

HARRIS SCHOOL WORKING PAPER
SERIES 06.05A

CHAPTER 9: RECIPROCITY PUZZLES*

Howard Margolis

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

Chapter 9. Reciprocity puzzles

Normal situations sometimes yield abnormal behavior. We know there are serial killers, pathological liars, paranoid psychotics. But they are uncommon enough that unless your line of work is psychiatry or criminal investigations, they would not be prominent in your understanding how people are behaving. But in data from experiments it is not so unusual to see a major fraction of choices violate what both common experience and essentially universal social norms would lead us to expect. The problem is not that players may behave differently in the lab. Something of that has to be expected, since no lab experiment can capture exactly the conditions of choice outside the laboratory. But the sort of problem I want to explore here arises when choices in an experiment look so completely different from what we might expect that there is a challenge in conjecturing what kind of context might prompt such abnormal responses from apparently normal people.

We know that can happen, since we have already been through a set of examples (in Chapter 7). But there the odd results had a comical flavor. What we have every reason to expect were reasonably intelligent, and certainly sometimes highly intelligent, subjects were making choices that are hard to describe as anything but very stupid. The material in this chapter has a different character, because essentially it is about *character*. We will be looking a set of games where subjects seem to reveal weakness of character that is as hard to believe as the weakness of intellect that seems to be revealed by the examples in Chapter 7. We will see players (students in Barcelona and Berkeley) who seem immune to normal human responses to generous treatment, and also normal human responses to bad treatment. Since the same players reveal both, we are not speaking of *bad* character but of *weak* character. I will be giving an account, as you will now expect, in terms of adverse defaulting interacting with NSNX motivation, to yield choices which are sometimes selfish, sometimes unselfish, but in both directions often inappropriate.

The Trust game is often cited as an experimental demonstration of reciprocity. A and B each are given 10 tokens. A sends as much as he wishes to B, which is then tripled. B sends whatever he wishes back to A. Trustees (B) by a large majority return tokens back to the trustor (A).¹ But define *full* reciprocity as when the trustee returns the tokens the trustor put under his control, and then fully shares profit from trustor putting

his endowment at risk. If A sends 10 tokens, full reciprocity occurs if B returns the 10 tokens and also shares the 20 token profit equally. That rarely happens. But when it does, if each starts with 10, each ends with 20, which if communication was allowed is what a competent trustor would demand and what a reasonable trustee would agree to. Trustor is doing what he knows is efficient. Trustee is reciprocating by doing what he knows is fair.

Against this standard, reciprocity can then run from 0 (no return to the trustee) to 1 (full profit-sharing). Allowing that reciprocity is a matter of degree, not dichotomous, there is still evidence of reciprocity in Trust game data, but it is usually evidence that reciprocity is feeble. In a large fraction of Trust game interactions, trustors do not even get back as much they send. The trustee pockets the profit and keeps some of the trustee's tokens as well. On average trustors are doing well in these games if they end up with some modest gain from putting endowment at risk. If this was the usual result of trusting in real life, we would see very little of the cautious, qualified, sometimes disappointed, but still substantial not feeble trust that plays an essential role in successful economies. But much more extreme failures of reciprocity are can be found in experimental data, and the focus of this chapter is on a salient example of that.

The anomalies which were the focus of Chapter 7 showed that experimental results could violate any coherent account of what to expect from subjects with normal intelligence. It should not then be surprising that subjects who fail to follow their own clear interests even when simple self-interest is all that is appropriate might also fail to follow norms to the extent that people like them usually do follow norms outside the lab. Conforming to norms is certainly imperfect, as stressed in Chapter 3, but a very long way from inconsequential. List, writing with Harrison (2004) and with Levitt (2005), has built a strong argument that subjects in the laboratory often behave better than people unobtrusively observed in natural settings. That is certainly true. But a broader look reveals that the opposite is also true. Players sometimes are less social in the lab. And sometimes, we have seen in Chapter 7, players are just stupider in the lab. In this chapter we will see the same holds for the moral dimensions.

In everyday life we take it for granted that if A is nice to B, then B will feel some obligation to be nice to A, and even if A is a stranger he will never see again. B might

not in fact do the nice thing, but if not we look for an explanation. If B can do the nice thing at very little cost to himself, but still doesn't, that is stranger. If B treats people who actually have been nice to him worse than he treats people who have done nothing for him, that is strange. Or if A is *nasty* to B and B in return is nice to A, that is strange. If C sees A is being wantonly nasty to B that would normally affect C's inclination to be nice to A. If she doesn't, that is strange. And so on.

It is easy to find examples in experiments of all the choices I have been labeling strange, other than the last. I will return to that significant exception. But given what we have seen already of self-damaging choices apparently prompted by adverse defaulting, we might consider how strange behavior with respect to reciprocity also might be governed by adverse defaulting.

Rabin (1993) proposed a model of other-regarding choice that turned critically on reciprocity. Charness & Rabin (2002) then reported on a set of 32 simple games which appear to have been designed to explore reciprocity effects. But reciprocity turned out to be almost invisible. CR note that, but neither they nor anyone else commenting on this very widely-cited paper has offered any explanation, or indeed explicitly viewed the results as puzzling, though most experiments indeed show at least the qualified reciprocity of typical Trust games. But if you think about what the data seem to show, you will agree that indeed these results demand some explanation.

Figure 1 shows the 32 CR games, the number of players in each, and the results. The notation $A(x,y \text{ or } B(w,z \text{ or } m,n))$ means Player A can exit the game by choosing *left*, which takes x for himself with y to Player B. Or A can chooses *right*, which passes the choice to B, who must choose between w for A, z for B or m for A, n for B. By changing the structure and parameters (the values of x,y,w,z,m , and n) many different situations can be set up. Sometimes (a change in structure) in games of this sort A makes a choice between a pair of possibilities (he has no “exit”) option, and then B makes a choice in response, as in the Prisoner's Dilemma game I described a few paragraphs back. CR did not use this structure, but they did use variants in which there is no A choice, just a B choice (B is “dictator”), or in which a third party (C) chooses. Throughout, players are in sessions of four games, where in most a player makes both an A and a B choice (or

in several an A and a C choice/3), anonymously matched at each choice with a different player among the several dozen in the experiment.

Start with CR23, where the choice was B(800,200 or 0,0). B is dictator. If he chooses left, an anonymous A would get the 800, and he would get 200. Or B could choose right, yielding 0 for both. Here B's unanimously (36 of 36) chose 800,200. The CR players, and players in choice experiments very generally, prefer equal payoffs if there is no reason or temptation to depart from that. Players in all these games show some preference for equality. But on any plausible model for other-regarding motivation, the preference for equality should not extend to no payoffs (which of course are perfectly equal) when there is no reasonable question of fairness. But here the player getting the larger payoff (A) did not in any way choose that. So this is an easy case to understand. But in a slightly more complicated game we begin to see odder results. Falk, Fehr & Fischbacher et al (2000) tested A(8,2 or 8,2);B(8,2 or 0,0). So they offered their A's the degenerate choice 8,2 or 8,2 immediately before B's made the proportionally identical choice just considered in the CR game. Translating tokens into dollar payoffs, the CR and Falk et al games involved much the same stakes. But now 20% of B's chose 0,0, apparently to punish A's for the one-sided allocation though A's had no more actual input than in the CR game. It is as if B noticed the only the situation A's choice had put him in, but not the alternative if he had not done that. B acts as if he did not notice that A in fact had no choice. A reader who recalls what happens in the puzzles of Chapter 6 and the degenerate games of Chapter might recognize this behavior. Somehow, players know what situation they are in but seem to be neglecting information highly relevant to responding to the situation. This possibility will play a large role here, since you will see that choices in the CR data over and over seem to be tied to this, on its face, exceedingly unlikely possibility.

FIG 9.1 (full page) HERE

But this case is only mildly puzzling compared to many others. In CR28, A can exit the game by accepting 100 while B gets 1000. Or he can force B to choose between 75 or 125 for A, with B getting 125 (instead of 1000) whichever he chooses. Since each point was (contingently/2) worth a penny, an A player who forced B to choose has deprived B of \$8.75, apparently in the hope that B will respond by rewarding him with an

extra 25¢ rather than punish him by 25¢. And half of the A players do throw away \$8.75 of B's \$10 payoff on a gamble that this will make themselves better off by 25¢. In the B role, players then respond to the possibility that A has done this. The apparent judgment of players who make the risky and greed choice in the A role then proves sound. Two thirds of the B choices do reward the A choice which throws away most of their payoff. We appear to have A players who are quite viciously selfish and apparently expect B players will mostly turn out to masochists who will respond favorably to this treatment. And they are apparently are correct to suppose that.

In the CR games, within each game (and across the four games in a session), each player makes an A choice, and (anonymously matched against a different player) also makes a B choice. A players and B players are the same people. Taken together the choices reveal that players are quite viciously selfish but also extraordinarily indifferent to losing most of their earnings. And looking across the subject pool, these player are also extremely altruistic as well as intensely selfish. In CR15, facing B(200,700 *or* 600,600), 73% surrendered 100 points of own payoff to provide a gain of 400 points for their partner. And in the A role in CR14, rather than keep all the tokens for themselves when facing A(800,0 *or* B(0,800 *or* 400,400)), a third were willing to share equally even though this not only involved a large sacrifice to benefit an anonymous other player, but also the risk that a B player might take advantage of their generosity and just keep all 800 for himself.

Which leads to a startling next result. Almost half (45%) of the beneficiaries of A's generosity did keep all 800 for themselves. In this A(800,0 *or* B(0,800 *or* 400,400)) game, even a pathological selfish person might see that reciprocity is called for. Being pathologically selfish, he will not reciprocate anyway. But it seems hard to understand how anyone could fail to notice what is called for. A could just keep 800 (worth \$8). But he doesn't. He allows B to split the 800 equally, even though this allows the chance that B will just pocket all for himself. And almost half then do pocket the entire payoff for themselves! So in addition to being viscously selfish and masochistic and altruistic, CR players are also about as likely as not to be utterly immune to the normal propensity to reciprocate kind treatment, even in a really clear case that calls for it.

And there is more. In CR14, the fraction who shared when the A choice risked 800 in hand to allow an equal split turned out to be almost exactly identical to the fraction who shared in CR18, where choosing *right* involves no risk at all. Here the choice is A(0,800 or B(0,800 or 400,400)). A can choose to get 0 or allow B the opportunity to share. Unsurprisingly, all subjects were willing to give B an option to share. The fraction of B's who then did share when reciprocity was irrelevant (56%) turned out to be higher (insignificantly, but higher not lower) than when reciprocity was plainly in order (55%). That in one case A had run a major risk by allowing B to choose to keep all or to share, and in the other took no risk whatever seems of no interest to these players. Any propensity to reciprocity is again utterly invisible.

And we can conclude with an even more extreme example. In CR2 and its replicate (CR17), B's choice is 400,400 or 750,375. B can improve the outcome for A by 350 if he is willing to give up 25. So contrast that with CR3 or CR4 or CR21, in all of which the B choice is the same as in CR2 or CR17, but now after A has risked 400, and in one case (CR4) also sacrificed at least 50, to gain at least 375 for B. We have a baseline for judging the effect of reciprocity by noticing that in the first case (no reciprocity at issue) half of B's were sufficiently generous to A to accept a loss of 25 (from 400) to gain 350 for B. But after A's have run the risk that they could be abused by a really selfish B's response, the result is different. Now less than 40% (not fully 50%) are generous. A clear majority (above 60%) of B's return A's favor with abuse rather than pay a very small price to reciprocate this generosity.

A Nobel laureate (Amartya Sen, 1977) wrote a well-known article under the title "rational fools", where the rational fool was the economic man of the standard model. But at least Sen's targets were rational. Sen, however, relied on everyday intuitions about everyday situations. As he intended, readers easily saw his postulated behavior not so much as exemplifying pure self-interest as exemplifying social idiocy. As his title anticipates, a reader sees Sen's examples of pure economic man as fools. However, what all the games reviewed here have in common is an absence of any familiar connection with the choices to be made. As is usual, in none of the games here are players allowed to talk to others in the game, or know who their partner might be. The only thing players can easily recognize is that they are players in a game, and we all know that in a game

you should try to win. Most often that means doing better than the other player. But everyone also has experience with games where you win by cooperating with a partner, as in bridge or charades.

FIG 9.2 HERE

Players in these games ordinarily sit at computer consoles, matched anonymously with some unknown individual at another console. Making a choice consists of looking at a diagram like one of the game trees in figure 2 and clicking a button on a screen. Although it is a strong tradition within the experimental economics community never to lie to subjects, players in fact have no way to be sure there is actually any other player to be affected by their choice. But since actually telling subjects (in the degenerate game described at the start of Chapter 7) that there are no real partners does not have a huge effect, that source of uncertainty is not likely to be a problem. But the severe shortage of cues that link to the world of visceral experience, could nevertheless yield odd effects. Even the simple tree diagrams in figure 2, though transparent to anyone who has worked with this sort of notation might be considerably short of transparent for many subjects, parallel to the difficulty subjects might have with the Prisoner's Dilemma matrix in Chapter 7.

Our brains are set to be alert for orienting cues. Absent strong cues even weak cues can have strong effects. And absent even sufficient weak cues we fall back on defaults that guide intuition when it is unclear what to do. From the defaulting argument and examples of Chapters 6 and 7, in the impoverished environment of the CR games, we might even expect defaulting to play a significant role, and we find it.

Individually, but especially in the aggregate, the CR games reveal multiple challenges to essentially universal moral intuitions, or to common sense, or both. The CR games are not unique in this. But it is especially easy to see in this data where we have responses from the same player in both the A and B roles (against different anonymous partners) usually making in all eight choices in a session of 4 games.⁴ These players turn out to be neither reliably self-interested, nor reliably other-regarding. Sometimes they behave the way their mothers taught them, but often they don't. Almost no one is a consistent conditional cooperator, though in other experiments that is often reported as the most common "type".⁴ Players seem to jump from type to type across

choices, like someone who sees the well-known gestalt drawing as a duck one moment but as a rabbit the next, and while seeing things in one gestalt comprehending things the other way is completely out of sight. I will be showing that gestalt shifts of very much this sort seem to resolve a series of puzzles in the CR data, where the gestalts are frames within the *cascade* developed in Chapter 7.

The *neglect* default developed in Chapter 6 comes into the argument because the data show us B choices that seem to make no sense unless players are in fact *neglecting* the payoffs they would receive if A does not choose to pass the decision on payoffs to B. Players in a CR game see a decision-tree like that on the left in Figure 2 (which is the game tree for CR1). But if the neglect default were left in place, a player might respond as if the game he was playing was the truncated game on the right. Logically this makes no more sense than the illusory responses that subjects overwhelmingly give to the puzzles reviewed in Chapter 6. In particular it makes no more sense than the very common response of sophisticated subjects to the puzzle that concludes Chapter 6, where they appear to neglect conspicuous and obviously relevant details of the simple diagram in figure 6.1, as here players respond as if they saw only their bare choice that might be coming from A, and neglecting the alternative A might have chosen. I will call that a response to the truncated situation on the right in Fig. 2, in contrast to a response to the situation in the actual game they face, on the left in Fig. 2. Sometimes B should be very pleased with the choice he has been offered, sometimes B should be appalled. But the B responses mostly seem to have gotten things backwards. It is instructive to consider in detail what would happen if indeed adverse defaulting (defaulting that on reflection the chooser herself would regard as inappropriate) was shaping responses.

Recalling the cascade diagram introduced in Chapter 7 (p. --), the competitive branch is the default relative to the cooperative (if in doubt, better treat the situation as competitive); and within the competitive branch, the zero-sum frame is the default. But a rational agent would be alert for a cue pointing away from the zero-sum (1a) to the payoff maximizing (1b) frame. It would be dangerous not to guard against the possibility that an unclear situation will turn out to be zero-sum. But it would be self-damaging to unnecessarily fail to maximize payoff. And a NSNX agent would also be alert for a cue that would prompt a shift across the cascade to the cooperative branch, where cooperation

subject to free-rider concerns (2a) would become the default (what to do if you're not sure what to do).

FIGURE 3 ("cascade" reprinted from N7.4) here, if needed -- or a page reference.

Cooperation is weak in frame 2a in the sense that a person sees cooperation as a possibility, but might not cooperate anyway. Choice is under the "neither selfish nor exploited" tension fundamental to NSNX motivation. But in a context where free-riding does not seem an issue, the strong cooperation frame (2b) comes into play. An agent who sees the context as one where he is not tempted to free-ride or worried that others will be tempted sees the choices as a pure coordination problem (like driving on the right in a country where everyone else is driving on the right). But this gestalt can also govern cases less simple. You tell guests at a party to toss their coats on your bed, without worrying that you had better follow them into the bedroom to make sure they don't steal anything.

Even in the strong cooperation frame, and of course also in the weak cooperation frame, an agent does not always make generous choices. If the choice is with respect to someone who has behaved badly, punishment not reward is what is likely to seem socially appropriate. But in the coordination frame, ordinarily choice would yield benefits for others, and in the weak cooperation frame choice sometimes benefits others even at a cost to self.

Here is the situation described in terms of the NSNX formalism introduced in Chapter 1: On the competitive branch an agent seeks to maximize $S(\text{elf-interest})$, but in frame 1a the argument of the self-interested utility function is the difference between own payoff and other's payoff, while in frame 1b it is own payoff. In frame 2b on the cooperative branch, an agent seeks to maximize $G(\text{roup-interest})$, governed by his own sense of what would be socially good, as discussed in some detail in Chapter 1. In frame 2a, however, agent does not directly seek to maximize anything, but rather tries to move as close as is feasible to a "neither selfish nor exploited" equilibrium, where $W = G'/S'$, again as developed in some detail in Chapter 1.

And applying this to the CR games:

(1) *Without* a potentially threatening agent in sight (meaning an agent with an opportunity to respond adversely), and where a player's choice would not affect own

payoff (so there is no risk of being exploited), the default context for a NSNX agent would be coordination (2b). Only social concerns are immediately salient. This is the situation for the several CR games (CR10, 12, 16, 20, 24) in which a referee (C) chooses the response, but also in four of the games in which B responds (CR5, 7, 28 and 32). A NSNX agent would favor a generous choice unless something recognized in the situation makes punishment seem more appropriate.

In the converse case where B's choice does change own payoff (with no active agent in sight) the NSNX default would be the "neither selfish nor exploited" 2a. B then might or might not in fact sacrifice something from own payoff to benefit his partner, contingent, as in the pure coordination case, on whether the partner seems to deserve that generosity, but also on how much a generous choice will cost relative to the benefit it could provide. This situation holds for eight games in which B's choice is 400,400 *or* 750,375, for three games where B's choice is 0,800 *or* 400,400, and for a variety of choices in ten other games (CR6, 9, 11, 19, 22, 25, 27, 28, 30 and 31).

(2) But suppose there *is* a potentially threatening agent in sight. The cautious zero-sum frame (1a) is then the default. This holds for all A choices in the CR games. But that the zero-sum frame is the default for this kind of situation does not imply that a player is likely to be caught by that default. Risk that own-payoff will be harmed in response to a zero-sum move would push a player towards the payoff-maximizing frame (1b). A shift from the zero-sum to payoff-maximizing would be particularly easy within the CR games, since in these games zero-sum play in fact makes no sense (there is no payoff for doing better than the other player), so that once pushed to look closer, a player would be unlikely to remain caught by that default.

And for NSNX agents, a cue suggesting a cooperative context could shift the frame all the way over to the cooperative branch of the cascade. In the CR games, B has no opportunity to provide an explicit cue. But there is a cue built into the structure of the games. Except for CR18, each of the CR response games essentially asks A either "Are you willing to run a risk to improve your competitive outcome?" as in CR1 and ten others, or "Are you willing to run a risk to improve the cooperative outcome?" as in CR3 and eight others. So the framing might suggest the perspective appropriate for assessing the risky choice: competitive in the first case, cooperative in the second. The nudge

towards the competitive branch in the first case would reinforce starting from the default competitive branch. But the nudge towards the cooperative branch in the second case might be enough to move a NSNX agent over to cooperative branch of the cascade. The question, I think, is not whether there is such a nudge in the framing of the games. Of course there is, and it could hardly be entirely avoided. But is it consequential? Theory cannot say, but as will be seen, the data give us a clear answer.

(3) But the sampling of CR results earlier in this chapter suggests we also need to consider the possibility that in these games players somehow are caught by the *neglect* default developed in Chapter 6. In a real situation anywhere near as transparent as these games, that should essentially never happen. But we have already seen many examples (in Chapters 6 and 7) where in the impoverished environment of an unfamiliar game things of that sort do appear to happen. And here, when we come to the data, what certainly looks like a blindness of B choosers (in the CR response games) to what is right in front of them will be apparent. Instead of responding to the situation, B seems to respond to what I have already called the *truncated situation*.

Chooser notices that A has put a choice to B, but neglects the payoff B would get if A had not provided that choice. In some of the CR response games, the A choice (no matter which way B responds) would make things much better for B. In the rest, the A choice usually makes things much worse for B, and again whichever way she chooses. Responding to the situation should prompt positive reciprocity in the first case and negative reciprocity in the second. But neglecting the alternative can reverse the response, and the particular games in the CR set happen to be such that indeed usually the affective response would be reversed.

In an influential paper, Zajonc (1980) reported a series of experiments showing an immediate, ordinarily covert, *affective* response to whatever is the focus of attention, which then colors the response to the situation. For a B player who neglected the alternative A could have chosen, it is the bare choice he faces (the truncated situation) that prompts Zajonc's visceral response. And the bare choices B sees do vary in their affective character. The choice 400,400 *or* 750,400 does not *feel* the same as the choice 400,400 *or* 750,375, even though the difference in B's payoff is barely more than 6%. It is not hard to see why that might be so. Making someone else better off at no cost (even

if this makes them now better off than you) is more comfortable than making someone better off than you by making yourself worse off. But we have repeatedly encountered the cognitive force of qualitative as against quantitative cues.

In terms of the cascade effects just introduced, 400,400 *or* 750,400 would prompt the coordination frame in the cascade, but 400,400 *or* 750,375 would prompt the weak cooperation frame, where choice is under "neither selfish nor exploited" tension. A NSNX agent would be seeing the situation from frame 2a, where he might make the generous choice, but unless he has a cue which makes one choice clearly appropriate, he would tend to feel uncomfortable about being confronted with the choice. With a little bit of luck, he would not have been home.

Similarly, 300,600 *or* 700,500 looks stressful relative to the quite similar 200,700 *or* 600,600. Even 0,800 *or* 400,400 does not look stressful relative to 400,400 *or* 750,375, though the cost in payoff of a generous choice is much larger. If so, that would presumably be because a person often faces such choices and comfortably handles them. If you look in your pocket and find two dollar bills you didn't realize you had, you would hardly offer one to the person next to you even if she is your best friend. But if you find two cookies in your box lunch and the person next to you finds none in his box, it would be quite odd if you did not offer to share even if he is a complete stranger. Unless seen as imposed by the other player, a choice between an inferior payoff and no payoff for either does not seem to be stressful. As mentioned earlier, 36 of 36 players chose the positive payoffs from 800,200 *or* 0,0, which hardly seems likely if players found this choice difficult.

A point to note in this discussion is that framing within the cascade relates to the Zajonc effect but does not simply govern it. A choice from the weak cooperation frame 2a is more likely to be aversive than a choice from the strong cooperation (coordination) frame 2b, since there is a "neither selfish nor exploited" tension to be deal with in the first but not in the second. But sometimes a risky choice is easy (there is a risk but you are confident about how you want to deal with it) and sometimes a pure coordination choice is difficult (you suspect someone who will benefit really deserves punishment, but you are not sure).

But, restating the reason for this Zajonc discussion, if B indeed is somehow neglecting A's alternative and responding only to the truncated situation, the Zajonc prompt covertly coloring the B choice could only come from the bare choice A has put to B, which might be the opposite of the Zajonc response to the full situation..

Finally, neglect defaulting would also play a role in A as well as B choices, and of an even odder sort. When a CR game provides an A choice, it is always between an *exit* option (choose *left*), which takes a known payoff, vs. taking the risk of passing the choice of payoffs to B (choose *right*), which yields payoffs which might be better or might be worse than just choosing to exit, conditional on what B then does. A has no way to assess this risky choice if she neglects what B's options would then be. So in the A role, players should escape the *neglect* default. As A, a player has a very strong cue to pay attention. The same holds for the two games in which a third party (C) with no personal stake responds to a choice by A. C acts as referee, deciding whether A gets a very big payoff and B a very small payoff, or the reverse. His only clue to which might be more deserving is the choice A made to set up C's need to respond. So it is again hard to see how this player could be vulnerable to the neglect default. Like A with respect to B's possible response, though for a different reason, C has a strong cue to focus attention on his partner's choice. But players in the B role have no such sharp prod to look at the other player's choice, not just at their own./5

But when A looks at how B would react to the choice if offered, he must be prompted by the same visceral affect as B. If B's immediate feeling about the choice is prompted only by the truncated situation, so will A's, and with the same impression as B. This would be a strong assumption if A and B choices were made by different subjects. But here no assumption is needed. This is a case in which A does not need any actual empathy to be prompted to the same visceral response to the B choice as B would feel. Recalling how the CR data was gathered, in these experiments the person making an A choice is the same person who also makes a B choice. Each subject makes both choices. If the visceral response to that choice is positive (prompting A to feel a tacit sense that a "nice" response from B is likely), that would embolden competitive play when the question is whether to risk gambling for more payoff at B's expense. And it would also

embolden cooperative play when the question is whether to risk own-payoff to try for a better social outcome. All CR response games ask one question or the other.

There is a point of ambiguity in this that the data here do not let us resolve

So now look at the data, keeping in mind that B responses concern what choice they want to govern their payoff in the event A presented the choice to them. So in CR1, that is rare (only 4%), but all B responses are to the A choice that would make a B choice relevant to the payoff.

A reader will notice that many of the results, even looked at in isolation, will obviously be statistically impressive and some others not. But neglecting results that piecemeal do not reach conventional statistical significance is not actually defensible. What most of all needs to be considered is the chance that all the effects reported here will so consistently go the way the analysis expects them to go, which plainly is zero to as many decimal places as anyone would want to know.

(1) Is the evidence clear that B (responders) are mostly neglecting to notice whether A has made a generous or nasty choice? Very much so. In noting that reciprocity seems to be missing in this data, I am only repeating what Charness and Rabin themselves reported. All I have added is that failures of reciprocity as extreme as seen in this data would ordinarily only be seen if either the situation is plausibly an actual zero-sum situation (if your opponent in tennis hits a weak return you do not feel you ought to be nice and hit an easy return back) or if responders were somehow unaware of what another person had done to create the situation. But here the situation is completely described by the game tree the player is shown, not reasonably seen as zero-sum, but somehow normal reciprocity fails anyway. Over and over, A takes a big risk solely to help B, and B's mostly respond with utter selfishness. Or A makes an aggressively selfish move, hurting B, and B's mostly respond by being nice to A.

Of nine games in which B responds to a generous A choice (CR3, 4, 6, 7, 9, 14, 19, 21, 25) only two show a majority of generous responses (CR14 & 19), and as will be seen, in both there is strong reason to doubt that the "nice" responses here are in fact mainly responses to A's generosity. And of eight games in which B responds to a nasty A choice (CR1, 5, 11, 13, 22, 28, 30, 32), just one (CR1) shows clear evidence of

reciprocity for a majority of players, and that in the game that makes the unreasonable character of that A choice especially hard to miss. The fraction of nasty responses to a generous move by A reaches 94% in CR9. And if you look at the games in Fig. 2, you will see there is nothing subtle about the A moves. The generous moves are *very* generous. The nasty moves are *very* nasty.

Nor could it resolve the puzzle to suppose that nasty responses to generous moves just show the importance of self-interest. That would not explain generous responses to nasty moves. Nor where the generous moves ungenerously responded to come from. Nor how such contrasts in motivation could make sense when all these moves are made by the very same people.

As Fig. 1 shows, there are many generous as well as many nasty choices by both A and B. What is puzzling is that B's choices, whether generous or nasty, so often seem inappropriate to the circumstance, and what is puzzling about the A choices is that so often the same player seems cooperative making one choice and narrowly self-interested making another. But the first puzzle could be resolved if indeed B's are frequently caught by the neglect default. And the second puzzle could be resolved if the cue to how a player should be seeing the context suggested by the question he faces is indeed nudging players towards the cooperative branch of the cascade in some games and towards the competitive branch in others. We want to see if the data provide some clear evidence of these effects.

(2) On the earlier discussion, B responses guided by a Zajonc-like visceral reaction to the *truncated situation* are responses to the choice B faces that neglect how that choice would look relative to the situation for B if A did *not* offer that choice. But if the covert visceral response to the truncated situation is aversive, that would color the choice in a way that suggests A does not deserve a reward for presenting that choice to B. Or if B's covert visceral response to the truncated situation is positive, that would color the choice in a way that suggests A does not deserves punishment for presenting that choice. In either case, the visceral response to the truncated situation could be – and given the particulars of the CR games in fact usually are -- the opposite of what a normal reciprocity response to the actual situation would be. The data provide striking tests of whether these effects are present and important.

I start with the most difficult case for this account. In CR1, B responds to a nasty A choice. In CR2, the B choice is the same but A is only a bystander. This provides the one example in this data of a substantial overlap between a normal and a truncated response from B. The effect is really big. Only 7% of responses are generous in CR1 but 50% are generous with the same 400,400 *or* 750,375 B choice in CR2. This is consistent with truncated responses, but also consistent with normal reciprocity. If a player escapes the neglect default, of course he will not surrender even more of the payoff A has destroyed to reward A's aggressively nasty move. But A's refusal to accept the conspicuously fair and efficiency 550,550 looks so conspicuously unreasonable that A's rejection of it might get B's attention even if in these games is hard to catch B's attention. So we might be seeing one game out of 19 CR games with a B response to A which actually yields a normal reciprocity response. On the other hand, a bit later you will see why some large part of what looks like a normal reciprocity effect might be in fact a *truncated situation* effect which in this case happens to coincide with normal reciprocity.

But in every other game there is very little ambiguity or no ambiguity at all. Under a later heading I will consider CR 14 and CR19, where we see appropriately generous responses that are clearly more plausibly attributed to the truncated situation effect, since a comparison is available that crowds out any significant normal reciprocity effect. In CR28 & 32, B faces a viscerally nice choice of whether to improve A's payoff at zero cost to himself. By two to one, B does what is generous, though in each case it is in response to an A choice that is very nasty. In CR22 a truncated view of B's situation looks *really* nice. B can choose to help A and at the same time help himself. B rewards A's exceedingly nasty choice by 97% - 3%, rather than forego a small fraction (1/8) of his payoff to punish a really nasty A move although in numerous replications of the Ultimatum game players are ready to accept much more severe costs to punish much less severe insults. In CR9, B responds selfishly to a really generous A move by a margin of 94% - 6%. And so on.

(3) The clearest evidence for the relative strength of truncated as against normal reciprocity responses from B comes when (a) the truncated and normal responses would go in opposite directions, and (b) the data allow a comparison between B choices with A as bystander and the same B choice responding to A. This yields an unambiguous

prediction that generous choices will increase if one element dominates and decrease if the other dominates. The CR data provide two opportunities to look at this test. Both show the truncated effect dominating any plausible normal reciprocity effect, though certainly at least some subjects in each game are seeing what is actually there and giving normal reciprocity responses, which here would reduce the apparent strength of the truncated situation effect. In both, quite amazingly in terms of what we could usually expect, B's are *more* likely to respond nastily to A when A has been very nice to B than when A has done nothing for B.

For CR6 (responding to an exceedingly generous move from A), normal reciprocity would yield more generosity from B's than in CR8, where B faces the same choice but with A as bystander. The A choice in CR6 definitely sacrifices 50 tokens, and puts an additional 400 tokens at risk, to help B get a 400 token increase in his payoff. But the bare B choice (300,600 *or* 700,500) is not comfortable, as discussed earlier. It is efficient (total payoff of 1200 rather than 900) but perhaps not fair, unless of course B attends to how generous A has been to make this choice available to B. But if caught by the neglect default, B misses that. A truncated response then would yield less cooperation here (where the Zajonc prompt affect would be negative) than in CR8, where A is only a bystander, not the agent who put this aversive choice to B. And indeed B is less generous (25%) to A when positive reciprocity would be in order than when A has done nothing for B (33%). One in three B choices sacrifice to improve the aggregate payoff when A has done nothing to earn that. But only one in four B choices do that in CR6 where only a moral idiot could fail to see what is in order. Since these games were within the same session in Barcelona, the reversal of normal behavior here involves exactly the same players across the two games.

An instructive auxiliary comparison is provided by CR19. This cannot provide an unambiguous result since normal reciprocity and a truncated response go the same way. But the B choice is similar to the choice in CR8, but now comfortable rather than aversive. The B choice is 200,700 *or* 600,600, rather than the CR6 B choice of 300,600 *or* 500,700. In CR 6, the generous choice is efficient but out of context somewhat unfair (and specifically unfair to B). But for 200,700 *or* 600,600 the generous choice is both efficient and conspicuously fair. This yields what looks like a normal reciprocity

response. Generous choices are higher in CR19 (with A presenting the choice) than in CR15 (with A as bystander). But the comparison between CR15 (where 73% of B's are already generous even though A is only a bystander) and CR19 (78% generous) leaves very little room for normal reciprocity once any effect of truncated responses is allowed. Given the striking truncated response effect where the conflict with a normal reciprocity response makes that starkly visible (in CR6), what could be interpreted as a mild normal reciprocity response that coincides with a truncated situation response looks entirely dismissible as just an artifact of that coincidence.

The second opportunity to examine conflicting normal and truncated responses yields the same morally perverse result as in CR6. In CR2 (and its replicate, CR17) half of B choices are generous to A, sacrificing 25 tokens to gain 350 for A, though A is a bystander who has done nothing to earn that. But in CR4 and again in CR5, after A has risked 400 or more tokens to give B a large instead of 0 payoff, the fraction of generous B responses drops to under 40%. As in the CR6/CR8 comparison, normal reciprocity would increase cooperation in response to a really generous A move. But a truncated response to the out-of-context aversive 400,400 or 750,375 choice would go the other way. And again, the data go the other way.

(4) A different test is available when we have a pair of games which are identical except that in one the B choice could be expected to yield a positive visceral (Zajonc) prompt response and in the other a negative Zajonc response. If the truncated situation argument is on target, we should get a clear result, which indeed we do. As background, note that comparing CR5 in Barcelona vs. CR29 in Berkeley, both of which offer the same B choice (400,400 or 750,400), the results are essentially indistinguishable though in CR5 the choice is in response to a nasty A choice while in CR29 A was only a bystander. No truncated situation effect is visible, but also no normal reciprocity effect. But the null effect could be because some fraction of normal reciprocity responses (here negative) are canceling the truncated responses which for this choice would be positive.

And we can see that indeed that is very likely by comparing CR5 with CR1. The only difference between the two games is that in CR1, B faces the aversive 400,400 or 750,375 choice instead of the agreeable 400,400 or 750,400 choice in CR5. B's payoffs from a generous B choice differ by only 1/16. But on the account here, this quantitatively

unimpressive difference shifts the frame in the cascade from 2b to 2a. And the effect turns out to be huge. In CR1 only 7% of B choices reward A for a really nasty move. In CR5 this soars to 67% in response to the identical nasty A move.

(5) In introducing oddities in the CR data earlier in the chapter, I already mentioned the anomalies that can be seen among the trio of games where B's choice is 0,800 *or* 400,400. B shares 55% of the time in CR14 and 56% in CR18, but only 22% in CR26. But the logic of the games says that sharing should be less common in CR14 (where it entails a risk of betrayal) than in CR26 (where it doesn't). And a normal reciprocity response would make B more likely to share with A in CR14 (where A could have just kept the 800 tokens) than in CR18 (where A had no such choice). But none of this happens. Mere common sense fails badly. But if B is caught by the neglect default he would see no difference between CR14 and CR18 (and from the results, it is apparent he doesn't); and if the truncated situation is what guides B's intuition, B would be more generous to A when A's choice puts him in that situation even when (in CR18) A has actually done nothing in any way generous to put him there. Again the data fit an analysis in terms of cascade effects and neglect defaulting.

(6) But, recalling the Falk et al set of experiments mentioned in introducing the CR data, Falk reports a very different result from a test which on its face is essentially identical to CR27. In both, the A choice is whether to accept an equal division of the total payoff (in CR27: 500,500, in Falk: 5,5), or make the risky aggressive move of presenting B with the ultimatum choice between giving 4/5 of the payoff to A or refusing and getting zero payoff. In CR27, 91% of B's then accept the severely unequal spit, far higher than in numerous equivalent "ultimatum" games, where this move could be expected to generate almost as many refusals as accepts. In Falk's game, 44%, rather than the 9% in CR27, refuse what they see as an unfair offer.

But in Falk et al there was a fork in the left as well as right branch of the A choice. A's generous choice was not just to end the game as 5,5 but to let B choose 5,5 *or* 0,0. So the nasty choice was identical to that of CR27, but the generous choice required explicit attention from B, who to get the even split had to explicitly choose it over 0,0. Further, both choices were made as part of a set of 8 choices over four games (with two choices in each game) where the A alternative to 8,2 *or* 0,0 was systematically

varied. The complete set of A alternatives to offering 8,2 *or* 0,0 included the degenerate choice mentioned earlier plus 10,0 *or* 0,0 and the self-sacrificing 2,8 *or* 0,0. So indeed there was a large difference between the CR game and the Falk game in fraction choosing 0,0 from 8,2 *or* 0,0 where A could have offered 5,5. But there was also a more than adequate difference in conditions to account for why truncated responses would be very much more likely in the CR game than in the Falk game.

(7) So far we have considered *truncated situation* effects on B. But for the reasons I've pointed to in setting up this series of tests, there should also be *truncated situation* effects on A, even though A cannot be blind to B's alternatives in the way that on the defaulting argument B can be blind to A's alternative. If as B a player usually attends only the truncated situation, then as A that same player (and recall that in the CR games it is the same player) would also usually see just B's truncated sense of the situation in considering B's possible response to his move. How could it be otherwise, since if looking at B's situation when considering the A move is *not* truncated, how could that fail to inform how the same situation looks to this same player when making the B choice? If A notices a conflict, a closer look will prompt a normal reciprocity response from B, since on reflection no one would be in doubt that a response to the actual situation not the truncated situation is what she wants.

That "look closer" shift in perspective certainly must sometimes happen. There is severely deficient evidence of normal reciprocity in the CR games, but far from a complete absence of it. There are always some, and often many appropriate responses. But if A mostly sees only the truncated situation B faces, a positive Zajonc affect that for B makes it more likely he will make the nice choice would make A more optimistic about taking a risk. This encourage a cooperative choice when the risk on offer is put his own payoff at risk to do better as a group, but it would also encourage a competitive choice when taking the risk might make A better off at B's expense. So we would see more nasty A choices, as well as more generous choices. And all this would point in the opposite direction when the B choice is aversive for B.

Is there evidence of this in the data? An appropriate test here is to consider A's tendency to risk putting the choice in B's hands conditional on whether B's truncated situation is 400,400 *or* 750,375 vs. 400,400 *or* 750,400. As already noticed, the B payoff

is almost the same in either choice, differing by only 1 part in 16. But from a small quantitative difference we have a big affective difference. The most striking illustration comes from a pair of games already discussed but with respect to the B choices. We want to compare A's propensity in CR1 to risk a nasty but potentially profitable move in CR1 to the same nasty choice in CR5, which differs only by that 1 part in 16 in what B will get from a nice response to this nasty move.

Both games are with the same subject pool (both in Barcelona). They differ only in that in CR1, B has the aversive 400,400 *or* 750,375 choice and in CR5, B has the nice 400,400 *or* 750,400 choice. We have already noticed that in CR1, only 7% of B responses to a nasty A move are generous, while in CR5 that skyrockets to 67%. And if, as the argument here expects, A's Zajonc affect when he looks at the B choice is like B's, that ordinarily covert but prompt affect will be negative in CR1 and make A cautious about taking the risk but positive and make A bolder in CR5. Very emphatically, that is what we see. In CR1 just 4% of the A choosers made the nasty but risky choice. But in CR5, 61% of A choosers made the same nasty but risky choice.

The effect is also very apparent in the converse situation, where the risk on offer is to risk own payoff to gain a better cooperative rather than a better competitive outcome. We see a sharp demonstration across three Barcelona games where choosing *right* is generous (not nasty, as in CR1 and 5). As already noticed, the B response to an exceedingly generous A choice in CR3 and 4 by B's facing 750,375 is 62% nasty. But in CR7, facing the nice 750,400 rather than the aversive 750,375, nasty B responses plummet to 6%. So look at A choices in these same games. A generous choice is a little easier in CR3 than in CR4, but the pair yields an average of 22% generous A choices when B will then face 750,375. In CR7, B faces 750,400 with its benign affect, rather than 750,375 with its aversive affect. Generous A choices now more than double, to 53%.

And note that these are indeed truncated situation effects, enormous for B but also big for A. Looking at the actual situation there is no reason for any significant difference in CR3 and 4 as against CR7 for either A or B choices. In CR3 and 4, the B choice would be from frame 2a in the cascade, not from frame 2b as in CR7. So B in CR 3 or 4 would have to resolve the "neither selfish nor exploited" tension characteristic of that

weak cooperation frame. But the situation is one in which it is obvious (responding to an A choice which risked a great deal to help B) that the generous choice is what is called for. B responses to the actual situation would be overwhelmingly generous not overwhelmingly punishing. And an A who shared that sense of the full situation B would face would not be inhibited by a sense that it would be dangerous to trust B because the generous B payoff would be 25 tokens less than in CR7. So without the truncated distortion affecting A as well as B, the difference of more than a factor of 2 between willingness to make the risky cooperative choice in CR3 & 4 as against CR7 would be mysterious.

The inferences here should also apply to CR21, which does show the effects, but just barely. But CR21 is a decided outlier on any view. See the note./6

(8) Can we distinguish whether the effects in (7) are really due to Zajonc's "emotions come first", where conscious processing of a choice is colored (usually covertly) by a very fast visceral response, as against A successfully intuiting how B will choose, and responding accordingly. This is a side issue for the account here. The former seems the more pure Zajonc effect (positive affect encourages A to take a chance). But the latter would also be contingent less directly on a Zajonc effect (A senses that B will be inclined to be nice, which comes from B's Zajonc response to the truncated situation). The adverse defaulting conjecture is not at risk. But the point is still consequential, since seeing which way this possibility goes is likely to prove helpful in interpreting what is happening in other experimental settings.

On the "payoff" view, A is not directly put in a positive mood by the covert visceral prompt, but correctly judges how likely B is to be generous, hence making his risky alternative look more risky or not so risky contingent on his intuition about B's sense of the situation. On the "emotions come first" view, A is influenced by sharing in B's covert visceral response, not by correctly intuiting how B is likely to choose. As with the competition between normal reciprocity and truncated responses, we want to look for situations in the CR data where the payoff interpretation would yield different behavior from the directly visceral interpretation. The data give us two opportunities to test this, both of which strongly go against the payoff possibility and support the direct version.

On the logic of the games a player making the A choice in CR14 should be *less* likely to be willing to share (since B then has the option to betray A and just keep the tokens) than to make the risk-free choice to share as the B choice in CR26. The choice 0,800 *or* 400,400 looks nice for B, as discussed earlier, but in CR26 it implies no particular propensity to be generous. Rather it looks nice because it offers B a guilt-free option to do whatever he feels like doing. By a wide margin (78% to 22%) B does *not* share in this game, as by a much wider margin even very generous people do not ordinarily share half of what they have with random strangers. So if A is responding to his reliable intuition of what B will do, that would further strengthen the already strong inference that the risky offer to share in CR14 must be rarer than the risk-free choice to share in CR26. On the other hand, if it is Zajonc's visceral response that is directly influencing A, the positive affect of this choice would encourage A to take the risk. So we have starkly contrasting inferences. On the payoff mechanism, we get an emphatic inference that the fraction inclined to share must surely decrease between CR26 and CR14. On the direct Zajonc mechanism, it should increase. And in fact the fraction willing to share does increase from 22% in CR26 to 32% in CR14.

We have another opportunity to test for a direct Zajonc effect in the results from CR9 in Barcelona and CR25, its replicate in Berkeley. The B choice in this game is 450,350 *or* 350,450, which is certainly agreeable. And it is not surprising that overwhelmingly, given a guilt-free choice between more for you and more for me I cheerfully prefer more for me. In CR9, 94% make that choice. But this would be a guilt-free choice only for someone who neglected to notice that his payoff would be zero if A had not very generously put part of his own payoff at risk by making it available. In the actual context, it is a nasty choice after A has taken a risk solely to help B. If A is intuiting that perverse B choice (the payoff motive), he would hardly offer the choice. On the other hand, on the direct Zajonc argument A is influenced by the visceral response to the choice, which is positive, not to the likely pick from this choice, which from A's perspective is outrageous. So on the payoff response, A should never offer this choice, but on the visceral response view, A should be tempted to take the socially nice risk. And almost a third do, so that instead of just pocketing 450 tokens they trust B to respond reasonably, which these same players, in their B choices, do not do.

This is the starkest example in the CR data of a gestalt shift that yields what appear to be utterly incompatible responses between a player's choice as A and the same player's choice as B. Since many more A choices are generous than B choices respond to that generosity, it is necessarily the case that a large majority of players who as A trust their partner not to betray their generosity then as B do betray A's generosity. In the trial in Berkeley (CR25) 12 of 32 players risked making the generous offer as A. Nine of the 12 betrayed as B. In Barcelona (CR9) 11 of 36 players took the cooperative risk as A, and as B every one of them betrayed a person who made that generous choice. The fraction of generous A's who betray as B is even larger than the raw data in Fig. 4 suggest since most players did not take that risk, not all of whom betrayed another player who did.

(9) But, as noticed in (8), except for the atypical CR9 & 25, a B choice that prompts a positive visceral sense will always favor a generous to A response.⁷ Consequently the visceral response to the B choice that prompts A to make the risky A choice (and again whether it is a risk to gain payoff at B's expense, or a risk of losing payoff to help B) will prompt B towards indeed making the generous choice. Hence if we set aside the two plays of the atypical game (CR9 & 25), the fraction taking a risk as A in all but these contrary cases should correlate with the fraction of B responses that reward that risky choice. Figure 4 plots all relevant tests, distinguishing the contrary cases just described. The figure speaks for itself.

FIG 3 (or 4 if cascade is reprinted as 3) HERE

(10) In CR10, a third party with no own payoff stake (C not B) makes the final choice facing 400,400 *or* 750,375. C's choices favor the larger but unequal payoff, but by only 54%-46%. So it looks like equal payoffs indeed are very nearly as attractive as larger aggregate payoffs when a cost -- even a small cost -- is imposed on the trailing player, and even when the choice is made by a referee with nothing of his own at stake. Choices favoring the larger but unequal payoffs are barely more common when the referee chooses than when B makes this choice about his own payoff in CR2 in Barcelona and its replicate in Berkeley, CR17. In both trials, B makes the choice equally both ways. But why? If these players could talk they would easily agree that the sensible

preference is for more payoff, and very obviously so in these games where everyone gets both an A and a B payoff.

But apparently in the impoverished environment of these abstract games, even a third-party chooser (C), with no own-payoff stake in the choice, cannot easily escape a sense that they may be making a choice about some continuing game where it would be unfair to give one player an advantage. We see yet another indication that in the impoverished environment of an experiment, players tacitly impute some richer but more familiar context than is actually there, in which they know what to do from ample experience. Sometimes this will be just what is appropriate or just what the experimenters had in mind in designing the experiment. But sometimes it won't, as seems to be the case for half the players here. But when B's payoff becomes slightly larger, so that giving A a larger payoff does not cost him anything (CR29), B is nice to A by far more (69%) even though this still results in a large disparity in payoff favoring A.

We like to think of experimenters providing a context and incentives to subjects whose choices then reveal their motivation, and the consequences of that motivation in the experimental context. But here as in examples throughout this study, we can see that it is the players not the experimenters who provide the context, influenced but not always controlled by the cues that experimenters provide. Even explicit statements are only cues. What governs choice is the tacit sense of the situation inside a subject's head, which is open to influence by more than what experimenters intend their instructions to convey. And a further complication is that if subjects respond to NSNX incentives, their motivation may appear incoherent in terms of standard theory, as it conspicuously does in this data, since no standard theory allows for the gestalt shifts in perspective of the NSNX cascade.

(11) On a related point, since zero-sum play makes no sense in the CR games (there is no payoff for doing better relative to other players), we cannot expect many zero-sum responses. Even when players are prompted to make the A choice from the competitive branch of the cascade, they should be quite easily nudged by a prospect of payoff loss over to the payoff-maximizing frame (1b) of the cascade. But if indeed players making the A choice can be caught by a zero-sum default, we could expect to see some non-trivial fraction of players who are slow to notice that the situation is not one

where a zero-sum choice makes any sense. There seem to be remarkably few mere errors in the CR games.

In two games with obvious B choices for any alert player, they never miss. In CR18, all 32 players see that choosing *right* (passing the choice over to B) cannot lose and might gain, and in CR23, all 36 players see no reason to punish A at a cost to themselves for a one-sided allocation of payoffs in which A is only a bystander. But if mere blunders are rare we can at least roughly estimate the fraction of zero-sum choices typical of the CR games by looking at games where a *right* choice by A can make sense only as a zero-sum move. There are two such games. In CR30, A cannot gain, only harm B by choosing *right*. CR32 the zero-sum motive is even more extreme, since A hurts his own payoff in order to hurt B worse. We see 23% of these perverse A choices in CR30, and 15% even if the more extreme CR32. So we do find a remnant of zero-sum choices, as would be expected as a cascade effect.

But as another indication of the dominance of truncated B responses in this data, note the fraction of B choices which respond with generosity to these aggressively perverse A moves. In CR30, 88% of B choices reward A for his nasty choice rather than give up the small remaining balance of their payoff (200 left after A has refused to let B have 1200). Even in CR32, where it costs B nothing to punish an especially nasty A choice, 65% reward him.

(12) Comparison across games within the same session (hence comparing choices by the same players across games) provides evidence bearing on the reality of the stable “types” commonly reported by experimenters, as in Hauser & Kurzban (20--). As I mentioned in Chapter 1, in terms of NSNX, highly atypical cases aside, these stable types should not exist. Of course there will be a wide range of individual difference, so that some people could be cued quite easily into the cooperative or competitive frame of the cascade and others would be hard to budge. Further every player comes to the game with experience of their own which may make them inclined one way or the other given the immediate circumstances. So it is not surprising that careful experimenters would find what seems to be clear evidence for “types”, especially since the type that is usually by a wide margin the most common is the type that in terms of NSNX everyone should be. A

NSNX agent is by definition a conditional cooperator, seeking to be "neither selfish nor exploited".

But if there are stable types, we should be able to see them in the CR data. The simplest test comes from comparing choices across individuals within a session. I will give just one example of many that can be teased out of the CR data, since as much as I want to do here is to set the stage for a more detailed discussion in Chapter 10.

If we can take the players in CR9 as an unbiased sample of the population (and it is not apparent why not), there would be almost no consistent cooperators and also almost no conditional cooperators, since 93% of B responses violate any plausible interpretation of how such players should respond to a selfless move by A solely on their behalf. And extending the argument to other games would only reinforce that. But since the claim here has been that B choices are distorted by a cognitive illusion (neglect defaulting), I could not want to make the argument based on B choices.

But defaulting effects would not weaken an inference about types from A choices. Intrinsically competitive players would never take a risk to benefit the other player, whether seeing the actual B situation or only the truncated situation. And a cooperative type would not make an aggressively selfish move merely because the Zajonc affect encouraged him to feel he could get away with it. So consider the replicate of CR9 in Berkeley, where the data format is such that I can be confident that in comparing choices across games I am dealing with the same players. The session covers CR25 (the replicate), 26, 27, and 28. Only twelve players (of 32) make the risky choice to allow B to choose the payoffs. There may be other cooperative types among the 32, but too cautious to run the risk of being betrayed. But clearly the twelve bold enough to accept that risk cannot be competitive *types*. Their choice cannot benefit themselves at all. It only risks losing 100 tokens in order to help the other player.

So we can identify 12 players who on the usual taxonomy of types must be either conditional cooperators or committed cooperators. But if that is so, we could predict that in CR27 within the same session, they would not refuse an even split of the gross payoff in order to put an unfair ultimatum to B. And in CR28 they would not destroy 875 tokens that would otherwise go to B on the chance that B will actually reward them for this by 25 tokens. But of the 12 players who make the risky cooperative move in CR25, 11

make an aggressively selfish choice in either CR27 or 28, or both, which I would think would be hard indeed to explain in terms of stable types.

And continuing on that line, I next want to show some consequences of the NSNX view of types as artifacts in other experimental data, and also show what looks to me like a potentially important, intrinsically social, cognitive illusion which can be teased out of the widely-discussed Minimum game.