

HARRIS SCHOOL WORKING PAPER
SERIES 06.05C

CHAPTER 1: THE NSNX MODEL *

Howard Margolis

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

Chapter 1. The NSNX model.

Almost a century and a half ago Charles Darwin (1871, p. 166) argued that natural selection would support the evolution of social behavior, since there would be selection between competing groups as well as selection between competing individuals within groups. In competition between groups, groups in which individuals cared something about what was good for the group, not only about what was good for themselves, would be favored. For biologists, claims of this sort went very much out of fashion after Williams' (1966) classic critique of group-selection. This was not because there is something wrong with Darwin's claim that between-group selection would favor some measure of social motivation. The argument that between-group as well as within-group selection must occur has never been in dispute. No competent biologist doubts it, nor that such selection would favor group-oriented rather than self-interested behavior where the two conflict. The counterargument is rather that the intermittent between-group advantage to social motivation would be swamped by the routine within-group advantage to self-interest. In the language of economics, the argument leaves open the possibility that the result will be a corner solution, or something very close to it.

But even within biology there has been some revival of interest in group-selection in recent years. Among other things, it is hard to make sense of the existence of sex in terms of the self-interest of mothers who squander half their genes in favor of a father. Even where fathers help with child-rearing (in most species they don't) why would not a self-interested female cheat? This is not a knock-down argument for group-selection. But it has never been decisively answered, and various models that respond to Williams critique of group-selection are now in play.^{/1} But whatever the standing of group selection in biology generally, that possibility is of particular interest for the special case of our own species, since the human ability to communicate and to improvise novel behavior makes the potential gains from a capacity for cooperation vast compared to any other species. And we can observe that indeed human beings manage to cooperate (*sometimes*) under conditions and on a scale far beyond anything that could plausibly be accounted for by motivation reducible to indirect self-interest (kin selection or reciprocal altruism).^{/2} Since cooperation far beyond plausible self-interest exists, it apparently is possible. So it is worth considering *how* it is possible.

What is required is an account of how (even in the specially favorable human case) the conflicting between-group vs. within-group selection pressures might yield a non-trivial

sustainable component of social motivation And a key to that (going now beyond Darwin's brief remarks) might be noticing that between-group selection would favor not only some element of social motivation, but more specifically social motivation qualified in ways that would increase the *sustainable* (given the always-present within-group pressure favoring self-interest) level of social motivation.

The NSNX rules

Social motivation favors what is good for the group, even when what is good for the group is not what is best for the individual. And any Darwinian process will tend to favor efficient use of resources. But the competition here between within-group vs. between-group selection would make that tendency toward efficiency particularly strong with respect to making the most of that special resource of social motivation. It would "economize on love" (Robertson, 1956). Which suggests:

Rule 1 (NSNX Efficiency) - Other things equal, the more good a bit of resources would do if used socially (compared with what it could do if used for private interests), the more likely an individual will use that bit of resources socially.

Without a rule of this sort, whatever social motivation can be sustained despite within-group selection favoring self-interest could be squandered, which would reduce whatever capacity this motivation had to resist the contrary pressure from within-group selection. Clearly one aspect of how social motivation would work which would be favored by between-group selection and which would *not* be undercut by within-group selection is the propensity captured by Rule 1. Sustainable social motivation would tend to conserve whatever reservoir of social motivation is available for when it would be most likely to be well-used.

But Rule 1 itself needs some qualification. *Sustainable* social motivation must operate in some way that shields the most socially-oriented members of the group from fatal exploitation by those most inclined to self-interest. I give a detailed discussion of this in Chapter 3 of SA&R. But the outcome is that an essential element of how social motivation is likely to be governed would include a second rule making social behavior less vulnerable to exploitation (hence the label for the model: "neither selfish nor exploited"). Specifically::

Rule 2 (NSNX Equity): Other things equal, the more an individual has already used resources socially relative to what others ---- in particular what

others who look "like me" or "in my situation" -- are doing, the more weight will go to private interests in spending the next bit of resources.

Rule 1 would incline a person towards cooperation to the extent her contribution looks (to that person) to be socially useful relative to its private cost. Other things equal, the more socially useful the behavior seems to an agent, the more willing she would be to do more. But the greater the private cost of that contribution, the less willing she would be. No plausible account of social behavior could fail to incorporate some functional equivalent of this rule, turning on the perceived ratio between the social values versus the private cost of a contribution to social goals. The private cost must include expectations about rewards or punishments associated with a choice as well as 'out of pocket' costs. The (net) cost of giving needs to be reduced if there is a prospect of reward and the net gain from free-riding needs to be reduced if there is a prospect of penalties. And actual social choices commonly involve some private as well as some social value. SA&R treats this in some detail, and some aspects of this will come into the discussion in Chapter 5.

But Rule 2 says that an individual response to social concerns will also depend on how a person's sacrifice of private for social values compares with what others are doing. When almost everyone is cooperating, it will usually be less risky to sacrifice some self-interest in favor of social cooperation than when few are cooperating, and also with usually better prospects.

And as I have been stressing, the Darwinian competition between within-group and between-group selection would favor some functional equivalent of both these rules. Neither rule is novel. Their equivalents have a long history. With respect to Rule 2, for example, more than two centuries ago, James Madison remarked (about federal arrangements): "... distrust of the voluntary compliance of each other may prevent compliance of any, although it should be the latent disposition of all." Madison (1787, note 7). He would not have been puzzled by Rule 2 here. And although a skeptical reader need not accept the Darwinian argument on anything more than an "as if" story, I want to discourage that, since thinking about the Darwinian logic often proves to be useful in interpreting the rules.

The simplest specification of an equilibrium condition consistent with the two rules would be:

$$W = G'/S' \quad (\text{NSNX equilibrium})$$

where G' is the value to the group of a contribution to the group (as perceived by the individual facing the choice), S' is the marginal private cost to the chooser, and W is the weight (never less than 1) given to self-interest.³ The value ratio (G'/S') can be thought of as especially tied to the “neither selfish” aspect of NSNX, and the weight to self-interest (W) especially tied to the “nor exploited” aspect of NSNX. This oversimplifies but perhaps helpfully as a first cut.. NSNX emerges from the interaction between the two rules, not as “NS” from G'/S' and “NX” from W . But thinking in terms of this simplification may make it easier to see why, if $W < G'/S'$ then an increased allocation to group-interest, which increases W (from Rule 2) and ordinarily also decreases G'/S' , moves the player towards NSNX equilibrium, though perhaps over- or undershooting the equilibrium. And the converse for $W > G'/S'$. From choices observed in entirely unrelated settings (not involving the group-interested vs. self-interested tension) we might expect that the guiding ratio of group-interested vs. self-interested spending will be some “local” (context-dependent) sense of that rather than a global accounting. So when we consider revealed behavior in Public Goods games (Chapter 8), it is not surprising that we see “neither selfish nor exploited” responses heavily influenced by what has been happening in the game at hand though that must be wholly inconsequential in terms of some global accounting.

This NSNX equilibrium looks like, but (as already stressed in the Introduction) *is not*, the usual 1st order condition for optimizing a utility function. The topic is treated in detail in Chapter 2, where I show (among other things) that *any* model which stays with the fundamental commitment of the standard theory to maximizing a single utility function implies empirically implausible behavior.

The “see-saw” figures here illustrate this equilibrium notion. Consider a sample citizen (call her Ellie), and simplify a bit by considering dollars the only resource at issue. From Rules 1 and 2, Ellie will allocate the dollar to social G -spending if $W < G'/S'$, and to self-interested S -spending if $W > G'/S'$. Figure 1 provides a picture of the intuition that drives the model (the balancing of the conflicting pressures from the *value ratio* in rule 1 and the *participation ratio* (how much the agent is already spending socially), which governs W in rule 2. Alternatively, we

could write the same equilibrium condition as $WS' = G'$, which prompts allocation to self-interest if $WS' > G'$, to group-interest if $W < G'/S'$, and is just in balance (in equilibrium, and so Ellie is not moved to shift the balance of her spending one way or the other) when $WS' = G'$. Later on, we will have occasion to use that form in working out the notion of social equilibrium mentioned in the Introduction, to complement the notion of individual equilibrium being developed here.

Fig. 1 here

In figure 1, the value ratio (G'/S') is on the right of the see-saw, and the weight to self-interest (W) is on the left. As Ellie uses resources, or as things happen to Ellie or to Ellie's society, G' , S' , and W all may change. If the seesaw is tilted down to the left, then she would allocate a marginal dollar to self-interested S -spending, if tilted down to the right she would allocate it to group-interested G -spending. She is in equilibrium, feels she has done her fair share, feels neither selfish nor exploited when the seesaw is in balance. SA&R shows how the account generalizes to treat various complications, especially the common (in fact usual) situation in which a choice provides some value to both group- and self-interest.

Fig. 2 here

A second seesaw diagram shows the dynamics of the situation in an alternative way. Here (fig. 2) S' is on the left, G' on the right, and W is the fulcrum. As you will remember from playground days, the heavier person does not necessarily tilt the seesaw her way. How the allocation goes depends on how far the fulcrum favors S' over G' . The farther to the right the fulcrum is, the greater the leverage given to S' . Other things equal, if Ellie allocates the dollar to G -spending, that pushes W more to the right (rule 2), increasing the leverage of S -spending in allocating the next dollar.

Or, finally, Figure 3 shows (on the vertical axis) scales for both the weight (W) and the value ratio (G'/S'). On the horizontal axis, we have the fraction of Ellie's resources that have been allocated to G -spending (g). At the origin Ellie is spending nothing at all socially, and at the right she is spending 100 percent of her resources socially with nothing left for her private interest. So figure 3 plots both W and G'/S' curves on the vertical axis against the fraction of resources allocated to group-interest on the horizontal axis, where the horizontal axis runs

from 0 (complete free-riding) to 1 (everything is spent for the group). This yields a diagram that looks like the familiar supply/demand diagram. In that market equilibrium diagram the axis that carries two variables (quantity supplied, quantity demanded) is the horizontal axis. Supply increases with increasing price, demand decreases, yielding the market equilibrium at their intersection. The NSNX rules yield an equilibrium at the intersection of the W and G'/S' curves. But here the intersection defines an individual equilibrium between self- and group-interested spending, not the social equilibrium of market prices and quantities of a supply/demand diagram. The Schelling diagram mentioned in the Introduction does define a social equilibrium, but as you will see (Chapter 5), it does not look anything like a supply/demand diagram.

Fig. 3 HERE

I've drawn the figure with two W -curves (solid W and dashed W^*) and two value ratio curves (G'/S' and G'/S'^*). But for now consider only the solid curves.

The G'/S' curve illustrates the usual situation, where diminishing marginal utility holds. For that usual case, G'/S' must decline as we move to the right in the diagram. For Ellie is spending less on herself, so S' (marginal utility of private spending) must be increasing as the contributing individual digs deeper and deeper into what otherwise would be her private spending. And although in the usual situation, G' will also be decreasing, so long as Ellie is one person in a large society, G' would not ordinarily be noticeably affected by her own spending. For any single person, G' is ordinarily perceptibly constant, as in the standard supply/demand diagram, the spending of an ordinary individual by itself has no perceptible effect, though the equilibrium is determined by the aggregate effect of all. But with G' perceptibly constant and S' increasing as Ellie's allocation moves to the right (towards more for the group, hence a smaller share of spending on herself), the marginal utility of giving up even more is increasing. The value ratio (G'/S') consequently, will be getting smaller. The curve for G'/S' will be downward sloping, as shown.

The rising curve measures W , since by definition W increases as g increases. We can set $W = 1$ when the share of Ellie's resources allocated to G -spending is zero (so the participation ratio is zero). But from rule 2, other things equal, W must increase as the participation ratio increases. Equilibrium in the figure occurs at Q_1 , where $W = G'/S'$.

Suppose Ellie found herself at Q_2 , where for the solid curves W G'/S' she is out of equilibrium on the high side of social spending. By shifting resources to favor her private interests (moving left in the figure, allocating less to G -spending), she could restore herself to equilibrium. For moving her left in the figure would decrease W and increase G'/S' , reaching equilibrium at Q_1 .

Alternatively, suppose Ellie's wealth increases. Then for any fraction of wealth allocated to G -spending, the amount left for S -spending will be larger, hence S' smaller. As before, if Ellie is an ordinary citizen in a large society, G' would still be sensibly constant. So with G' unchanged but S' decreased at any share allocated to G , the increase in Ellie's wealth would shift her value ratio upward (say to the dashed G'/S' curve in fig. 3), and her equilibrium allocation would therefore shift to the right, to Q_2 . So an almost immediate implication of NSNX is that, other things equal, as wealth increases, the share of resources spent on what an individual takes to be group-interested spending increases. Social spending turns out to be a superior good, though in another sense Ellie is also becoming more selfish: the weight to self-interest, W , is increasing.

The two W -curves (solid, dashed) and similarly two G'/S' curves allows some simple comparative statics. Of the two W curves, the solid is more favorable to contributing. Since it is lower, it intersects a downward-sloping G'/S' curve further to the right. Both W -curves (repeating that point) necessarily slope up, since W is the weight to self-interest (always ≥ 1), which (from rule 2) increases as the fraction committed to social spending (on the horizontal axis) increases.

Parallel to the remark about the two W -curves, of the two G'/S' curves shown the dashed is more favorable to social spending, in the way required by Rule 1.

Suppose again that Ellie's initial situation is that of the solid curves, with equilibrium at Q_1 . Now her neighbor's house catches fire. On any plausible account, the value of acting socially must increase when your neighbor's house catches fire. G' must increase. The W -curve is not shifted, but the G'/S' curve must shift up (G' 's is bigger but S' is the same), here to the dashed G'/S'^* curve, implying the shift to the less self-interested equilibrium at Q_2 .

Alternatively, suppose an actor is initially on the solid W -curve as before, but on the dashed G'/S' curve. He is in equilibrium at Q_2 . But now his own house catches fire, which suddenly increases the marginal value of effort in his own self-interest, shifting the G'/S' curve down, say to the solid G'/S' curve, with an obviously reasonable implication for how the actor

will react. Perhaps he was going to take his Boy Scout troop on a hike. But now he finds he has better things to do with his time.

Next, suppose again the initial situation finds an agent facing the dashed W and solid G/S' curves, with equilibrium Q_0 . But now his country is attacked. G' shifts abruptly up, moving the actor to G'/S'^* with equilibrium Q_2 . But this “social emergency” shift in marginal value of social effort affects everyone, and the W curve depends on how much social effort this actor is making relative to what others are doing, and now others are doing more. Consequently the whole W curve must shift up, say to the dashed W -curve, resulting in the even higher equilibrium allocation at Q_3

Simple illustrations of this sort show the basic NSNX mechanics, and the results are easy to interpret as the kind of behavior we would expect if the model is workable.

And what if there is a hurricane or earthquake so all houses, including your own are simultaneously damaged? Now the model is ambiguous. G' is sharply increased. But since your own situation is worse, S' is also increased. What happens to G'/S' will be contingent on the local circumstances. And in fact communities are observed to react in dichotomous ways to such calamities. People do not go on behaving (with respect to competing social and private concerns) as if nothing had happened. Many but not all communities exhibit an exceptional amount of cohesion and commitment to common interests. On the eve of the first game of a World Series, an earthquake near San Francisco forced cancellation of the game (in Oakland, across the bay). Around the same time, a hurricane struck Charleston, SC. News coverage showed people in San Francisco rushing into the streets to help their neighbors, and people in Charleston rushing into the streets to loot. The dichotomous response to such calamities is very well documented (as in Quarantelli, E.L. & Dynes, R.R. 1977). So that the model does not give a general result on this point is not a weakness. The it-could-go-either-way implication corresponds to what we see in the world.

As already mentioned, a later chapter works out an extension -- by way of Schelling's ingenious diagrams -- to dynamic social situations. This yields a set of tipping point processes which find ready application to real cases of the evolution of novel social norms, establishment of new religions, ethnic conflict and political revolutions. The could-go-either-way implication for a situation of general calamity applies to all these cases, as indeed it must unless there is something fundamentally wrong with the model.

Another way to sharpen intuitions about how NSNX works is to consider what happens, on the logic of the model, under odd conditions. It sometimes happens that G' is negative. Under circumstances in which it is awkward or risky to say "no", a person might be asked to contribute to a political campaign that she privately intends to vote against. Or with vastly more severe consequences, she might be asked to collaborate with an occupying army she passionately hopes will be driven out. So she sees the social value of cooperation (G') as certainly negative. But the net private cost (S'), considering adverse consequences that can be escaped by cooperating or a reward that might be obtained by cooperating, might also be negative. It is then declining to cooperate, not cooperating, that accepts some private cost. So although S' , like G' , is ordinarily positive (there is usually a social gain, as judged by the chooser, in cooperation, but a private cost), this can be reversed. As just illustrated, the cooperation at issue could be, to the chooser, social perverse (so G' is negative), but the reward obtained or risk averted from cooperation could more than offset the direct cost (so S' is negative).

And sometimes there may be sufficient offsets to the direct cost of cooperation to make S' negative when G' is positive. You must contribute to what you regard as a good cause to attend a party. But if you are keen enough on the party, the price of contributing may be more than offset by what you would pay even if no cause you valued was to benefit. So here in another way, the value ratio (G'/S') in the NSNX equilibrium condition could be negative.

This yields three atypical cases ($G' < 0$ but $S' > 0$; $G' > 0$ but $S' < 0$; and $G' < 0$ & $S' < 0$). In each either the value ratio is negative (because G' or S' but not both are negative), or the value ratio is positive in an unusual way (because both G' and S' are negative). But each implies easily-interpreted and appropriate behavior. For either of the first two cases, G'/S' is negative, hence less than W , since W by definition is ≥ 1 . So the chooser will allocate to self-interest. But for $G' < 0$, $S' > 0$, it makes sense that a chooser will not sacrifice privately to damage her sense of group-interest. And for $G' > 0$, $S' < 0$, a self-interested choice again makes sense, since now there is no conflict between choosing in self-interest and choosing in group-interest. Cooperation in this case has a negative cost: it is actually profitable. So NSNX equilibrium choices in both these cases are self-interested (since $W > G'/S'$), but in the first case, chooser does not cooperate, and in the second she does cooperate, in both cases as makes sense.

And for the $G' < 0$ & $S' < 0$, for sufficiently small negative cost we would have $W < G'/S'$ (since for $S' = 0$, G'/S' would be infinite, and for an arbitrarily small negative increment to S' ,

G'/S' would still be greater than W). So up to some point (until $W = G'/S'$) chooser will favor her sense of group-interest. She will not cooperate with an activity she sees as socially perverse, even though she could gain by doing so. But for sufficiently large negative S' , we must eventually reach $W > G'/S'$. Then chooser will cooperate even though she recognizes that it is inconsistent with her sense of group-interest. If what made S' sufficiently negative (what made the cost of cooperation negative) is a reward for cooperation we could say she was bribed. If what made S' sufficiently negative (what made the cost of cooperation negative) is avoiding a punishment by cooperation we could say she was coerced.

But that overstates the case. We can observe, not often but unmistakably, that even the most severe cost to self-interest does not necessarily make cooperation the dominant choice. An exceptional chooser may see G' as so extremely negative that cooperation would leave that chooser even further from NSNX equilibrium than would a choice that would avoid cooperation with what (to this chooser) is a highly perverse mode of cooperation. And we can recognize such extreme situations, most clearly for individuals who choose extreme suffering or death rather than cooperate with their captors. This becomes highly relevant when we come eventually to think about application of the argument to the challenging case of terrorism.

Since the S -function in the denominator of the value ratio (G'/S') is just the self-interested utility function of standard theory, nothing needs to be said about it beyond stressing that point. The S -function is indeed strictly self-interested. All compromises of self-interest in favor of other-regarding motivation enter through the G -function as moderated by the weighting function, W .

The G -function needs to be explored, but its salient components are entirely unsurprising. There are three inputs: (1) We can look at the norms that rational individuals might jointly adopt if they could mutually bind themselves. Most obviously, social preferences that would make everyone better off (pareto-improvements) must have positive value in the G -function, (2) But on many matters, multiple equilibria are available, so we would often need to observe revealed social preferences in the actor's society. And (3) path-dependence will certainly sometimes yield prevalent behavior that fits neither (1) nor (2) but an acceptable explanation calling on path-dependence must offer a plausible story (or better, but it is not always even a possibility, actual historical evidence). Plausibility may lie in the

eye of the beholder. But the qualification is not vacuous. If a phenomenon is seen widely across many cultures, it will not be easy to find a path-dependent explanation. And concerns that this is too loose (in particular any looser than allowances taken as reasonable within standard theory for path-dependent oddities in the realm of market and auctions) do not arise unless turning to path-dependence to account for anomalies turns out to be common. But it doesn't.

So the G-function is not open-ended. Individuals will vary in what they see as social value, but (without embarrassing the theory) not in ways that would easily violate both (1) and (2). And falling back to (3) is something that is inevitable even if the theory as I am presenting it should be perfect, but it is something that cannot happen very often. The complications should not require any more generous allowance for context-dependence and path-dependence than would be taken as unobjectionable in the case of mainstream economic theory applied to core domains such as markets and auctions. To give a couple of presumably uncontroversial examples: No one feels bound by social norms against deception while playing poker. And while tipping is common to many societies, the appropriate amount and occasion for tips emerges in a path-dependent way that leads to different expectations in different societies. I will say a bit about both examples in Chapter 3 (on norms) as well as about some odder cases, such as why in no society is it treated as reasonable for people to sell their place in line to a late arrival. If you have plenty of money and I have plenty of time, we may find a deal to make. But we know it is not really the way people are expected to behave, and we will keep it out of sight. On the other hand, people are almost always more tolerant of ticket scalping, which is close to the same thing. So can we say why that might be so? I will try to do that in Chapter 3.

What do criteria 1 and 2 of the previous paragraphs imply? At the most general level they imply that NSNX agents will value *efficiency* and *equity* and (of course) *self-interest*: *efficiency* because other things equal, more resources for group-interest is better than less; *equity* because a sense of unfairness in distribution of resources could only harm group solidarity and increase resource-dissipating conflict; and attention to *self-interest* follows directly from the basic Darwinian argument.

Here are some rules of thumb that develop these ideas, always subject to the *other things equal* proviso.

A. A bigger pie for the group is better than a smaller (which just restates the bare efficiency principle.)

B. Equally deserving members of the group ought to get equal slices (since what could be as fair.) This leaves a lot of context-dependent and path-dependent room for interpretation of "equally-deserving", but not unlimited room since interpretations are constrained by criterion 2.

C. Diminishing marginal utility implies that the value of helping a member increases as that member is worse off than others in the group.

Criteria (1) and (2) also imply that norms of fairness and reciprocity will be positive values in the NSNX G-function. But just how they operate in practice here (or, equally, within a more standard model where fairness or reciprocity might be arguments in the individual utility function) is not implicit in the theory, especially allowing for cognitive complications, which might be severe as will be grossly apparent in some of the experimental data we will take up in later chapters.

Allowing for complications (like the possibility of overshooting, of gestalt shifts, and others) a prediction could be only that the tendency (with $W < G'/S'$) will be to allocate more to social values, or in a comparative statics inference across situations, give as much or more as in the comparison situation. But as will be seen, comparative statics is sufficient to yield many predictions with bite.

To conclude this introductory sketch of the model, here are some simple illustrations of the NSNX logic. Each deals with some familiar puzzles among economists.

Suppose you are in a restaurant in a distant city and the people you are with are trying to figure out why they will tip even though they happen to people who tell each other they know it is irrational to tip in such a situation. But if normal motivation is NSNX, then there is no puzzle here. A waiter's tip is part of, indeed most of, his compensation. Everyone knows that. So not leaving a tip, even though it is legal not to leave a tip, is unfair, and everyone knows that. A society in which people feel free to abuse elementary standards of fairness would not be a nice place to live, so there is special social value (elevated G') to resources used for an appropriate tip

relative to just giving some of your money to a stranger. Since leaving the usual tip is virtually universal, W will be low: a person who doesn't will feel selfish, and a person who does will not feel exploited. And a tip, being a tip, is not a great compromise with self-interest, so S' will not be high. But as will be discussed in Chapter 3, this equilibrium would be under stress when many people -- restaurants seem to always judge *six* as where "many" begins -- share a common check, and voluntary tipping comes under stress.

Or consider the puzzle of repeated games of Prisoner's Dilemma. Suppose there will be 100 plays, with a payoff of \$100 for each play when both cooperate; \$0 for both defecting; and \$101 for a defector against a cooperator, with -\$1 for the cooperator. On a backward induction argument, perfectly rational players would defect every time, and end up winning nothing, since they would realize that at play 100, with no future opportunity to be punished, each will defect. But since defection is certain at play 100, there is no reason to cooperate at play 99. And so on. But two naive players would be likely to win \$10,000 each. Bringing this puzzle to wide attention half a century ago, Luce & Raiffa (1957) allowed that rationality here leads to a recommendation that would be stupid, but they were not able to exactly explain why. The favored explanation (of why a rational person could cooperate, contingent on the other player also cooperating, at least up until some point near the end) is that it is not certain that the other player will be rational (Krebs, D. et al, 1982). This "trembling hand" account provides a rather shaky foundation for a model of rational choice. It implies that the more confident players are that they are playing with someone competent, the less money they will make. This seems neither reasonable nor true. That the trembling hand account has become the standard explanation reveals how difficult it is within the standard account to avoid a pragmatically intolerable inference (that the correct play is to defect at move 1).

In terms of NSNX, the dilemma does not arise. The backward induction that creates the dilemma has each actor seeing it as obviously rational to prefer \$110 for himself and -\$1 for the other player to \$100 for each, even if they had cooperated fully for the first 99 plays. That each player has just won \$9900 by cooperating on the previous 99 plays is irrelevant. But one of the things that generates a sense of group loyalty (borrowing a bit from a cognitive discussion to come) is working together on a project -- especially working together against a common adversary, though that frequently important aspect is irrelevant here. If NSNX is right, players who have been enjoying enormous success from cooperation will not be indifferent to the other

player's situation.

By now they are partners in a very successful enterprise, and it would seem quite bizarre if either was seriously tempted to want to end by defecting since he had no further use for his reputation for cooperating. Perceived G' would be high relative to a game played only once; and with both players far ahead of their starting point, perceived S' (the cost of cooperation relative to the gains in hand from cooperation) would be low. And W would be low. Neither player is doing a bit more than everyone else is doing. Cooperation would be likely to be felt easily more comfortable (closer to equilibrium) than defecting. And supposing the contrary to be sufficiently certain to warrant defecting on move 1, would seem quite idiotic, with no trembling hand needed. It is failure to cooperate that would be a puzzle, which indeed sometimes occurs in experiments with this game. We will see later other examples of failures of cooperation under conditions where cooperation should not be very difficult. That poses a puzzle for NSNX, with interesting consequences. But nothing as fragile as the trembling hand is needed to manage the difficulties.

Finally, consider a bit more elaborate situation which has been the focus of more recent discussion in the American Economic Review. The “traveler’s dilemma” (Basu, 1994) engages two players who must (without communication) choose a number between 180 and 300. Each will get a payoff equal to the lower number chosen, but unless both make the same choice the payoffs are adjusted to punish the player who bids more and reward the player who bids less. Some amount – call it R -- will be taken from the payoff of the higher chooser and given to the lower. So this is a mixed motive game akin to the Prisoner's Dilemma. And again on the standard theory a backward induction argument yields as the unique Nash equilibrium that no matter how small R might be, both players should choose 180./4 So again, completely rational play, on the standard account, yields each player the lowest possible payoff. Basu observed that this hardly makes sense in terms of how people would plausibly behave when R is small.

Subsequent discussion has tried to resolve this difficulty and other apparent departures from strict rational choice by a process (QRE: see McKelvey & Palfrey, 1995) which turns on each player calculating optimal responses on the assumption that other players are *not* making optimal responses. So there is a bootstrapping effect in which everyone assumes that everyone else is vulnerable to error (akin to the “trembling hand”) on their primary response in a simple situation, and then calculates a sophisticated response to that.

But convolutions of that sort would not be needed if players behave as NSNX would require. With small R the potential gain to both self and group of choosing high is large and the risk of exploitation is low, but the converse when R is very large. So while standard theory predicts 180 as the equilibrium choice whatever the value of R , with NSNX "neither selfish nor exploited" motivation, for sufficiently low R both players become likely to choose high. Both numerator and denominator of the NSNX equilibrium condition ($W = G'/S'$) move interactively in favor of the socially better choice in a comparative statics assessment of G'/S' for $R = 180$ vs. $R = 5$. Where it is an embarrassment for standard theory that players choose high for $R = 5$ but low for $R = 180$, the same result is just what must be expected if NSNX is right. Goeree and Holt (2001) ran trials with $R = 180$ and $R = 5$ finding that the very high penalty indeed yielded a modal response of 180, but $R = 5$ produced a modal responses of 300, with a bit more than 2/3 of all responses at 300 or 299.

On the other hand, puzzles like the three reviewed here do not provide a sufficient argument for taking NSNX seriously. NSNX easily handles them. But so would almost any model that broadens the strictly self-interested motivation of the standard theory to allow for other-regarding motivation. In general, if we are focused on some particular case, or some restricted range of cases, it will always be possible to define a less radical departure from received theory than NSNX that will do whatever is needed to handle what is on the table. Postulate a propensity to value reciprocity, for example, and either puzzle just discussed has a solution without any departure from standard theory as drastic as the dual-utilities of NSNX. So if there is a case to be made for NSNX, it has to turn of how this theory is able to handle the full range of other-regarding choice we can observe *better* than a more conventional alternative.

APPENDIX: Similarities and contrasts

A few comments on a particularly prominent review (Fehr & Fischbacher 2002) may be useful, noticing both some points where NSNX coincides with and where it contrasts with the FF survey and interpretation of the experimental work.

As already discussed, we certainly must expect that pareto-improving choices would be favored in a NSNX world: obviously so given a neutral effect on self-interest (on the S-function), and significantly so even when there is a cost to self-interest. And we should expect that pareto-damaging choices would be discouraged. Both points follow from the earlier

discussion of the G' -function: clearly, rational actors explicitly agreeing on norms to guide a group would agree on these. Although we can find what seem to be pareto-damaging norms (potlatches, and a few other cases), these are certainly odd, as their notoriety attests.

Positive reciprocity, which allows a society gains from trade which transaction costs would otherwise render infeasible, should certainly be favored. Further, criteria (1) and (2) require that this extend to a *strong* form of reciprocity, where a socially-motivated choice helps someone (possibly a stranger) who is never likely to be in a position to repay the choice, but in a context where prevailing norms warrant anticipation that on another occasion, were the situation reversed, someone would return the help. And we would expect negative reciprocity also to be evident but less marked. The social value of punishing those who fail to follow social rules (2nd order norms) has often been noticed in discussion of rational norms. But for norms of punishment the possibility of errors is more serious (perhaps there was a misunderstanding, perhaps there were extenuating circumstances), and if that might be so there is an enhanced possibility of that the punishment will turn out to be socially perverse, or that it will turn out to have unexpected costs to the punisher. Risk aversion would play a role here as well. For positive reciprocity, uncertainty about possible secondary effects is essentially all on the upside: good behavior might have been noticed beyond what the actor expected. But for negative reciprocity there is a clear downside risk as well: socially-motivated punishment might be seen by some (and not only the target) as unreasonable or mean or badly judged. Overall, negative reciprocity ought to be less well-marked than positive reciprocity.

Finally, on two other possibilities often discussed in the recent literature, a propensity toward allocating marginal resources to favor equality across equally deserving people ought to be clear, but not ordinarily a propensity toward harming those well-off merely to move toward equality (difference aversion). For the former favors helping those most in need, which would ordinarily be pareto-improving in utility, but the latter will be pareto-damaging.

FF do not explicitly comment on the last of these points. The experimental evidence is mixed (see the discussion in Charness & Rabin 2002). But as in a NSNX world, FF find clear and persistent evidence for strong positive reciprocity, and somewhat weaker effects for negative reciprocity. This in fact is sometimes grossly defied in experiments, with the just-mentioned Charness & Rabin paper a striking example. Experiments which seem to defy reciprocity norms need an explanation given the FF survey, and for our purpose they especially need an explanation consistent with NSNX. We will take up the really striking example provided by Charness & Rabin in Chapter 9.

FF stress the potential of a minority of free-riders to over time lead to the disintegration of cooperation when no opportunity for rewards or punishment is available, and they see strong evidence that the ability of others to reward cooperation or punish free-riding can sustain cooperation (as in the lessened sharing propensity in dictator as against ultimatum games).

On the most fundamental points NSNX is consistent with the FF reading, and both also are largely consistent with what shrewd observers have noticed over the centuries (as in the Madison quote earlier). For example, the drop in sharing which FF note (but not nearly to zero) between the ultimatum and dictator games necessarily holds under NSNX. For in ‘ultimatum’ the net cost of sharing is *less* (relative to ‘dictator’) due to the risk that what the chooser keeps she might not actually get. In other words, for any given level of sharing, S' adjusted for risk is smaller in ‘ultimatum’ than in ‘dictator’, where there is no risk. But G' is not smaller. Hence in terms of the NSNX diagram (Figure 1) the G'/S' curve is shifted up in ‘ultimatum’ relative to ‘dictator’. And to the extent that a player’s expectation of what others – including the very recipient in this choice! -- would give in this situation is (correctly, after all) higher for ‘ultimatum’ than for ‘dictator’, then from Rule 2 the W curve would be shifted down. So the situation would be parallel to the ‘wartime’ illustration in the comparative statics exercise earlier, with the clear effect on propensity to give that FF observe.

FF see the theory and evidence they review as suggesting ‘that a combination of altruistic and selfish concerns’ is what motivates choice.... ‘if this argument is correct, we should also observe that altruistic acts become less frequent as their cost increases’ (Fehr & Fischbacher 2003, p. 788) as is transparently correct if the pair of NSNX rules and the equilibrium they define are sound.

But on two points there is some contrast between the NSNX and FF views. The first concerns the important role given to contrasting ‘types’ in the FF account and in many other economic models; the second concerns the role of gene-culture interactions.

On the NSNX account, pathological cases aside, there is only one ‘type’. Players may vary a great deal, contingent on individual differences, on conditioned expectations of how others will play, on (perhaps subtle and unintended) differential reading of contextual cues, and more. But in terms of NSNX there is no division between selfish types who cooperate only if severely enough threatened with punishment, and cooperative types who are willing to bear costs to punish free-riders. All players are NSNX players. This claim has consequences, and those immediately at hand seem to me to favor it. For example, as will be seen in the third of the tests to be discussed, there are very large differences in the frequency of complete free-riding, contingent on game parameters, even when players are randomly drawn from a common pool. This makes sense in terms of NSNX, but may be hard to explain on a notion of fixed types. A more familiar kind of evidence is the powerful effect of communication in these games. This yields a quite huge increase in cooperation, which makes sense in terms of NSNX, where the usual decline of cooperation over a sequence of rounds is substantially due to coordination problems: players who want to be neither selfish nor exploited face difficulties in feeling confident about how far others are playing the same game they are. But it may be hard to make sense of this powerful effect of ‘cheap talk’ if there are fixed selfish and altruistic types. We will see (in Chapter 9) a striking example of how far from what might be expected from stable types experimental evidence sometimes reveals.

On the gene-culture point, FF argue that Darwinian influences alone cannot account for socially-motivated departures from self-interest. But on the NSNX argument these are deeply-entrenched human propensities that surely are influenced by gene-culture interactions but do not fundamentally depend on that. Indeed how could we otherwise explain why the propensities FF report are found in *all* cultures for which solid documentation is available? One possibility is that what we see may be a result not of group-selection as usually thought of, but of higher order selection at the species level. Of the several (or perhaps even many) proto-human species in the past several million years, one turned out to dominate all others. But since the advantages of cooperation are vast for a species able to communicate and improvise, a species which was genetically bound to a socially-motivated component of

behavior would have an enormous advantage over sibling species for which the inevitable and persistent within-group competition favoring self-interest could swamp the between-group social advantage. The cross-group migration that makes it difficult to model Darwinian social motivation would be absent when it is migration across species that would be needed.