Wolfgang Spohn

Application for a Koselleck Projekt

Zusammenfassung / Summary:

Die Figur des *homo oeconomicus*, expliziert in der modernen Entscheidungs- und Spieltheorie, prägt weite Teile unserer Sozialwissenschaften. Diese Theorien werden als ein im Kern vollständiges normatives Ideal angesehen und sind daher kritischer Bezugspunkt von Verhaltens-, Psycho- und Neuroökonomie, die die empirischen Unzulänglichkeiten dieser Theorien überwinden wollen. Im Widerspruch zu diesem Mainstream hält das vorliegende Projekt die Entscheidungs- und Spieltheorie für normativ unzulänglich und will daher das normative Bild des *homo oeconomicus* verbessern und damit auch die Ansatzpunkte empirischer Kritik verschieben.

Das Projekt leistet dies, indem es den ‚reflexiven Aufstieg' in aller formalen Strenge konzeptualisiert und theoretisiert. Demzufolge denkt eine Person nicht nur über ihre möglichen Handlungen und deren möglichen Wirkungen nach, sondern auch über ihre möglichen (zukünftigen) Entscheidungssituationen, die diese Handlungen bedingen. Das führt zu einer Systematisierung der dynamischen Entscheidungstheorie, zu einer systematischen Behandlung von Selbstbindungsphänomenen und insbesondere zu einem neuen grundlegenden Gleichgewichtsbegriff in der Spieltheorie, der eine Neubehandlung von Kooperation und in der Tat eine Vereinheitlichung von non-kooperativer und kooperativer Spieltheorie verspricht.

The paradigm of *homo oeconomicus*, as explicated in modern decision and game theory, shapes extensive parts of our social sciences. These theories count as delivering a basically complete normative ideal and hence serve as the critical reference point of behavioral, psycho-, and neuroeconomics, which attempt to overcome the empirical deficiencies of those theories. In contrast to this mainstream, the present project takes game and decision theory to be normatively deficient and thus attempts to improve the normative ideal of a *homo oeconomicus* and to thereby shift the point of attack of empirical criticism.

The project does so by conceptualizing and theorizing 'reflexive ascent' in a formally rigorous way. According to it, a person considers not only her possible actions and their possible consequences, but also her possible (future) decision situations, which entail those actions. This will provide a systematization of so-called dynamic choice, a systematic treatment of (pre-)commitment, as widely discussed in the literature, and in particular a new fundamental equilibrium concept for game theory, which promises a novel treatment of cooperation and indeed a unification of non-cooperative and cooperative game theory.

Sketch of the Project

(1) Introduction

This project is very risky, since it opposes a widespread mainstream in economics and philosophy. If I am right, it is utterly ground-breaking, since it intends to elaborate reforms of decision and game theory at their very bases; this is bound to have utterly rich consequences. And I need support, since this basically philosophical, but in effect interdisciplinary project goes more deeply into economics and mathematics than I can afford by myself; that's why I am applying for a Koselleck project. These are grand announcements. Too grand? What are they about?

Standard decision theory, as paradigmatically developed by Savage (1954), and standard game theory, as paradigmatically presented early by Luce, Raiffa (1957) and much more comprehensively, e.g., by Myerson (1991), provide the theoretical foundations of microeconomics—and of macroeconomics, too, insofar methodological individualism holds. These theories explicate the tremendously influential model of *homo oeconomicus.* On the one hand, this model states a normative conception of rationality—we *ought* to conform to it. On the other hand, the model provides an idealized empirical picture—we actually conform to it, approximately—and thus serves as foundation of economic theorizing. This normative-empirical double role is peculiar to all theories of rationality (Spohn 1993).

In its empirical role, this model has increasingly been criticized for decades. Economists have not excelled in prediction. Their theories do not seem well made for this. The basic model seems to contain so formidable idealizations that it can't but mislead empirically. It's simply not true that we conform to it at least approximately. This has led to the rise of behavioral, psycho-, and neuroeconomics. Three Nobel prizes underscore the importance of these developments (Selten 1994, Kahneman 2002, Thaler 2017, though Selten did not receive it for his groundbreaking work in behavioral economics). Those disciplines attempt to modify the model of *homo oeconomicus* and thus to arrive at empirically sounder theories.

My essentially philosophical project is to pursue a different line of criticism, which attacks standard decision and game theory as normative theories of rationality. As such they seem basically perfected and thus serve as steadfast reference point for empirical criticism. However, I take them to be substantially *deficient from a normative point of view*. This project attempts to considerably amend and generalize our normative model of *homo oeconomicus*.

Clearly, if correct, the project is highly significant for normative theorizing. However, it would also be of great importance for empirical research, because it would shift the normative reference point of that research. If the ideal to be empirically corrected shifts, the corrections shift as well or may even be superfluous. In this way, the reference frame of behavioral economics will be deeply concerned.

To mention just a teaser, which will be slightly expanded below: The ultimatum game has always been considered as a strong case in favor of behavioral economics, because its only Nash equilibrium is extremely unfair and actually rarely chosen. We usually find a fair, or a fairer, division of money. A basic point of my project will be to develop and defend the alternative and more general notion of what I call a dependency equilibrium. And it will turn out that both, a fair and a merely fairer division, as well as the apparently irrational rejection of an insufficient offer are dependency equilibria in the ultimatum game. This is to serve as one example of how the entire dialectical situation may change through this project.

(2) The Basic Idea

The basic ideas of the project are few and simple and, I think, cogent. Let me start with decision and turn to game theory only at the end. I wish I could refer here to some formal representation of decision theory, but I have to do my best in remaining informal:

A familiar distinction is that between chance events (an unhappy term) and actions, which are also events. Some (chance) events have happened, the world is in a certain state, you do something, then some further events happen, you do something else, etc. In such a way an entire course of events unfolds. Many such courses may evolve. They please you or conform to your desires to varying degrees. That is, you have a *utility function* over such courses. Now you can influence the courses through your actions. Hence, you have varying *subjective probabilities* for the possible courses of events given your possible courses of actions. The basic and convincingly defended normative rule then is to choose some course of action which *maximizes conditional expected utility* (*CEU*). Let me call this mental set-up, consisting of a

conceptual structure of possible actions and events and of a cognitive/epistemic attitude (= probabilities) and a conative/optative attitude (= utilities) towards them, a *decision situation*. To emphasize, this is something internal; being in a certain decision situation means having it in mind.

I am well aware that there are many discussions about the format of those cognitive and conative attitudes; indeed I have contributed to them (Spohn 2012b, 2017). The project will conservatively stick to probabilities and utilities, because its concerns lie elsewhere. Another delicate issue is that economists tend to assume utility functions to be selfish or egoistic and then discover that utilities so interpreted need to be amended by social or other-regarding preferences. However, this assumption is nowhere written into decision and game theory. Hence, I will interpret utility functions throughout as these theories do, namely as representing the overall action-guiding preferences, however they may be analyzed into various components.

Now, of course, you don't fix a given course of action in advance, although you may do so. What you usually do is to develop a *strategy* for actively responding to various courses of events. And what you should do then is to maximize the CEU of possible strategies.

The **first key point** of my project is that this standard notion of a strategy is too narrow. According to it, a strategy responds to various external events. However, you can respond to those events only if you learn about them. That is, what you really respond to is your knowledge of those events, which changes your mental set-up. This immediately suggests to generalize the notion of a strategy as something responding to arbitrarily changing decision situations, at which you may arrive not only through (uncertain) learning, but also through forgetting and other epistemic changes, and also through changing your (intrinsic) desires or utilities; there are many conceivable causes of such changes. *This* is what you have to consider strategically.

The only way to do so is to extend the conceptual structure of the decision situation at hand by all those possible decision situations that you might reach and in which you decide about the subsequent actions. In other words, the conceptual structure now contains not only event or chance nodes and action nodes, but also decision nodes representing those possible future decision situations. (Beware: what I just called action nodes are standardly called decision nodes, and what I just called decision nodes doesn't exist in the standard theory. One reason for the latter might be precisely the conflation of actions and decisions.)

It is this extended structure and hence this extended kind of decision situation that I want to study in my project. It should be clear why I call this *reflexive decision theory*: precisely because I assume the agent to have higher-order beliefs reflecting possible future first-order attitudes which in turn rationally determine future actions. To the best of my knowledge, this reflexive structure has never been considered in full generality. I could give various explanations why this is so. In any case, the point that we have to consider such reflexive structures seems entirely inescapable to me, once we realize what our strategies really respond to.

(3) Reflexive Decision Theory

The central problem now is to find a decision rule for those reflexive structures. In principle, this problem has been quite extensively discussed, after the ground-breaking paper of Strotz (1955/56), under headings like dynamic choice or endogenous preference change, with no real consensus, as far as I see, and never within my general setting. The central difficulty here is that optimization is now governed by different points of view (= mental set-ups = decision situations), your present one and your possible future ones, which may diverge conatively and cognitively. That's why theorists, at this point, always felt to transcend the confines of standard decision theory, which is about optimizing only one point of view. One finds various approaches here; let me only mention the widely accepted rule of so-called sophisticated choice or

McClennen's (1990) minority rule of resolute choice, etc. Let me only say that it does not seem unfair to say that according to all approaches the agent disintegrates into separate stages which somehow seek their own interest. I find this very strange. *Persons do not disintegrate into stages*, they somehow try to integrate the various points of view they possibly have into one. To do so is an important aspect of rationality.

So, the **second key point** of my project consists in a proposal for generating that integration by a basically subjective second-order evaluation of the various points of view that we possess. For instance, learning usually moves us to a better or superior point of view, while forgetting moves us to an inferior one. Addiction and brain-washing move us to an inferior point of view, while education usually moves us to a superior one. Aging changes our point of view, sometimes to the better, sometimes to the worse, and sometimes neither. Or so most of us would say; as said, I don't claim objective standards. I think I know by now how to formally explicate the role of this second-order evaluation in a precise and general way.

One may suspect that these reflexive decision models contain a terribly complex recursive structure, that a decision rule for such structures referring to second-order evaluations is even more complicated, and that all of this results in a totally unrealistic exaggerated rationalization. Yes, I admit this. Still, it is, I think, a great achievement to have the general structure and the general rule, not in order to develop the general complex theory—I presently don't really see the point in doing so, even if my work prepares for this—but in order to see how it covers a great variety of very diverse phenomena and their theoretical treatments:

One issue will be to systematize the many contributions to the issue of endogenous preference change, including a treatment of the neglected phenomenon of positive addiction (more amiably described as becoming a connoisseur) and a resolution of the dispute between sophisticated and resolute choice. Another issue is the so-called preference for flexibility (Kreps 1979), the desire to leave room for tomorrow's preferences. This has implications for portfolio theory. Clearly, the theory I am trying to develop has a close relation to the theory of time preferences, which in turn is basic for the theory of interest; I am still unsure how to conceive of that relation. The economics of education might be quite a different application (see, e.g., Gintis 1974). There are also applications too obvious to even think of their rational foundation such as wearing glasses (in order to acquire sharper information) or the evolution of script (in order to fight forgetting). My hope—according to my preliminary work a very plausible hope—is that all these variegated topics may be subsumed under, and thus be systematized by, my general framework. Clearly, I need economic expertise to carry out all these details and thus to prove the usefulness of the general framework. This topic makes for a very substantial dissertation and/or several substantial papers. This work will be co-supervised by Prof. Dr. Urs Fischbacher from the Economics Department of the University of Konstanz.

(4) Unattended Causal Relations

I didn't mention so far that the agent's mental set-up, the decision situation, also contains a causal picture, which is usually implicit in the subjective probabilities. Of course, it must do so; the agent must think to be able to causally influence the events by her actions. As far as standard decision theory is concerned, this point is well accounted for in causal decision theory. However, the novel reflexive decision nodes must also find a place in that causal picture.

To begin with, these novel decision nodes are subject to variegated causal influences; our mental states are complex things with complex causes. We may even influence them by our own actions (e.g., prevent forgetting or guard ourselves against seduction). Another obvious point is that decision situations cause the actions that are decided or intended within them; that's the widely accepted causal theory of action (Davidson 1963). Both points are already taken care of in reflexive decision theory as sketched in the previous part (3).

The **third key point** of my project now consists in the observation that those mental set-ups, decision situations, have also effects in the world not mediated by the actions they cause. This possibility has been paradigmatically instantiated in the Toxin puzzle (Kavka 1983). However, it is not just a fancy case. Granted, the natural world is receptive to our mental states almost exclusively through the actions caused by them; we *act* upon the natural world. But the social world is very different. Our fellows know about our mental states not only through the ensuing actions. We express them linguistically, and also through gestures and facial expressions. Our conative-cognitive set-up is accompanied by emotions, which we cannot hide. And so on. Frank (1988) has deeply pondered about the economic relevance of emotions.

Again, the question is how to accommodate such side effects of possible decision situations. We have to reckon with them, this must have repercussions for our mental set-ups, and thus it must have significance for rational action. And again, I think I can state the fundamental decision rule adequately dealing with those tricky causal situations (Spohn 2012a).

This point will turn out to have great powers of systematization. For instance, it seems to allow a systematic treatment of (*pre*-)*commitment*, which precisely consists in letting the others see that one is determined early on. Since Schelling (1960), economists know about the relevance of this point, but it is not unfair to say that its theoretical treatment has remained a delicate affair. Again, resolute choice must be discussed under this heading. Another important aspect of this point is explained in Kusser, Spohn (1992), where it is argued that our own pleasures and pains, or hedonic states, are not only caused by our actions and external events, but are also direct effects of our mental set-ups (= decision situations) not mediated by our actions. This causal structure prevents the derivation of extrinsic utilities from the intrinsic utilities of hedonic states and thus prohibits simple maximization of CEU. This point explicates deep concerns regarding the maximization of utility already voiced by Butler (1729), and it helps clarifying long-standing confusions about the notion of utility (which has been conceived in hedonic, motivational, and other ways). Here the project will develop considerable extra potential.

So, the third point is full of deep philosophical consequences, which I have not argued here, but thought through to some extent. They clearly need precise and detailed elaboration, something I would like to achieve by myself within this project.

(5) Consequences for Game Theory

Perhaps the most dramatic consequences of the previous point concern game theory. There, non-cooperative game theory has always been taken as fundamental. It is basically characterized by the causal independence of the actions (strategies) of the players. This causal independence, in turn, is encoded in the notion of a Nash equilibrium and its assumption of the probabilistic independence of the actions of the players. That notion (and its modifications) is the base of non-cooperative game theory.

However, this notion rests on a fallacy. (I am fully aware of the heretical nature of this claim.) Causal independence does not entail probabilistic independence. Two causally independent actions may nevertheless have a common cause and thus be probabilistically dependent. This may sound odd, but there can be no doubt that this is a pervasive social phenomenon. The individual actions are caused by the individual decision situations, but the individual decision situations (= mental set-ups) of the players are almost always causally entangled, and that causal entanglement is then a common cause of the players' actions.

The point is that this phenomenon is invisible as long as one does not take the reflexive step to be explored here. And this step has never been properly taken in 70 years of game theory, not even in epistemic game theory (of which Spohn (1982) is the first clear precursor). If one takes it, one must allow probabilistic dependence between the actions of the players

(which, to repeat, does not mean causal dependence). Maximization of CEU and the familiar common knowledge or publicity assumptions of game theory then entail a new type of equilibrium, which I call dependency equilibria (Spohn 2003, 2010). This is the **fourth key point** of my project.

Each Nash equilibrium is a (limiting case of a) dependency equilibrium (so that all of Nash equilibrium theory is maintained as a special case), but not vice versa. The new type differs from correlated equilibria, as is displayed by the fact that cooperation is a dependency equilibrium in the single-shot prisoners' dilemma (PD), but not a correlated equilibrium. Also, my claims about the ultimatum game in part (1) turn out provable. Of course, players always have the freedom to make themselves independent of the other players; then we are back at Nash equilibria. However, *players also have the freedom to stick to a mutually advantageous dependency*. I take this possibility to be fundamental for all human affairs. For conceptualizing it we need the notion of a dependency equilibrium.

Obviously, it is a huge task to rebuild game theory on these new foundations. So far, I have done so only most rudimentarily. But the idea can only convince if it is elaborated in deep and precise detail. Clearly, one may copy a lot of existing theory, but there will also be a lot of new issues. Also, one need not fear that game theory will be completely overthrown. Far from it. Standard game theory has developed many interesting attempts to account for cooperation (Myerson 1991, e.g., is full of them). The most popular attempt by now at explaining cooperation is to assume social or other-regarding preferences; this means, though, to assume that utility functions differ from what they were supposed to be in the critical examples (like PD). It will be interesting to compare all this with the approach taken here, whether we find tension or support.

All in all, it does not seem unfair to say that cooperation has always appeared difficult or problematic for game theory. This is strange, since cooperation seems to be an entirely natural phenomenon. Dependency equilibria promise to do better in capturing this phenomenon head-on. They even have the potential to unify non-cooperative and cooperative game theory. Developing game theory on the new basis in comparison with existing theorizing is potentially endless work. And it will be difficult mathematical work for which, again, I will need expert help. This makes for another very substantial dissertation or several very substantial papers. This work will be co-supervised by Prof. Dr. Markus Schweighöfer from the Mathematics Department of the University of Konstanz. The goal is to carry the development at least so far that it can no longer be dismissed as a viable alternative conceptualization of game theory.

(6) End Statement

I am well aware that the previous pages contain a lot of immodest claims. They may seem to come from an outsider, to address familiar problems that are already extensively treated and to propose new solutions that may appear to be based on misunderstandings. I cannot argue here that this is not so. I can only refer to my long-standing occupation with the topics treated. Spohn (1978) is the first German dissertation in philosophy about modern decision theory. The thesis also contains the foundations of causal Bayes nets, which have developed considerably later into the most widely used account of causation. I have continued working on causation ever since (Spohn 2012b), an important requisite for the present project. The dissertation moreover contains the first glimpses to the reflexive step taken in part (2), which I have further improved, but not published. As mentioned, I have made important contributions to game theory, and I have worked at the other elements mentioned in my proposal. I have, so to speak, all pieces of the mosaic. However, I have worked on many other topics in the meantime, where I have proved my ability of rigorous constructive theorizing without losing contact to intuition and to the phenomena. That's why the present project hardly progressed.

So, I have a lot of material, I have even written preliminary versions of two chapters (200 manuscript pages) of the big research monograph, which is to result from this project. Now I will find time to fully devote myself to this topic. There is tremendous work awaiting me for which I need the five years support provided by a Koselleck project.

Bibliography:

Butler, J. (1729), "Fifteen Sermons Preached at the Rolls Chapel", in W.R. Matthews (ed.), *Butler's Sermons and Dissertation on Virtue*, London 1949.

Davidson, D. (1963), "Actions, Reasons, and Causes", *Journal of Philosophy* 60, 685-700.

Frank, R.H. (1988), *Passions Within Reason. The Strategic Role of the Emotions*, New York: W.W. Norton & Co.

Gintis, H. (1974), “Welfare Criteria with Endogenous Preferences: The Economics of Education”, *International Economic Review* 15, 415-430.

Kavka, G.S. (1983), "The Toxin Puzzle", *Analysis* 43, 33-36.

Kreps, D.M. (1979), “A Representation Theorem for ‚Preference for Flexibility’”, *Econometrica* 47, 565-577.

Kusser, A., W. Spohn (1992), "The Utility of Pleasure is a Pain for Decision Theory", *Journal of Philosophy* 89, 10-29.

Luce, R.D., H. Raiffa (1957), *Games and Decisions*, New York: Wiley.

McClennen, E.F. (1990), *Rationality and Dynamic Choice*, Cambridge: Cambridge University Press.

Myerson, R.B. (1991), *Game Theory. Analysis of Conflict*, Harvard University Press, Cambridge, Mass.

Savage, L.J. (1954), *The Foundations of Statistics*, New York: Wiley.

Schelling, T.C. (1960), *The Strategy of Conflict*, Oxford: Oxford University Press.

Spohn, W. (1978), *Grundlagen der Entscheidungstheorie*, Scriptor, Kronberg/Ts. 1978, out of print; pdf: http://www.uni-konstanz.de/FuF/Philo/Philosophie/Spohn/spohn_files/GE.Buch.gesamt.pdf

Spohn, W. (1982), “How to Make Sense of Game Theory”, in: W. Stegmüller, W. Balzer, W. Spohn (eds.), *Philosophy of Economics*, Berlin: Springer, pp. 239-270; wieder abgedruckt in: Y. Varoufakis, A. Housego (eds.), *Game Theory: Critical Concepts, Vol. 4, Discontents*, London: Routledge, 2001, pp. 213-241.

Spohn, W. (1993), “Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein?“, in: L. Eckensberger, U. Gähde (eds.), *Ethik und Empirie. Zum Zusammenspiel von begrifflicher Analyse und erfahrungswissenschaftlicher Forschung in der Ethik*, Frankfurt a.M.: Suhrkamp, pp. 151-196.

Spohn, W. (2003), “Dependency Equilibria and the Causal Structure of Decision and Game Situations”, *Homo Oeconomicus* 20, 195-255.

Spohn, W. (2010), “From Nash to Dependency Equilibria”, in: G. Bonnano, B. Loewe, W. van der Hoek (eds.), *Logic and the Foundations of Game and Decision Theory – LOFT 2008*, Texts in Logic and Games, Springer, Dordrecht, 2010, pp. 135-150.

Spohn, W. (2012a), “Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box”, *Synthese* 187, 95-122.

Spohn, W. (2012b), *The Laws of Belief. Ranking Theory and Its Philosophical Applications*, Oxford: Oxford University Press.

Spohn, W. (2017), “Knightian Uncertainty Meets Ranking Theory”, *Homo Oeconomicus* 34, 293-311.

Strotz, R.H. (1955/56), "Myopia and Inconsistency in Dynamic Utility Maximization", *Review of Economic Studies* 23, 165-180.