

HARRIS SCHOOL WORKING PAPER
SERIES 06.05B

INTRODUCTION *

Howard Margolis

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

Cognition & Extended Rational Choice

Howard Margolis, University of Chicago

TABLE OF CONTENTS

Introduction

1. The NSNX model
2. Dual-utilities
3. Norms
4. The Schelling diagram
5. Using the Schelling diagram
6. Adverse defaults
7. The NSNX cascade
8. Public Goods experiments
9. Reciprocity puzzles
10. Social illusions
11. What we see in the world that looks like what we see in this theory: the case of terrorism

Appendix: The data template

Introduction

It is now 25 years since I published a proposal (Margolis 1982) for broadening the standard model of economic theory to build in what few people would want to deny: that human beings respond to interests that go beyond self-interest. Even then, what I later labeled the NSNX ("neither selfish nor exploited") account was not the only proposal on the table.^{/1} But at the time *Selfishness, Altruism and Rationality* was published, among "rational choice" social scientists (most conspicuously, economists) the prevailing view was still that the standard account of self-interested choice would ultimately prove adequate to deal with the accumulating evidence against that. Indeed until recent years, and even today for many writers, "rational choice" was by definition self-interested choice.

But by the early 1980's, allowing for motivation beyond simple self-interest had at least moved to the second stage of an often-noticed sequence in science: from "it's too absurd to be taken seriously" to "it's not crazy but it's wrong." The end of the sequence, which may or may not be reached in any particular case, is "we knew it all along." My sense of the situation as this book was written was that broadening a rational choice model to allow for motivation beyond self-interest was well into the third phase of this transition from "it's crazy" to "it's wrong" to "we knew it all along". But the NSNX approach which is the focus here was a step behind the broader notion of extending the notion of rational choice to allow for social ("other-regarding" or "norm-obeying") as well as self-interested motivation.^{/2}

What makes NSNX seem odd is that it violates something ordinarily taken to be so obvious that it seems incoherent to even suppose it *could* be wrong. A formal (mathematical) model of rational choice has always entailed maximizing a utility function. The crucial word in the previous sentence is the shortest: "*a*". For as you will see, NSNX does not do that. On the Darwinian logic that underpins the theory a person would not have *a* utility function. What in a more conventional account would be a utility function with additional arguments to capture other-regarding motivation here *turns out to be* -- for it is not an assumption underlying the account but something that follows from the argument to be introduced in Chapter 1 -- a pair of utility functions.

Each meets the usual conditions of a utility function, but there are *two*: one capturing the self-interested motivation that used to be referred to as "economic man", and the other capturing the person's social preferences.

That a person has distinct social and self-interested preferences is a very old idea, easily traced back not only to Adam Smith's *Theory of Moral Sentiments* but all the way to the Greeks. But the simple equilibrium condition you will encounter shortly, as a matter of purely mathematical considerations, does not permit the two utility functions (self-interested and social) to be merged into some total utility function. As a scientific point, there is nothing at all surprising in that. Science is full of equilibrium conditions where competing forces tend to come into in balance though nothing is being maximized. In the Darwinian argument to come here, it would be rather a miracle if there was some quantity that was being maximized by the tendency towards equilibrium.

What can be said to someone (I have run into such someones many times) who says that if it is actually true that there is no single utility function in NSNX then that shows NSNX has to be wrong? An important part of the answer is that while it is hard to draw firm lessons from the history of science, one thing that does stand out is that it does not work to put prior constraints on what Nature is allowed to do.

But to someone thoroughly immersed in a universe of discourse where rational choice *means* maximizing a utility function, then the NSNX failure to provide this will seem as unreasonable as seeing cars drive on the *wrong* side of the road seems to someone from America or the Continent on first encountering automobile traffic in Britain. So in due course I give that substantively trivial but psychologically potent concern the attention it needs.

The book is organized as follows: The first five chapters are based on my occasional papers on NSNX since the 1982 book, but prior to my current effort to use the extensive accumulation of data from cooperation experiments to test and refine the NSNX account. Since the 1982 book remains in print I do not attempt to cover here anything like the full range of topics discussed there. But everything needed to support the discussion in this book will be covered. What is new here, in particular, is an account of how the individual equilibrium developed in SA&R, balancing self-interest and social motivation, can be elaborated to yield an account of how a social equilibrium emerges

out of the interactions among many individuals. This elaboration of NSNX as explored in SA&R emerges from applying the NSNX logic to extend the range of the ingenious diagram for analyzing simple social choices devised by Tom Schelling. This turns the Schelling diagram into what I call the "S-diagram", which turns out to capture some fundamental dynamics of social change.

But the title divides attention between NSNX and cognition. And the cognitive effects of special interest here are those that channel intuition in ways that violate what on reflection the chooser would judge sensible in the situation. Mostly, cognitive shortcuts take us to quickly to some reasonable approximation of where step-by-step reasoning would get us. But sometimes, of course, the shortcut must go awry. Cognitive effects of that perverse sort plays no role at all in the first five chapters. But if the NSNX proposal captures something essential about interaction between social and self-interested motivation, that must further interact with cognition in the response of individuals to social issues and through them affect the response of societies. And cognitive effects which are sometimes perverse are likely to be of particular concern in social contexts, where they may be harder to correct, and where adverse effects can be multiplied.

NSNX turns on a balancing of propensities to self-interested choice against propensities favoring socially-motivated choice. But social choices will often turn on how a person sees large-scale contexts which are far from transparent to the individuals whose interactions determine the social result. Analysis of social cooperation is then not likely to be adequate without insight from both of what developed as two separate projects but are here brought together. The NSNX project seeks to define how individuals balance between their propensity to act in self-interest and some propensity to act in the interest of a salient group. But, very prominently in the later chapters here, that NSNX account interacts with the *habits of mind* view of cognition I've pursued in other work.

Chapter 6 begins with a bit more discussion than I have given here of why cognitive effects that are usually benign but occasionally adverse can take on enlarged importance in contexts of social choice. But it quickly turns to a particular kind of cognitive shortcut (what I will call *neglect defaulting*) which has not played any role in the very extensive attention to cognitive effects that has developed since Kahneman &

Tversky's work in the 1970's. As with cognitive shortcuts generally, *neglect* defaults usually are effective as well as efficient, but not always. They are usually benign, but sometimes adverse. Then the individual then would have done better to plod through in the time- and attention-consuming step-by-step alternative to the shortcut, but she is unlikely to notice that.

The argument starts from simple puzzles which overwhelmingly elicit responses which are not smart even from subjects who are. Cases of "adverse defaulting" will play a direct role in later discussion of social contexts. But seeing how defaulting can have remarkably strong adverse effects in the context of the simple puzzles (in Chapter 6) prepares for consideration in Chapter 7 of another sort of *defaulting* effect which is intrinsically tied to the social contexts.

The analog of the simple puzzles in Chapter 6 is experimental data in Chapter 7 which defies reasonable interpretation in terms of either self-interested or social motivation or of any plausible mix of the two. Experimental data in which *most* choices are of that puzzling sort turn out to be unexpectedly easy to find. And the account of what is happening suggested by NSNX turns on noticing that *adverse defaulting* akin to the neglect defaults of Chapter 6 must have counterparts in the cognitive structure I call the NSNX cascade. The cascade governs how a social context is seen (as competitive *or* cooperative, with the secondary distinctions you will see in Chapter 7.) Parallel to the neglect defaults encountered in Chapter 6, the NSNX cascade entails a defaulting structure that can be expected to influence social intuitions in contexts that are ambiguous or unfamiliar or otherwise difficult to interpret. But contexts of large-scale social choice, where individuals must make choices about matters far outside their normal range of experience, often have that character.

Chapters 8, 9, and 10 then consider how the cognitive effects introduced in Chapters 6 and 7, interacting with NSNX, can be applied to further experimental data. I try to show how NSNX + cognition effects can yield insight into data from Public Goods and related games that is interesting but not bizarre (Chapter 8), and how it can produce what seem to me even more striking results when applied to experimental data that is also interesting but verging over to choices that sometimes looks quite bizarre (Chapters 9 and 10). But even bizarre choices come from somewhere. The subjects in the experiments

are not ordinarily bizarre in their behavior outside the lab. Somehow in the lab they are being cued into choices that ordinarily would not be made.

The experiments are intended to show us something about what is surely the fundamental problem of the social sciences, which is understanding cooperation among agents whose behavior is ordinarily conspicuously marked by pursuit of self-interest. What is novel here is not the notion that experiments can show us something about how cooperation works. That has been the point all along. But here new tools are applied to the data, yielding what I (of course) hope can be seen as substantial new insights.

This puzzle of how cooperation can be (sometimes) sustained is mostly a puzzle about the extended cooperation of large-scale societies. If we want to understand how the simplest human societies work, we can learn a good deal by treating them as more sophisticated versions of other animal societies, and especially other primate societies. The culture of even the simplest human hunter/gatherers is vastly more complex than the culturally richest non-human community. But anyone who has been to a zoo can see there is no absolute divide between the social behavior of chimps or gorillas and human behavior in small kin groups. Whatever insights can be gained from cooperation among primates, however, does not get us very far at all in trying to understand the vast scale of human cooperation since (at least) the emergence of literacy 5000 years or so ago. Somehow cooperation can be sometimes sustained even when it extends far beyond small groups who know each other and have continuing interactions, and often with strong kinship relations. We can see substantial cooperation extending to people who are strangers but not enemies

In social choice experiments, subjects are recruited to play games, usually for non-trivial amounts of real money, that are arranged to mimic incentives characteristic of important social interactions. The data produced by these experiments therefore provide us with the choices human beings have made faced with carefully organized incentives, conditional on variations in the structure of the games, conditions of the game, player's characteristics, and so on. But interpreting what is going on in the games is a challenge, given the isolation of choice in the artificial environment of the lab from the always far more complicated and never fully specifiable, often even chaotic conditions of choice

outside the laboratory. And almost always, handfuls of players are proxying for the many player interactions common outside the lab.

So it should not be surprising that choice in the lab does not always look very much like choices we see in the world among people who are strangers but not enemies. In contexts outside the range of everyday experience, cues that would ordinarily correct faulty intuitions might be missing or blurred, so that anomalous behavior that pretty easily arises in the lab might be showing us something about real behavior in unfamiliar conditions outside the lab. Large scale political and social cooperation is *mostly* about issues beyond the scale of familiar experience, where how issues are perceived can stray from how things will someday look to historians. In a concluding chapter, I try to give some flavor of how the formal insights of earlier chapters and the insights tied to experimental data in later chapters might be put to work.

That concluding chapter sketches applications of ideas from both the social equilibrium discussion developed through to Chapter 5 and from the NSNX + cognition discussion of Chapters 6-10 to the particularly salient (as this is written, and perhaps long after it is written) of fundamentalist terrorism. I will be considering how the “tipping point” dynamics of the S-diagram can yield insight into the recruitment and longer-term prospects of terrorist movements, and then of how these dynamics could be expected to interact with the defaulting effects developed in the later chapters.

All this is followed by an appendix (on the *template*) which concerns an entirely independent matter. The *template* is software for data analysis developed for use in this project. But this tool for analysis is not at all tied to the theoretical and empirical arguments that comprise the bulk of the book. I consequently defer details until the opening of that discussion, and present it as an Appendix to the main text. This is appropriate, since until you have seen some results from use of the *template*, you are unlikely to be interested in this software exercise anyway.