cooperating more than others, lest they be exploited, though there is actually no way they can be exploited. Players act as if there were a free-rider risk, even though there is no free-rider risk. Illogical though that may be, it is apparent in figure 3 that that is how players are behaving. The cascade conjecture tells us how it could be that choosers do that.

It is a conjecture that what we are seeing are NSNX choices revealing the force of adverse defaults. But it is not a conjecture that players act *as if* that were so. Whatever the cause, the data, not the cascade conjecture, shows the restrained cooperation characteristic of NSNX responses to free-rider incentives in a context where that make no sense. But that is consistent with the adverse defaulting which seems to govern responses to Neugebauer's convertible Prisoner's Dilemma though here the adverse default is on the cooperative branch, not a defaulting failure to even reach the cooperative branch. And if there remains, perhaps, some plausibility to the contrary conjecture that somehow the choices we see in this game, and also in the convertible Prisoner's Dilemma game, can be explained with the help of a *spite* motive, that pretty dim prospect becomes vanishingly remote in the light of the concluding example coming next. Spite in the two games

considered so far would be quite stupidly pursued, and hardly typical of how students usually respond to anonymous interactions with other students, but if you are desperate enough for an explanation, that might seem tempting. But we can turn now to an example where "spite" is not available even as a perverse and barely plausible possibility.

The degenerate Minimum game

Strategy uncertainty has been the standard explanation of how rational choice in the Minimum game (figure 4) can lead players to their worst possible outcome. Within a very few rounds coordination is focused on the lowest possible payoffs, which are about half of what is available. But each player is tempted to choose low by what indeed turns out to be correct anticipation that someone else will choose low. I will return to the game in Chapter 10, to show that this standard explanation might in fact reflect intuitions among analysts of the game captured by the same illusory mechanism that captures players within the game.

But here we deal with a entirely uncontroversial case of illusion. In the degenerate game we have the "money back guarantee" described earlier in this chapter. A player can only gain, never lose, by offering to contribute the maximum. There is no spite incentive even for a player who would be spiteful if she could, and also no chance of being exploited. Every player in every round gets exactly the same payoff. But we have just considered how neglect of qualifying information could miss that. Then a can't lose context is left looking like a free-rider context. And the same possibility here would have the same consequence in a slightly different form.

The data starkly show players treating an opportunity for profitable, risk-free cooperation as subject to a free-rider concern that does not exist. Since this is overwhelmingly evident in round 1, it cannot be somehow explained as a reaction to what other players have done. Nor can it be revealing some perverse propensity of players to spitefully value doing better than others even when the alternative is doing just as well as others. That cannot happen no matter what the chooser does or what anyone else in the game does. Everyone gets the same payoff, which can be larger or smaller contingent on player choices, but is always exactly the same for every player. In terms of the cascade,

players act as if they caught in the risky cooperation frame (2a) when in a completely unambiguous way they are a pure coordination context (2b).

A modest degree of translucency turns out to be enough to yield the logic-defying choices in this game. After 10 rounds of the standard Minimum game, Van Huyck offered his subjects a no-risk game in an additional five rounds. That was essentially the same game as the Valencia no-risk game, but more starkly transparent. Van Huyck's players had just played 10 rounds of a real Minimum game. Now they play 5 rounds of the degenerate game. The explicit payoff table of Figure 4 now is changed into a schedule that no longer imposes a penalty for choosing more than the round minimum. And with this degree of experience and transparency, even though there are many more players in a group than in Valencia, only 14 of 91 Van Huyck's players started with the blunder made by 19 of 24 FNJ players. But although one game makes the situation harder to miss than the other, this is a distinction between speaking clearly and shouting. Why do players have to be shouted at to notice something clearly in their interest to notice?

What should we make of what looks like severe aggravation of vulnerability to adverse defaulting, which we have been seeing repeatedly, due to a modest change in the transparency of the situation? Think of a tennis player instructed to keep her eye on the ball, and having no doubt about how to respond to a test question on this point. Nevertheless, as every tennis player knows, she will sometimes take her eye on the ball. Why is it so hard to avoid that? Presumably because outside a few specialized contexts (like playing tennis) we do best keeping our eye on what is ahead of us. We watch the road, not our hands on the wheel. So it is understandable that keeping your eye on the ball, not on where your opponent might be moving, might go against some well-entrenched default. Then, if concentration flags momentarily, the usually advantageous default can slip into place, even though we know that while playing tennis that is not helpful at all.

And here, on the cascade conjecture, the default in a potentially cooperative situation is to guard against exploitation by free-riders. Players in the no-risk game have received instruction and testing about how the game works, but they have not had anything like the repeated experience that a tennis player gets but still is insufficient to

avoid lapses even among professionals. A bit of translucency turns out to be sufficient to make players vulnerable to a default that prompts responses that make no more sense in this context than the usual illusory responses to the puzzles in Chapter 6.

If the situation is stark enough, we get a jolt sufficient to displace the default propensity to neglect a piece of secondary information. But if the clue is only translucent, not completely transparent, the jolt is cushioned, and the default might stay in place. Here we can observe that for most players, it does stay in place. But, reprising a point already made, if effects like these can prompt players to damage their own interests in an only mildly translucent experiment, then in far less transparent natural situations the same effects which here prompt players to visibly nonsensical choices might also generate enough unfortunate choices to make effective social cooperation difficult. It could hardly take more than a moment's discussion to correct the misbehavior in the degenerate Minimum game. But in a difficult natural setting, where nothing will be so starkly apparent, correcting a misperception of what makes sense may not be easy at all. Situations marked by perversely inappropriate responses should be uncommon. But especially under novel conditions, they ought to be sometimes encountered, so that something consequential might be learned by careful attention to anomalous results in the vastly simple, artificial, but far more amenable to analysis realm of social choice experiments.