Philosophy and Economics I: Methods and Models

by

Hartmut Kliemt

Preamble and overview

Preamble

These lectures in philosophy and economics intend to bring together several strands of inquiry on the level of an advanced undergraduate or an introductory graduate course for students of PPE (Philosophy, Politics, and Economics), MPE (Management, Philosophy, and Economics), P&E (Philosophy and Economics), and L&E (Law and Economics). Hopefully, others interested in (practical) philosophy, political theory, or (normative) economics will find the discussion useful as well. The aim is to give an integrated presentation of philosophical and economic topics that is accessible with a rather modest training in any of the aforementioned academic fields. For the more advanced and the specialists, the chapters of this two volume book can serve as an invitation to "look at the world through an interdisciplinary window." The first volume deals with methods – in the wide sense including meta-theoretical issues – and models – in the decision theoretic sense. The second volume will again start with some reflections of a methodological kind and then discuss morals on the basis of the methods and models of the first. The ambition is to illustrate that "economic philosophy" (or "philosophical economics") does indeed exist as a discipline in its own right.

This project began at the European Forum in Alpbach (Tirol) in 1984. On this occasion I recommended to Thomas von Cornides of Oldenbourg that this publishing house might translate Axelrod's then upcoming book on the evolution of co-operation. Cornides immediately went on to get the rights to the book but also suggested that I write an introduction to the field of philosophy and economics in German. He wanted it for his series "scientia nova." Based on the lectures that I gave together with James M. Buchanan at Alpbach at the time and using the material of lectures presented in Munich where I served as a substitute lecturer for Wolfgang Stegmüller shortly after, I did follow Cornides' advice to "write it all up." The first text was then used for courses offered in the philosophy department of the University of Frankfurt in 1987. Later on, after moving to the University of Duisburg, Bernd Lahno and I expanded the German text. We use the text still as lecture notes distributed for free to our students.

Given this history, it seemed to me appropriate not to give in to the sirens of an English publishing house and to stay with the original proposal of Oldenbourg as publisher of this philosophically somewhat more and technically somewhat less advanced text. Much of it is the result of lectures and courses held at the Universidad Torcuato Di Tella, Buenos Aires, in 2003. The work on these lectures started after Guido Pincione and Horacio Spector invited me there. Subsequently lectures of the same kind and beyond were held at the Center for Study of Public Choice, George Mason University, Fairfax in 2003, 2004, and 2005 and the Vienna Circle Institute, Vienna, in 2006 (together with Geoffrey Brennan). I am greatly indebted to the hospitality and inspiration that the three institutions and their staff offered. Some parts are drawn from my lectures in the MBA program of the Frankfurt School of Finance and Management, too. Bernd Lahno and I have recently transferred there from “the calm but dark fields of philosophy” to set up a Management, Philosophy, and Economics program. I am grateful to the school for providing an excellent environment in which this work could be completed.

On a more personal level I should like to mention in particular my friends Hans Albert, Max Albert, Michael Baurmann, Geoffrey Brennan, Jim Buchanan, Bernd Lahno, and David Levy. As can be seen from the lectures, these excellent philosopher economists influenced my thinking fundamentally. I do owe much to another seasoned economic philosopher and old friend, Manfred Holler, for kicking me to get this book done. His generous advice as well as his written and oral criticisms greatly improved the text. Without Manfred’s persistent interest, I never would have finished the project.

That much of the following is most strongly influenced by my collaboration with Werner Güth should be obvious to those who know his publications. Werner took me by his “invisible hand” almost to the extent of co-authorship and many of the thoughts expressed here I owe to him. Our joint work on foundational matters started after Reinhard Selten brought us together in his “game theory in the behavioral sciences” research project at the Center for Interdisciplinary Research of the University of Bielefeld (ZiF) in 1987/1988. Since the early 1990’s, Werner has become my game theoretic mentor and closest friend. I do owe much to the ZiF, not only in this regard. It is also the institution where Joachim Frohn, Werner Güth, Reinhard Selten, and I organized a one year project on “making choices” in 1999/2000, which had a considerable effect on some of the following chapters and on me personally, by bringing my colleague Marlies Ahlert into my life.

Last but not least I owe a great debt to my old American friend Allison Blizzard who – nomen est omen – went through the text “like a storm”. She did not merely improve the English but also the clarity and intelligibility of the text. Her criticism has been invaluable. Needless to say that the remaining blunders are all mine.

Overview

Systematically, the following text brings together elements from three broad subcategories. *The first part* on method starts with a paradigm of a basic rational choice analysis of social interaction (chapter 1). It goes on with a methodological reflection on how we "make the world" from an objective and a participant's point of view, respectively (chapter 2).

The second part, devoted to the models, takes up the tools from what has come to be known as the "rational choice approach to human (inter-)action." It relies on many sources, but the influence of Werner Güth and, more indirectly, the foundational work of Reinhard Selten will be felt throughout. Elementary models of action (chapter 3) form the beginning of part II. Then, likewise, elementary models of inter-action are introduced (chapter 4). Both chapters can be read either as a rehearsal of or as an introduction to elementary models of opportunistically rational choice making. They put the models into a distinctive philosophical perspective by asking what their basic – often tacit – assumptions are and what they "mean". To that effect, it is shown (chapter 5) how commitments can be modeled explicitly and legitimately as part of the rules of the game. The next chapter (6) illustrates how in principle "internalist" and "externalist" accounts of interaction can be brought together in an "indirect evolutionary approach" as proposed by Werner Güth. The final chapter of the first volume (chapter 7) brings the discussion of modeling techniques and their interpretation to a preliminary close by relating the two basic perspectives that run thorough this book to each other and to issues discussed in the second volume.

The third part, which is presented in the second volume, will deal with, very broadly speaking, moral and welfare economic theories. The reader might look at the chapters forming the parts of both volumes as an invitation to join the search for a "reflective equilibrium" on foundational issues of economic philosophy. The search method is interdisciplinary. It makes an effort to integrate many strands of modern philosophy and economics via decision theoretic language.

Its decision theoretic background notwithstanding, the discussion is nontechnical. However, the present text is in many ways more technical (in a nonmathematical, philosophical sense) than, say, a good first introductory text to game theory like Dixit and Nalebuf (Dixit and Nalebuff (1991)). The text will not be easily accessible to those who have no training in rational choice at all. Gerald Gaus (see Gaus (2008)) might serve as a complementary introduction to the present text for those who have no former acquaintance with decision theoretic arguments. (Binmore (1994, (1998, (2005)) as well as Skyrms (1996) and Sugden (1986) are highly recommended and so are Taylor (1976; Taylor (1987), Young (1998) to mention just a few great texts from a flourishing field. Some more specific aspects are dealt with in several extended discussions (some of my favorites include Baurmann (2002), Brennan and Lomasky (1993), Brennan and Hamlin (2000), Brennan and Pettit (2006), Buchanan (1999), Smith (2008)). My aspiration here is to present a coherent narrative covering core issues of the field from a point of view that is sympathetic with rational choice modeling (RCM) but aware of the limits of rational choice theory (RCT).

Reading the book

1. **Readers with some background in elementary micro-economics and/or elementary game theory** may want to start with chapter 2 and read it in full. They then should eyeball chapters 3 and 4 and focus there on the distinction between strategies as plans that are made and strategies as moves that can be chosen. It will become obvious that a concept like sub-game perfectness is not a technical device to sort out solutions of games but rather expresses a fundamental philosophical insight. I will refer to this insight as the "principle of intervention." It characterizes the rational forward-looking choices of choice makers who are able to distinguish between what is and what is not a causal effect of their "interventions." Then, chapters 5, 6, and 7 should provide both a smooth read and a hopefully interesting argument for such a reader.

2. **Readers with no background in elementary micro-economics and/or elementary decision and game theory** should start with chapter 1, move on to chapters 3 and 4, then go back to chapter 2 and finally continue with chapters and 5, 6, and 7.

The second volume will also be a directly accessible text on normative issues that will be presented both from an ethical and a welfare economics point of view. As in the first volume, all chapters of the second volume are reasonably self-contained such that they can be used separately within courses on philosophy and economics or as lectures on several topics of interest for those who are working in one of the fields and are interested in looking at matters from an interdisciplinary "rational choice" angle.

Inhalt

Part 1: Looking at the World Through Different Windows

The simple Robinson Crusoe and Friday world laid out subsequently shows why we humans, in a very fundamental way, have an interest in co-operating and, at the same time, have so many problems accomplishing what is in our interest.[1] Co-operation is practically always "antagonistic co-operation."[2] It is rational for us to wish that we co-operate. "In foro interno" we have good reason to desire that we do so.[3] At the same time, "in foro externo" it may be rational not to act according to our wishes. As will be shown, *rationality itself rather than any deficiency of it stands in the way of our getting what we have a rational reason to wish be true.*

In Part I, after the introductory "appetizer" in chapter 1, the first main course follows in chapter 2. The chapter addresses two fundamentally different "ways of world making" that should be better distinguished in philosophy and in economics than they normally are. The two perspectives are associated with the traditional distinction of understanding and explaining. It is sometimes believed that a gulf parts the two, but there is none. The best way to understand things is to explain

[1] A more economic introduction to the same kind of argument is presented by James Buchanan in his "Property as a Guarantor of Liberty," in Buchanan (1999 ff.), volume 18. I highly recommend this strikingly elegant and insightful paper, which has been undeservedly neglected by many, including myself. As the editor of the volume of Buchanan's collected works, in which it is reprinted, however, I do not have as good an excuse as others.

[2] A term used already in Sumner and Keller (1927)

[3] See Hobbes who uses the category of *foro interno* in a sense that alludes to strategic interests or their absence in his Leviathan. Interpretations focusing on the conscience of the actor miss the point entirely by leaving out the strategic aspects of the problem; see chapter 13 in Hobbes (1651/1968)

them. Nevertheless, there is still a fundamental difference between alternative modes of explanation in social theory.[4] They either explain matters in terms of future expectations or in terms of past causes. The acting individuals are modeled either as led by their aims, ends, values, and expectations of the future or by their experiences in and of the past. How the two perspectives – the teleological aim-oriented and the behavioral law-based one – influence actual modeling as well as substantive moral views will be addressed in the parts devoted to models and to morals, respectively. – Now, however, first things first, let us begin with the "appetizer" for the non-economists, which is followed by the first "main course" in which the economist might want to join as well.

[4] Basic elements of a critically rational world view are presented succinctly and clearly in Albert (1985)

1 The hidden side of the "invisible hand"

Robinson Crusoe has landed on his island. Though after a while he finds out that some man-eaters occasionally visit this spot of land, he is basically on his own. He can do whatever seems fit to him unhampered by any "artificial" (i.e. socially produced) restrictions. Crusoe plays a so-called "game against nature." Daniel Defoe vividly describes that game and how Crusoe allocates his time to diverse tasks, including leisure activities.

The story is interesting in itself, but let us try to reduce it to its mere theoretical bones: As a rational individual, Crusoe will allocate his time such that the satisfaction he derives from what he does will equalize across all his occupations "at the margin" or concerning the last unit of effort. For, if he could derive more satisfaction from devoting his last unit of effort to occupation *a* rather than to occupation *b*, then he should re-allocate at least the last unit invested in *b* to occupation *a*. He would derive the higher satisfaction of *a* from the same effort as he would need to invest in *b*.

For the sake of simplicity, let us assume further that Crusoe basically has three options. He can produce steaks, beer, or invest in defense activities. As long as Crusoe is on his own, we may neglect defense. Though in Defoe's original Hobbesian plot Crusoe is afraid of an outside invasion all the time and spends some resources preparing for that eventuality, let us assume, initially here, that Crusoe is only interested in producing either beer or steaks. He has a fixed time budget for productive activities. (Of course, strictly speaking, he would seek to allocate his time such that at the margin spending it on leisure would be as satisfactory as to spend a marginal unit on creating the production cum consumption bundle – but we will let leisure also be forgotten for the time being.)

Crusoe, as long as he is on his own, will devote his time exclusively to the two alternative productive activities of "making" steaks and beer. More specifically, let us start from the premise that Crusoe reaches maximum satisfaction within his own production possibility space if he produces 5 units of steak and 5 units of beer in each time period. Should he specialize on one of the products, he could, due to "economies of scale,"[5] produce up to 20 units of each of

[5] He can, say, by devoting one unit of time get one unit of output, but using two units of his time he can get, say, three units of output. The increase is more than proportional to the additional time

the products while having none of the other. But, subjectively, he prefers a consumption of (5, 5) over a consumption of (20, 0), a consumption of (0, 20) or any of the other consumption bundles (or vectors) that he could realize on his own – whichever these might be.

Crusoe's preferred feasible consumption bundle:

(5 units of beer, 5 units of steak)

Crusoe's (extreme) production possibilities:

(20 units of beer, 0 units of steak) or (0 units of beer, 20 units of steak)

Crusoe would prefer to have more than five units of each of the products if only he could overcome the natural constraints of his production possibilities.[6] However, for Crusoe on his own there is no way to accomplish this, for he has fixed technological means. Consuming (5, 5), Crusoe would have to give up more of either beer or steaks than would be worth it to him in terms of the other good. The opportunity costs – i.e. what he has to forego in terms of the alternative – will deter Crusoe from any re-allocation of his efforts at his optimal consumption cum production position (5, 5).[7]

After a while the rules of Crusoe's game become modified. Some man-eaters on a picnic trip arrive by boat. Friday is on their menu. Crusoe, following a sudden inclination, rescues Friday. Now there are two on the island. In Daniel Defoe's story, it is assumed that Friday submits to Crusoe immediately. However, in this new game between two human individuals who "play" each other rather than play against nature, Friday cannot be sure whether Crusoe wants to spare his life or whether he is out for some change in his own diet too.

Lacking any specific information about Crusoe's preferences, Friday has to assume that Crusoe may have a hidden agenda when intervening. A rational Friday should infer that a rational Crusoe would not risk his own life to save the life of a person completely alien to him. Within the constraints of human nature, as we all experience it quite independently of our cultural backgrounds, such acts of unselfishness as we observe between people close to each other are unlikely among strangers. Therefore, Friday should think "cross-culturally" that unless Crusoe would stand to gain substantially from his act, he would be unwilling to

spent. That may happen for many different reasons. For example through practicing an activity he becomes better at it, concentration on one task rather than several reduces transition costs etc. Adam Smith and his example of the pin factory in which workers become much more productive due to the division of labor does not apply yet because Crusoe is still alone on the island.

[6] It is assumed throughout that he is not satiated with respect to any of the dimensions of value.

[7] A comprehensive philosophical treatment of "opportunity cost" is Buchanan's "Cost and Choice", volume 6 of Buchanan (1999 ff.)

incur substantial risks. Thus, Friday must come to the conclusion that a hidden agenda is involved.[8]

A rational Crusoe, in view of his ignorance of Friday's motive and character, would also have good reason to be afraid of Friday as a threat to his own life, which is indeed stressed by Defoe in his plot. From this point of view, only a dead Friday can be a good Friday for Crusoe. Again, a rational Friday, who understands all this, should conclude that Crusoe does what he in fact does out of some ill will to him. Friday, therefore, should rationally distrust Crusoe and – contrary to what Defoe and Hobbes might tell us – should not, as the story would have it, put Crusoe's foot on his neck. Moreover, even if Friday believes in Crusoe's good intentions, this gesture would be pointless. For, how could he surrender without any institution rendering the act of surrendering trustworthy? Putting Crusoe's foot on his neck as Defoe suggests might not be an unambiguous gesture. And, above all, if performance of the gesture is not an institutionally defined act, then after the gesture is performed, the state of nature still prevails; there are no causal effects that are institutionally transmitted into the future. If, on the other hand, the gesture is institutionally defined in advance, then an institution must exist and, thus, the state of nature cannot exist as far as this institution is concerned.

In view of the preceding, let us assume that Friday does not even try to signal surrender. He rather escapes to that part of the island that is not controlled by Crusoe. Now two individuals are permanently on the island. As we shall presume, they divide the island in two, and on *his* part, Friday starts to produce steaks and beer too. From the productive side as well as from the point of view of his tastes, it turns out – at least for the sake of our theoretical story – that Friday is Crusoe's identical twin as far as their options are concerned and in that they are mutually aware of their situation.

Friday's preferred feasible consumption bundle:

(5 units of beer, 5 units of steak)

Friday's extreme production possibilities:

(20 units of beer, 0 units of steak) or (0 units of beer, 20 units of steak)

Both Crusoe and Friday are basically in the same position. If they were alone on two separate islands without any chance of exerting causal effects (externalities) on each other, they would each produce and consume (5, 5). However, they are not on their own and not on separate islands;[9] they are on the same island. Knowing this, they could spend their time on other activities too, for they could acquire more of the goods that they desire by simply stealing them from the other

[8] Spontaneous acts may be special, see in particular Frank (1988)

[9] John Donne, Meditation 17 "*No man is an island, entire of itself...any man's death diminishes me, because I am involved in mankind....*"

producer. As rational individuals, they will devote some of their time to this "noble" activity until its marginal returns are equal to the returns of their other activities. Expecting this, both will invest in some defense activities, diminishing the returns of the stealing activities of each other until their time invested in each of the activities is expected to be equally advantageous.

After complete adaptation of their allocation decisions to the new circumstances of interaction, a "natural equilibrium" of the Buchanan-Bush type will emerge.[10] In an **equilibrium**, neither of the two actors can do better as long as the other does what he in fact does and is expected to do. If an equilibrium of an interaction like the preceding emerges, then nothing can be gained by further unilateral behavioral adjustments as long as the behavior of the other actor remains fixed. For instance, the unit of time devoted to one's own productive activities brings in at least as much as the unit of time in preventing stealing given the time allocation of the other actor. Likewise, an additional investment of time into the activity of taking from the other actor will yield less satisfaction than devoting that unit of time to "directly" productive activities in one's own territory (factoring in the leak that exists due to the takings by the other).

In the equilibrium state, none of the actors can reallocate his time and, thereby, do better as long *as the other actor persists in his equilibrium allocation.* Like the stones in a Roman arch, the activities and expectations hold each other in place. If Crusoe expects Friday to act in equilibrium, then this expectation will not prevent him from going on with his own equilibrium behavior. Analogously, expecting that the other will continue to act in the manner that equilibrium behavior suggests, Friday will have no reason not to act according to the equilibrium behavior that Crusoe expects from him. After both show equilibrium behavior, neither would have a reason to regret what he has done given what the other did. Among all options available for each, there is none that could have been used to reach a higher level of satisfaction given the equilibrium behavior of the other. In this state of equilibrium of directly (creating goods) and indirectly (taking or defending goods) "productive" activities, both actors might have, say, a satisfaction level derived from a net consumption of (4, 4). This would be less than what they would have if each were completely alone.

The preceding sketch of the seemingly exclusively negative external effect of the presence of "other" is, however, not the whole story. The relationship between Crusoe and Friday is almost exactly the same as between two sovereign nations that both know that they would be better off if they could only reach a state of disarmament (i.e. no time used in defense or taking). However, as between sovereign nations, Crusoe and Friday face the problem of uncertainty or a *lack of trust*. Nevertheless, they, being confined to their island, may learn to play something like "tit for tat" with respect to stealing and other kinds of

[10] See on this Buchanan's "The Limits of Liberty", volume 7 in Buchanan (1999 ff.)

infringements on the other player’s sphere. Then, like between neighboring nations, a kind of “peaceful” solution might emerge. In this solution, both players are “under arms,” i.e. still lose the resources for being armed, but keep the peace.

Neglecting defense and aggression efforts in subsequent considerations, let us assume that Crusoe and Friday are reciprocally refraining from infringements on the other individual’s sphere. The interesting point is that even under the unfavorable conditions of equal tastes and equal production possibilities, there are some chances of trade for Crusoe and Friday.[11] Remember, the two are identical twins as far as their options are concerned and in that they are mutually aware of their situation. Both are able to understand what is involved. That each has an understanding that there is another likewise understanding individual who can also act upon his understanding of the situation (a model of the situation) transforms the game in essential ways. This understanding can induce players to reason about the reasoning of each other and to consider acting differently than they would without such reflection. Factoring in reasoning and knowledge brings the discussion closer to a game theoretic one. Nevertheless, several additional steps must be taken before a game in the proper game theoretic sense emerges and that is the path we will now follow.

Imagine that Friday and Crusoe meet each other at the border of their territories. They, quite understandably, do not trust each other. From a safe distance, they start to communicate. Let us assume that complete information about the production possibility spaces of the two rational “players” on our fictitious island is the result of this more or less friendly chat.

The assumption of **complete information** as made by theorists of strategic interaction goes beyond **common knowledge** (a knowledge of which each knows that each knows that each knows etc.) of the actions open to each actor. It also includes what is often called “common knowledge of rationality”. At this stage of the discussion it is neither useful nor necessary to go into details concerning this assumption (we will come back to this in several ways later on). Suffices it to note that it means basically that the actors, whose behavior is being discussed, are assumed to know the theory that we develop in the model. In this sense, the model of rational choice analysis that is formed “in theory” is assumed to exist “in the world” (in the minds of the actors) described by that theory. A self-referential structure is implied in such rational choice modeling (theorizing). The model of the theorist is by the theorist’s assumption also in the heads of the individuals on whose behavior the theorist forms a theory. The theorist assumes that the individuals know what he knows and that they know about each other’s knowledge. They know that they know it, and they know that they know that they

[11] I am indebted to James Buchanan here who pointed out that the standard assumption of comparative advantage as in v. Mises’s “Ricardian law of association” may be weakened somewhat further; cf. v. Mises 1949, 158.

know it, and so on ... The theory becomes a kind of reasoning about knowledge rather than a reasoning about “nature.”[12]

In this vein, Robinson Crusoe and Friday know the whole setting and everything that is described here in the text, and they know that they both know it. They will immediately see that there are gains from trade that could be acquired should they both completely specialize. For the sake of specificity, we can imagine, for instance, that they may jointly consider that Friday could specialize completely in the production of steaks while Crusoe could completely specialize in the production of beer. Then there would be a total of 20 units of each of the products instead of 10 of each. Under a regime of equal distribution, for example, each of the two could consume (10, 10) if cooperating in both specialization *and* exchange. Each would clearly prefer this to a consumption of (5, 5), or, for that matter, (4, 4), for both would be better off.

Being endowed with the faculty of forming models of the situation, Crusoe and Friday might understand and, as we assume, will in fact understand all this. As rational individuals, they can and will know where the “bonanza” of co-operation and specialization lies. But, as we shall see next, because they are fully rational individuals they have problems to profit from the “treasure”.

Assume that from their safe distance the two come to agree on co-operation. Each of them agrees to specializing fully and to exchanging afterwards half of the fruits of his labor for half of the products of the other player’s toil. They intend to exploit the economies of scale that can be tapped by their division of labor to get from (5, 5) to (10, 10).[13] After the agreement, both leave for their production sites somewhere within their own territories.

It is important to note that Crusoe and Friday are *not* in the position of two men in a rowboat, who can (almost) perfectly monitor each other all the time. The two oarsmen instantly react to whether or not the other one is pulling as strongly as he should.[14] If one of the two lets go a bit, the other, observing this, could let go too. Then, the first may think it better to go back to the higher level of effort he showed before. As in our previous example, there will be some equilibrium level of effort that is optimal for the two oarsmen in the ongoing interaction. At least

[12] An advanced text on this is Fagin et al. (1995)

[13] Of course, if they were not identical, they could also exploit comparative advantages and what von Mises called the Ricardian law of social association would kick in, Mises (1949/1966), and extensively Kliemt (1986a)

[14] cf. on this Mackie (1980) who discusses Hume’s famous example. As Viktor Vanberg has suggested, it is very interesting to imagine different boats. It makes a difference for instance whether each of the men has only one oar, both of them sitting side by side, or whether both have two oars, sitting in a line. Then, if both want to reach some spot by rowing along a straight line, the rowing process would control itself insofar as every failure to pull sufficiently would lead to a deviation from the straight line. If both have two oars then the danger of shirking is much higher. Monitoring costs rise. Should there be a hundred oarsmen in a boat, it would be very hard for them to control the efforts of the others; cf. also Hume’s example of draining a meadow in 1948, 101 (book three, part 2, section vii of the “Treatise”).

within a common sense framework, we might expect it to stabilize itself in a “tit for tat” manner.[15]

Getting back to our island, we may expect the same mechanism to operate in principle in the case of the production and co-ordination game. However, the rounds of play in this case are further removed from each other and leave a broader scope for decision making and deliberation within each of the rounds. It is necessary or at least useful, therefore, to consider a single round of play, first. Such a round can, at least conceptually, be isolated from the context of the ongoing interaction in which it is embedded.

So, let us assume that Crusoe and Friday take such a round of play separated from all other interactions as a one-shot interaction. While each is engaged in his own production process, neither has any information about what the other is doing on the other side of the island.[16] Therefore, neither can make his own actions *directly* contingent on what the other is doing. And, for the time being, neither expects consequences in later periods of interaction as a result of the acts he performs “in-period,” so to say.

Under certain plausible assumptions about the material pay-offs resulting after deviations occur, the characteristics of the situation sketched before expose Crusoe and Friday to a so-called (prisoner’s) dilemma.[17] For the present stage of the argument, a brief outline will suffice. The prisoner’s dilemma basically rests on the presumption that Crusoe and Friday both understand that the production game is distinct from the later bargaining about the distribution of the fruits of specialization.

Crusoe and Friday agree on specialization and exchange as a “package deal” but have to execute their agreement incrementally. This conceals the more general problem of “the division of labor”. There is not only Adam Smith’s “invisible hand”[18] guiding humans to cooperate in exploiting the “gains from trade”, there is also the back side of that invisible hand. Let us look at this hidden side of the invisible hand more closely.

Assume that Crusoe and Friday agree that they will both specialize and then exchange the fruits of their labor as specified in the agreement. However, as rational beings, they distinguish between two problems, first, specialization in production and, second, bargaining about distribution. This makes it hard for them to achieve what they understand to be in their common interest.

Because of the symmetry of the decision situation, it suffices to look at the interaction, or the “*game*,” from the point of view of one of the participants. Let

[15] As we shall see below, game theory may suggest that among strictly rational players “tit for tat” and other similar strategies might not work.

[16] The game provides complete but not perfect information.

[17] Nowadays the character of exchange is more commonly understood as a pd. For an early presentation in English see also Hardin (1982)

[18] See for a good collection of relevant excerpts from the Scottish Moralists, Schneider (1967).

this be Crusoe (all considerations apply to Friday analogously). As a rational actor, Crusoe must consider basically two events that he cannot causally influence by his own actions after going back to his side of the island. The two events are dependent on the choices of Friday, who may or may not keep the agreement.

Crusoe will reason thus: "On the one hand, Friday may specialize as promised, i.e. Friday keeps the agreement and shows co-operative behavior C_F. On the other hand, Friday may refrain from full specialization, i.e. Friday does not stick to the agreed terms and shows potentially exploitative behavior or defection D_F. I can either co-operate myself ($C_{Cr)}$ or not (D_{Cr}). What should I do?"

For the sake of specificity, let in the case C_F, Friday's partner Crusoe consider specializing only marginally to, say, (11, 2) as his defection alternative D_{Cr} (assuming that this is an alternative in his production possibility space; see for all this table 1.1 below). Crusoe, thereby, realizes his alternative of breaking the agreement. He is himself not anymore in his own most preferred consumption position (5, 5), which he would have realized under isolated production. However, he anticipates that he will be in a strong bargaining position in the final share out game which is to be played after production. In anticipation of this, he speculates on D_{Cr}.

In this scenario, due to his own deviation from his promise, Crusoe comes to the exchange with (11, 2). Friday, whom Crusoe assumes has chosen C_F as the first contingency to be considered in his, Crusoe's, considerations, enters indeed into the exchange game with (0, 20). Crusoe can expect to be able to exploit Friday's weakness. He might hope to exchange, say, 1 of his 11 units of beer for, say, 18 steaks and end up with (10, 20) while Friday would have to live on (1, 2), which we assume, for the sake of our simple numerical illustration, Friday would still prefer to (0, 20). Had Crusoe kept his promise and had the final bargaining led to the result agreed upon, Crusoe would have ended up with (10, 10) and so would have Friday. However, Crusoe prefers (10, 20) to (10, 10). Therefore, under the present extremely simplifying assumptions, Crusoe must conclude that it is better for him to realize D_{Cr} *if* Friday keeps his promise (plays his cooperative strategy C_F).

The other contingency that Crusoe could not influence anymore after the two agreed and departed for their separate parts of the island is the possibility that Friday may deviate, D_F (again assuming only this one possible deviation). If Friday plays (2, 11), then Crusoe would be better off deviating himself to (11, 2). For, playing cooperatively or specializing as promised to (20, 0), he would only expose himself to subsequent exploitation and end up, symmetrically, with (2, 1). The position (11, 2) is better than that (and (5, 5) would have been even better). Starting exchange with (11, 2) if Friday brings (2, 11) to the negotiation, Crusoe could – in view of a plausible solution of the bargaining game ensuing from (11, 2), (2, 11) – hope that from the final share out a distribution of, say, [(7, 6), (6, 7)] might emerge.[19] This is clearly better than winding up with (2, 1).

Therefore, *regardless* of what Friday does, regardless of whether Crusoe expects the event C_F or the event D_F to occur, it is always better for Crusoe to break his promise. Either he can hope to exploit the weakness of Friday, or he at least insures himself against exploitation by Friday.[20] As remarked above Friday will reason symmetrically. **In sum**:

Column = Friday
Crusoe = Row
	$C_F = (0, 20)$	$D_F = (2, 11)$

$C_{Cr} = (20, 0)$	(10, 10) (10, 10)	(2,1) (20, 10)
$D_{Cr} = (11, 2)$	(10, 20) (1, 2)	(7, 6) (6, 7)

Table 1.1: The Prisoner’s Dilemma structure of exchange

The upper left entries in each cell of the table refer to Crusoe while the lower right entries show Friday’s pay-offs as evaluated by the measuring rod of the goods received after playing either co-operatively or not. We have “preferences”[21] with

Crusoe : (10, 20) $>_{Cr}$ (10, 10) $>_{Cr}$ (7, 6) $>_{Cr}$ (2,1) [$>_{Cr}$ (20, 0)]

Friday : (20, 10) $>_F$ (10, 10) $>_F$ (6, 7) $>_F$ (1,2) [$>_F$ (0, 20)]

where “$>_{Cr}$” indicates that Crusoe strictly prefers the result on the left to that on the right of the “>” symbol, and “$>_F$” indicates the same for Friday.

If an act or a strategy *a* is strictly better than an alternative *b* regardless of whatever else may happen independently of that act, we say that *a* **strictly dominates** *b*. Since the result of D_{Cr} is strictly better than C_{Cr} from the point of view of Crusoe regardless of what Friday does, we may state that D_{Cr} strictly dominates C_{Cr}. No matter what Friday actually does, the dominant D_{Cr} leads to better results than C_{Cr}.

Since D_{Cr} dominates C_{Cr}, no strategic thinking on the side of Crusoe is necessary. He need not put himself in the shoes of Friday. If, in his own shoes, he understands that D_{Cr} dominates C_{Cr}, he can choose D_{Cr} regardless. No matter what

[19] Assuming as before that the units of goods are indivisible.

[20] M. Farrell pointed out that we implicitly assume here that Crusoe does not expect Friday to reverse his specialization and to specialize in the same way as Crusoe himself. This perceptive remark raises interesting questions about the complete game tree and the distinction between “trust” (which is lacking) and “expectation” (which is present as a factor of coordination). We cannot go into details here but would only like to suggest that the distinction would be closely similar to that between a convention in the sense of David Lewis and a prisoner’s dilemma norm.

[21] More on this below. For the time being, an intuitive grasp of the fact that there is an ordering according to better or worse suffices.

an actor assumes or knows the other actor will do, the actor who is in command of a strictly dominant alternative knows what he should choose no matter what. His own expected results tell it all.

In sum: A minimally rational Crusoe exclusively interested in beer and steaks for himself will have to choose a dominant strategy. Should he choose otherwise, it would indicate that he would be interested in aims, ends, or values other than beer and steaks for himself (i.e. other dimensions of value play a role).

That Crusoe be only interested in the two goods and, for that matter, only in himself is not a requirement of rationality; it is a substantive assumption of a different kind. However, if – as we assumed – these actors are only interested in the aims we name "beer" and "steak" as affect themselves, then it is a fundamental requirement of minimally rational behavior that a dominated strategy is never chosen.[22]

Friday is as rational as Crusoe. He has a dominant (defective) "D_F" strategy too. If both play their dominant strategies as suggested by the precepts of individually rational behavior, then the two of them will end up with (7, 6) and (6, 7), respectively, though they could realize (10, 10) for each if sticking to the terms of their agreement. Though both act fully rationally, there is a result in which both are better off than in the result reached by acting individually rationally. Such a result is called a **Pareto superior state**. They unanimously must hope that this result be realized instead of the one they reach by playing their non-dominated strategies. However, the result can only be brought about by the use of dominated strategies and, thus, by a violation of the fundamental principle of individual rationality to never use strictly dominated strategies.

Rational individuals are aware of this. In pre-play communication they may be quite willing to agree that in order to reach the Pareto superior result they should both play a dominated strategy. However, since their agreement specifies a dominated strategy for them to play and since their rationality is common knowledge among them, neither can trust that the other will stick to the agreement.

The very same faculties, from which the potential gains of cooperation and exchange stem, can hinder the realization of those potential gains. If humans could not take opportunities, they could not improve their situation by co-ordinating on new forms of the division of labor or of specialization and exchange. Crusoe and Friday may recognize the opportunity that is offered by the division of labor, but, because as rational individuals they cannot rationally trust each other, they cannot seize the opportunity. A prisoner's dilemma is the back side of the coin showing on its face the potential gains of the division of labor, of joint use of resources, of cooperation, of specialization and subsequent exchange[23].

[22] After, all dominated strategies are, no matter what, worse on the actor's own terms than at least one alternative.

[23] Contrary to traditional theory, it is maintained here that joint use of resources does not emerge *on* a market only. The market *itself* is viewed *as* a form of joint use of resources.

Traditional economic theory conceals the problematic side of human cooperation by modeling the phenomena of specialization, exchange etc. as a so-called "co-operative game" from the outset. Economists tend to assume that the agreements will be kept because there are social mechanisms in place that make keeping promises more advantageous than breaking them (not necessarily state-sponsored law based but possibly convention based mechanisms).

Controlling the problematic side of human cooperation is the very essence of any institution of human cooperation. The market is no exception to this. Like a company, a club, a government, or any other institutional competitor that we may analyze in a comparative institutions´ framework, the market has to provide remedies for the monitoring (one cannot directly and instantaneously observe what others do) and hold up (one may be confronted with a kind of ultimatum) problems which give rise to the fundamental specialization dilemma described before.[24] Neglecting this dilemma will distort our view of the functions and workings of the market. It will conceal from our views rather than reveal to us the fundamental problems that can emerge from human rationality itself.

In sum, rationality can prevent rational actors from realizing prospects opened up to them by means of their own rational faculties.

Interactions of the kind described above occur rarely in isolation. In social reality, many if not almost all games are not one-shot but are rather embedded into an ongoing interaction (cf. Granovetter (1985)). For instance, Crusoe and Friday are on an island. Being naturally confined to the island is a functional equivalent to a contract relationship. As if they were chained together by the legal relationship of a long term contract, they are locked on the island and, thus, cannot avoid interacting with each other in the future. In *this*, they can – and must – rationally trust. The players know in advance that the same game may be played over and over again.[25] They are involved in a "supergame." Though their productive activities cannot have a causal influence on the productive activities of the other player at the same stage of the ongoing interaction (or at the same "normal" or "stage" game), they can have a causal influence on subsequent actions of the other player. For instance, after being exploited, the other player may be expected to retaliate on some or all future rounds of play either by deviating or by refusing to enter into any further negotiations.[26]

What has been called the "shadow of the future" prevails, and so-called "conditional supergame strategies" like the aforementioned "tit for tat," which

[24] E-bay has helped in driving home the message of the necessity of organizing exchange; see on this Ockenfels (2003) and Güth and Kliemt (2004).

[25] Of course, a marriage is another example of this genre. Still another is the founding of an ongoing common enterprise offering quasi-rents after devoting resources specifically, see Alchian (1984) Alchian and Woodward (1988)

[26] This would be an iterated prisoner's dilemma with an exit option as studied in Schüssler (1990) or Vanberg and Congleton (1992).

relates behavior in a previous period to behavior in a later period, can account for some amount of co-operation if the other player has co-operated before. In this case, co-operation will no longer require the use of a dominated strategy. This makes it possible that regularly co-operative behavior on the parts of both players emerges as the outcome of individually rational strategic behavior, or so it seems. (We will see counter arguments against this standard view of the matter later on, see chap 5).[27]

Being confined to an island (i.e. being forced to stay there by a non-rational external factor), Crusoe and Friday can reach the Pareto improvement stemming from specialization or the division of labor rationally. However, social interaction in general does not take place exclusively between players who can trust that they will be permanently confined to dyadic interaction. Social institutions rather create larger games that break up ongoing bilateral monopolies and allow for transactions between changing partners. The bargaining position of those who have specialized is much stronger in these games since they can trust that they will find somebody else who is willing to exchange goods and services with them. Exploited partners can exit from interaction after they have been exploited. In that way, large markets “automatically” economize on the trust (and virtue) needed to induce specialization and the division of labor.

Still, there will always be a realm beyond which specialization and the division of labor cannot (or for reasons of “transaction” costs “should” not) be based on enforceable contracts. It is exactly here that other mechanisms become useful substitutes for explicitly specified externally enforced contracts. What this actually implies will be discussed in more precise terms in the subsequent discussion. For the time being, it will suffice that we have taken a first look at social interaction through the window of “informal rational choice theory.”

There are many other interesting aspects that could be observed here. For instance, the table 1.1 already shows that both Crusoe and Friday may be better off by the presence of the other even if there is no trust between them and each uses their dominant strategies. If they specialize marginally to get the better of the other one, they may not reach that aim but still be better off than when “bowling alone.” Though this is merely a possibility exemplified by arbitrary numbers, it should still make us think twice.

Other aspects are also well worth some, perhaps even many, second thoughts and second looks through other modeling windows. Some of them hopefully will be opening up in the next chapters. A philosophical window will be opened first.

[27] The seminal social science contribution after the even more seminal work of David Hume is Taylor (1976) The term “shadow of the future” and other suggestive models can be found in a more popular form in Axelrod (1984)

2 Dual Ways of World Making

The human actor's ability to distinguish between what are and what are not causal effects of choices and to imagine himself as the "uncaused" author or maker of choices is put into a broader philosophical perspective in this second chapter. Two basic economic perspectives, one framed in terms of *choice making from a participant's point of view* and one framed in terms of *objective explanations of choices* according to behavioral laws, are distinguished. To understand better what is at stake, the "subjectivist" approach, which seems to go against the grain of modern "objectivist" science, is placed within a perspective of the history of thought. Relating it to Kant, Strawson Morgenstern, and Buchanan will hopefully induce those who are inclined to discard the first perspective altogether to have some second thoughts about their own views "on the nature and significance of economic science." It may have a stronger humanities streak than expected, even in some of its most mathematical variants.

Economists want to be able to take the concept of reasoning about knowledge to the extreme by adopting an internal point of view to idealized thought processes of fully rational actors (which some of the most sophisticated and intellectually interesting economic theory does). At the same time, they claim that they are working within behavioral science and approach human behavior from an external point of view in which the reasoning and knowledge of the "behaving" individuals does not play a role and need not be understood by the researcher (often objecting even to cognitive psychology).

Looking at the world through different windows, economists can, perhaps, have it "both ways." On the one hand, they can see the world of interactive decision making as a scenery of reasoning about knowledge while, on the other hand, framing interactive decision making as if it took place in a Skinner box and could be explained in basically the same, though somewhat more complicated, ways as the behavior of animals. Both views of the world may have their place in economics as well as in philosophy. As long as we are aware of the fact that these different perspectives exist and are able to distinguish between them, the ability to adopt them both may outweigh the dangers of confusion. How the two worldviews relate to each other remains an open and difficult problem though. To this we turn next.

2.1 Layers of things and theories

The view that humans are citizens of two worlds, a lower “material” world and a higher one of reason has a long history. In the most modest formulation of this, we can state that man perceives himself as a physical being subject to the “normal” laws of nature and at the same time as a rational being guided by reason. The basic dualism can conceivably amount to two different things. On the one hand, the two worlds are assumed to exist in some sense or other “out there,” and the membership in them is taken as real. On the other hand, there is only one world, but we can look at it through two different windows.

In the first ontological interpretation, the risk to end up with contradictions between different laws operative in the two worlds is obvious.[28] In particular, as far as overt behavior is concerned, we cannot have it both ways. In the end, there is one kind of overt behavior that must be predicted and/or, after its occurrence, be explained. Accordingly, there should be one set of law-like regularities that all operate in well-specified realms without contradiction.

Though it is conceivable in principle that in some instances laws of the one world and in other instances laws of the other world prevail in explaining overt behavior, this view does not seem to be too appealing. The mind set at least of modern times is used to thinking of one “world.” This world is of one kind of “material” to be explained by one type of explanations.

Nevertheless, the stuff of the world is composed of layers of phenomena. The layers are hierarchically ordered, and some are more fundamental than others. As far as this is concerned, reductionism in some ontological sense reigns. Certain “things” are more basic than others. The less basic “things” are made up of the more basic “entities”.

The scientific perception of the ontological layers parallels the layer of things. Putting it rather crudely, physics spells out our basic views on how physical entities interact (and also what those basic entities are). Then comes chemistry, which is based on its own law-like regularities. However, according to the prevailing view, chemistry – though practically an independent realm of inquiry with its own methods – is “reducible” to physics. This means that its entities can “in principle” be construed from the objects of physics and its regularities can be “derived” (approximately) from the laws of physics. Then as a next theoretical layer, we have biology, which “in principle” is held to be reducible to chemistry

[28] Parallel world problems, occasionalism etc. come to mind here.

and, thereby, in the last resort, to physics. In the next step, we reach psychology, which studies (mental) processes "in the psyche" but possibly also their dependence on chemical and physical phenomena as occurring within biological entities (see on reductionism reduced to the essentials Stöckler (1991)).

Particularly with respect to the mental, there are considerable doubts about its reduction to the physical even in theory. Still, certain difficulties notwithstanding, according to the prevailing view of the world, a hierarchy of theories and a hierarchy of "things" basically run parallel. Thus, the most fundamental elements are physical entities. Of these the chemical substances are formed. From the chemical substances, we reach the biological beings, and finally, within these biological beings, we encounter the mental phenomena that may or may not be mere epi-phenomena of the physical or chemical substances.

If we accept the preceding outline, we must admit that all advances in our knowledge of the laws of nature and the underlying natural processes and substances have not brought us philosophically much further than we were at the beginning of modern times. To use the succinct and apt formulation of Hobbes, "matter in motion" is all there is (Raphael (1977)). To the modern mind – at least in its secular, commonsensical manifestations, this is still "what there is." Yet, many questions remain open. Those who try to deal with human behavior are left with the task of clarifying the role of the mind in this picture (either as an epi-phenomenon of material events or as an entity in its own right). If we accept the one world thesis, we must somehow account for the naïve dualism of our day-to-day experience. Even if it is no longer taken as an ontological one, the phenomenological distinction between the mental and physical must nevertheless somehow be explicated. And, this must be done in ways that do justice to our phenomenological experience along with our scientific convictions about how the world works.

Like the chemists who can get along quite well with the physicists, since the reduction of chemistry to physics takes place only "in principle," the psychologists discussing mental or social phenomena may well believe that the mental is merely an epi-phenomenon of chemical processes without coming to the conclusion that psychologists should now take to chemistry. The claim to conceivable reduction may matter as a basic scientific outlook, in which the basic paradigm of a science is fixed but will as a rule have no direct impact on day-to-day scientific practice within the established paradigm. As long as reduction is viable "in principle" it does not mean much for beings with our limited knowledge. As far as ordinary explanations and predictions are concerned, no reduction is possible.

There are, of course, instances in which specific phenomena in the psyche can be triggered by some chemical or electrical impulse. This underpins, at least in a way, the basic view of layers of things that hang together in a specific hierarchical way. Nevertheless, there is no way to explain everything that can be captured on a higher level of theory formation in terms of the lower level.

In sum, whatever "reduction" of theories and the so-called unity of science deriving from it may mean, in practice it cannot amount to doing only physics. There is a legitimate role for very different outlooks on the world. This being said, let us turn to some classical views on basic dualisms.

2.2 Being is perceiving (esse est percipii)

It is a brute phenomenological I-fact (fact from the first-person perspective) that to predict a choice from an onlooker's or an observer's point of view is different from making a choice from a choice-maker's point of view. Using Herbert Hart's well established philosophy of law terminology (see Hart (1961)) in a slightly different (though related) way in our context, we can say that, with respect to an actor's choice making, adopting an "**external**" point of view is different from adopting an "**internal**" one. The external point of view is associated with showing what Strawson called an **objective attitude** of treating other individuals and their behavior as parts of nature (see Strawson (1962)). The behavior of others is seen as part of natural events that are caused and can be predicted according to causal laws. As opposed to that, the internal point of view starts from understanding the acts of another individual with a **participant's attitude** (see again the terminology in Strawson (1962)) in the way they appear to that individual herself. In that perspective, choices do not happen to the actor; she is rather "making her choices."

In particular, predicting (external perspective) and doing (internal perspective) something are experienced as categorically different. Adopting an "objective attitude" corresponding to the external point of view, we can accept modern science, live in one world under causal laws, and predict and explain its course accordingly. Adopting a participant's attitude, we perceive our choices not as caused but as made, while their making is understood from an internal point of view even if it occurs in actors other than ourselves.[29]

Yet, can the objective and the participant's attitude be coherently adopted by the same person? Can "I" fully accept natural science and keep up human self-conceptions as a free and responsible actor? Must "I" in the first person perspective substitute choice by prediction if "I" accept the causal view of the world? Can "I" endorse the view of a causally closed world given the brute fact

[29] Sometimes the categories of "understanding" and "explaining" are seen as referring to broadly the same distinction. However, it should not be forgotten that explaining things from an objective point of view is a method – and perhaps the best we have – of understanding them. Moreover, understanding in non-objective categories often amounts to an effort to establish a competing explanatory concept. The latter may be futile, but that is not to say that that effort is not made and, if made, would not influence our view of the world and possibly the world itself.

that – at least for me – doing some of my acts is categorically different from predicting them? Can the scientific view of actions as events "happening to me" and the conceptualization of actions as "made by me" co-exist? In answering such questions, we need not dive into deep ontological waters. There are more down to earth ways to deal with these issues. For instance, we can opt for an interpretation of the two worlds thesis in terms of "perspectivism." We take different points of view when approaching what may well be one and the same thing.

As is the case with that well-known picture that shows a young woman if you look at it one way and shows an old woman if you choose to perceive it differently, you can look at the one world through two different windows. Depending on which basic "theoretical" approach or perspective you share, you will see different phenomena when you look at the next figure.

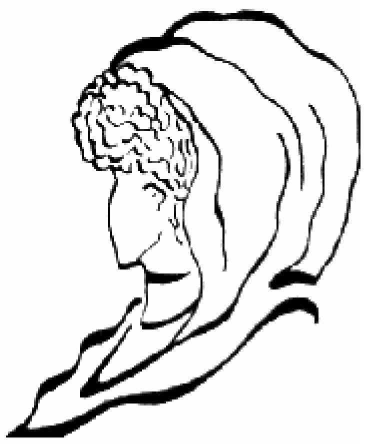

Figure 2.1: Old vs. young woman

What do you see in the next picture, antilopes looking down to the left or birds looking up to the right?

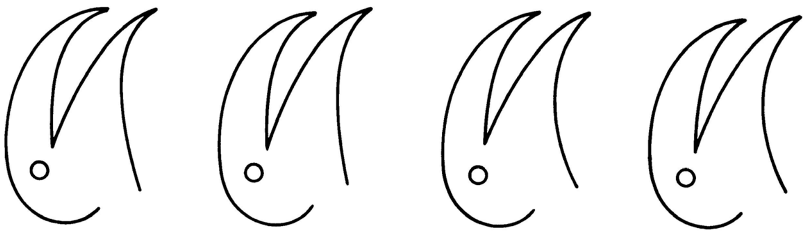

Figure 2.2: Antilopes vs. birds

You can see both and switch at will. However, there is more to the possibility of looking at things differently and seeing different phenomena without any change in the underlying objects as such; there may also be errors that are hard to get rid of as the next figure shows.

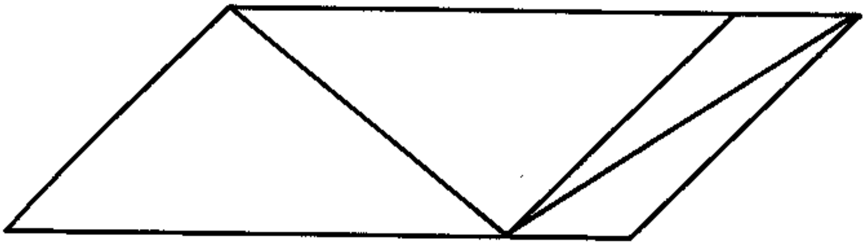

Figure 2.3: Sander parallelogram

Looking at the so-called Sander parallelogram of figure 2.3 you could be deceived into believing that one of the diagonals, the one that is in "objective truth" shorter, is considerably longer than the other one. Different than in the case of the "young vs. old lady" deception, in the Sander parallelogram case, even if the perceiving individual is actually informed about the deception, she will not be able to perceive things otherwise. The perception is still there even though the perceiving subject may now judge herself to be misled. Expressing what we perceive, we are aware that we are expressing a falsehood. Yet, the judgment notwithstanding, we "perceive what we perceive." We perceive the diagonals as differing in length, which may lead to wrong assertions, but we nevertheless can only perceive it the way we do. The practices we have been brought up with are so strong that we would have to be retrained in a new practice rather than merely be informed about the truth.

In the case of the two worlds' thesis, we may assume that, as in the case of the parallelogram, there is one true world out there and that we must be deceived into believing something opposed to the true view of the world. What there is, is in the last resort decided according to the facts and from an objective point of view. According to other views, the whole concept of a world existing out there independently of how we perceive it does not make sense or is not very relevant. We always approach it within the context of established practices.

The deeper philosophical issues are again not of great importance for the concerns at hand (for a critically rational defense of "realism", see Albert (1985)). In a context in which human action is influenced by what the individuals perceive, the fact that they do perceive certain things in certain ways is crucial whether it be due to a deception or not.

In sum, to trigger different behavior, it suffices that the two worlds of which we are speaking both *exist in the eye of the beholder*. We may talk about the two perspectives as existing *for us*. They exist as phenomena without any claim to the effect that they *only* exist in the eye of the beholder (though that may well be the case).[30]

After brushing aside all deeper philosophical issues, what remains of the two worlds thesis is, from an ontological point of view, a very moderate claim. It does not amount to more than assuming that human beings as a matter of fact can approach the world in two different ways. We can chew gum and walk at the same time. Even though there are some individuals who cannot proverbially chew gum and walk at the same time, there are others who can; the fact that they can do both affects their behavior.

Subsequently, I will assume that economists and philosophers can in fact both look at social interaction as participants from an internal point of view (i.e. internal to the decision making entities) and objectively as external onlookers. Yet, even if we insist that adopting both perspectives *within* economics and philosophy is possible, it is still true that the two ways of “world making” are distinct. This has implications for rational choice approaches to human behavior in general and for economics in particular. According to the one side of the discussion, economics is or at least needs to become a hard science based on objective observations of overt behavior and explanations of that behavior in terms of behavioral laws. According to the other side, there is a subjective world of intentions that must be understood from a participant’s point of view.

It is somewhat unclear how economics from a “participant’s point of view” relates to overt behavior.[31] Even if we disregard such exaggerations as may come up in the more extreme so-called “Austrian” quarters of economics[32], a look at the history of some of the most respected parts of modern economics demonstrates that the non-behavioral point of view should not be brushed aside lightly. The present state of rational choice theories of social interaction, which form part of, if not the core of, the discipline of economics, can be adequately understood only if

[30] Whether in the end we conceive of “the” world(s) as made or found (or both made *and* found) may also be left open for the purposes at hand – even though it is, of course, another interesting epistemological and theory of science issue.

[31] In the more extreme forms of expressing that there is a special role for understanding, we will find in philosophy the hermeneuticists and in economics the a priori theorists of human action. Economic theory becomes an a priori theory of rational action (a la Mises (1949/1966)) or a “hermeneutical” endeavor of understanding deeper meanings of actions and intentions and the like. It can be done analytically by spelling out the logic of human action and it is here that notions of the “logic of the situation” as endorsed by Popper become somewhat problematic.

[32] Austrian economics is a subjectivist approach to economics which is critical of some crucial aspects of main stream neo-classical economics. Ludwig von Mises, see the preceding footnote, would be the first to be mentioned here. Many others could be noted but since I will merely focus on the intimate relationship between the subjectivist view and the origins of game theory I need not dwell on Austrian economics any further otherwise.

we take into account some version of the two worlds conception.[33] Therefore, let us turn to a stylized, if somewhat coarse, account of a few of the many expressions of the two worlds thesis and related dualisms in philosophy and in economics. This will show how the two worlds of philosophy and economics run parallel and, for that matter, will lead to a better understanding of how these two worlds run in parallel within economics as well.

2.3 From and to the skies

2.3.1 Philosophers coming down from the sky

When it comes to a world of ideas or a world of the rational apart from everyday phenomena, the name of Plato comes to mind immediately. Platonic ideas are assumed to form their own world. In modern times, Descartes was the most crucial figure in developing dualistic conceptions of the relationship between mind and matter. Yet, like Plato, Descartes is associated with ontological claims to an independent existence of the non-material world that are highly problematic. At least from the point of view of the present author, they are hardly acceptable. So let us turn directly to Kant, who is most relevant for our present concerns. For, he had a surprisingly strong, if mostly unnoticed, indirect impact on some of modern social theory, including economics.

2.3.1.1. Kant

Kant clearly understands that the world view of modern science coincides with the aforementioned (also Hobbesian) one according to which there is nothing but matter in motion, and the motions of matter are subject to the laws of nature.[34] Kant was struck by the problem, presumably unduly so, of how to account for human moral responsibility, liberty, and the like in a world seen as governed by causal laws. For him, the problem was dramatic. On the one hand, he was willing to accept that the world conceptualized by modern science is one subject to general causal laws operative across the board. On the other hand, he was deeply committed to more or less Christian notions of responsibility and freedom of the will that seemed incompatible with the view of a causally closed world.

How could there be a sphere of freedom within the one world reigned by natural laws? If all events are subject to causal laws, then human acts as events should also be subject to those laws. It should, for instance, be possible for an

[33] Quite tellingly so-called behavioral game theory is a follow up of game theory.

[34] These are not the Hobbesian “laws of nature” in “The Leviathan,” which are norms, but descriptive laws of nature in the modern sense of regularities; on this also see Beth (1965), referring to Hans Kelsen’s analysis of “law”.

omniscient scientist to predict and to explain human actions in terms of causal laws only. In the extreme, it could be predicted whether a person would rather go to the cinema or to a boxing match in the evening by finding out, say, whether she had eaten beans or peas in the morning.

Such predictions would be viable if there were laws that made it more likely to develop a preference for boxing after eating peas than after eating beans. Should such laws exist, then the choice behavior of an individual could be predicted from an external point of view the very same way as the flight curve of an individual who just jumped out of a window.[35] What the individual herself would be thinking about her choice behavior as well as her intentions would *for the external observer* be as irrelevant for her overt behavior as corresponding thoughts about her flight after she jumped off the windowsill.

In sum, adopting a purely externalist view to human behavior is clearly possible; however, this is not how we conceive of ourselves, nor is it the way others conceive of us. Expressed in Kant's somewhat weird mixture of Latin and Greek terminology, we conceive of ourselves simultaneously as "homo noumenon" and as "homo phaenomenon."[36] As actors, we behave according to the laws of nature and at the same time act according to the "laws of freedom" or, less emphatically stated, according to our intentions, aims, ends, or values (or what we conceive of as such).[37]

How "true" freedom can be possible without being merely an illusion is a difficult problem. We should not burden our discussion with Kantian notions of "transcendental presuppositions" or with assuming some causal influence on nature not accessible to natural laws. We are better off turning to what may be seen as a down to earth variant of the Kantian approach that can be used to the same effect but more or less without such ontological commitments. And, this brings us back to the aforementioned categories of objective and participants' attitudes in world making.

2.3.1.2. Strawson

The aforementioned Peter F. Strawson presented a variant of the Kantian approach in his "Freedom and Resentment" (see Strawson (1962)). In this seminal article, he uses the distinction between a participant's attitude and an objective attitude towards others in a somewhat stronger norm-laden sense than we will

[35] The promise of creating completely externalist explanations is, as may be noted in passing, the strongest attraction of neuro-eonomics for those economists who resent the "internalist" elements of explanations that seem part of the humanities. Even cognitive psychology seems too far from direct observation.

[36] "Homo" for "man" and "noumenon" for the non-physical. In a somewhat different vein, see also Heinimann (1987/1945)

[37] Of course, in Kant laws of freedom would have their own special technical meaning related to self-legislation by the actor.

subsequently, but the inspiration lies in his distinction. On the one hand, we treat others as free and responsible actors towards whom we feel resentment and the like. On the other hand, we know that the others may be subject to forces of a causal nature that can drive them to act in ways for which the ascription of responsibility seems at least doubtful.

The example of inflicting punishment is useful to illustrate what is at stake here. Punishment can be administered like medicine. In this perspective, it serves a future-directed purpose of behavioral modification. The aim is to prevent future harm as may originate from the offender himself or others who observe the behavior of the offender and/or responses to it. So-called specific and general prevention are the aims of exerting such causal influences on another individual.

With an objective attitude, punishment is guided by concerns other than retribution. Though such punishment may be administered only within certain constraints of justice, its instrumental usefulness entirely relies on behavioral rather than normative laws. Within the legal normative limits, the objective attitude towards the individual prevails. There are behavioral laws that predict how the punishment will affect individuals. If, according to those laws, punishment is the best way to bring about the desired effect, then it should be imposed. If instead of punishment a reward could bring about the same effect at a lower cost, administering the reward would be preferable (other things being equal). If behavioral training and education, perhaps even brainwashing, worked best, then – perhaps within some external constraints of a normative nature again – relying on such instruments would be the right strategy of intervention.

Adopting an objective attitude, the task of designing institutions of punishment becomes one of social engineering altogether. However, administering punishment with a participant's attitude towards the offender should not be seen within the context of social engineering only. There may be a social engineering component involved, but, besides this, retributive aims play an essential role. The other individual, or offender, is approached as a person who is held responsible in a way that goes beyond administering punishments and rewards strategically. Retributive emotions are expressed here.[38] Yet, these emotions are not simply blind drives that spring up. They are rather triggered within the framework in which freedom of action is ascribed to others and resentment is expressed and felt.

In sum, when participating in an interaction with another individual, we treat our counterpart as a person in the full sense only if we approach him as a free and responsible actor. Thereby, we ascribe to an "other" qualities that go beyond our objective attitude towards natural phenomena and develop corresponding

[38] See Mackie (1982), and on the modern experimental literature on this rather old insight, for example, Fehr and Gächter (2002). Concerning the relationship between retribtively holding others responsible for consequences or for their intentions some experimental evidence can be found in Güth et al. (2001).

emotions. To that effect, the interaction situation must be framed such that the acts of others are not seen or perceived simply as emergent under natural laws. Yet, it is important to note that this does not commit us ontologically to much. According to the basically Strawsonian and empiricist account adopted here, this perspective does not necessarily carry with it the ontological commitment that a world other than the one subject to natural laws does in fact exist.

2.3.1.3. First and other persons

It may well be that all our own acts as well as all the acts of others are events occurring under the laws of nature. It may well be that all there is "matter in motion" and that the springs of action are of the very same kind as the forces that make an apple fall to the ground. However, when engaging in inter-personal relations, we can frame the interaction otherwise.

To approach another individual and the interaction with that individual as an inter-personal one with a participant's attitude implies that we see him as a "*doer*" (see Pearl (2000)), as a source from which causal effects on the world originate according to reasoned choices. We do *not* think of his actions as the result of causal forces even though we can switch to that perspective. As long as we adopt a participant's attitude towards some other person, that person is for us not simply the victim of his impulses, he is rather framed as the *author* of his deeds.[39]

We ascribe to others an I-perspective and intentionality, both of which we project from our own experience. This brings us back full circle to the aforementioned brute phenomenological I-fact, according to which predicting a choice from an onlooker's external point of view is different than making a choice from a choice maker's internal point of view. In the first-person perspective, we conceive of ourselves as the authors of our acts. Participating with others in interactive decision situations, we ascribe to them the same self-conception we ascribe to ourselves. (To make a bad pun, we are "eyeing them" that way.) Whenever we do that (ascribe an I-perspective), we subjectively *participate* in a world of interaction that is different from the world we experience when looking at interactions from an objective point of view.[40]

[39] As we know from other contexts, framing has effects, see Kahneman and Tversky (1984)

[40] That we all have corresponding feelings is obvious from the fact that we all regard basically the same kinds of actions as expressing category mistakes in such matters. For instance, the Persian prince who had the sea beaten up after he lost his fleet is smiled at by practically all of us. According to our view of the world, one does not express resentment towards the sea. The sea is not a free and responsible actor. One should note also that the category mistake is something other than merely a simple error. For, had the Persian prince conceived of the sea as a goddess, we would not have blamed him for a category mistake but for endorsing a false theory about the world (assuming that we think that there is no such goddess).

2.3.1.4. From philosophy to economics

If "science" were restricted such as to imply an objective external point of view on explanation and understanding, then there is much more non-science in economics than in other fields of social inquiry (and this, though making life for the researchers difficult, seems to be one of the reasons why economics and philosophy get along so well with each other). If economists think otherwise and assume that their theories are more scientific than other social theories, then this is due to their use of high-powered mathematical tools. Yet, quite ironically, these tools are used in particular in realms of economics that originated in approaches, like classical game theory, that can be adequately understood only in terms of adopting a participant's attitude to social interaction.

In the case of classical game theory with its focus on reasoning about knowledge (starting from common knowledge), the point that there are elements other than those that meet the behaviorist's eye seems fairly obvious. However, before turning to classical game theory a closer look at the revival of classical political economy in the work of James M. Buchanan may be helpful. It will serve as *an example* of an enterprise that, according to its own self-understanding, is *not* a behavioral science and nevertheless claims to be, and is generally accepted as, "economics" rather than "philosophy".[41]

2.3.2 Economists up in the sky

2.3.2.1. Buchanan

2.3.2.1.1. The participant's point of view in Buchanan

In his "What Should Economists Do?" (see Buchanan (1985)) Buchanan presents the example of Crusoe and the chimpanzee. Crusoe intends to keep the chimp off his fields. He, therefore, draws a curly line in the sand to produce the image of a snake. This is manipulation pure and simple. Crusoe adopts an objective attitude towards the chimp. He exploits his knowledge of natural psychological laws governing the psyche and, thereby, the behavior of chimpanzees in general. These laws apply to the specific chimp Crusoe intends to manipulate and yield the relevant technological predictions in a means-ends framework here.[42]

Buchanan is well aware that Crusoe could adopt the same attitude towards Friday. However, as Buchanan points out, unlike to the chimp, Crusoe can relate to Friday in ways other than manipulation. Treating Friday as a person rather than a "normal" part of nature seems possible without committing a category mistake.

[41] Classical game theory is endorsed by Buchanan as the most important innovation of 20th century economic theory, see Buchanan (2001).

[42] This is a direct application of a kind of covering law argument, see Hempel and Oppenheim (1948)

Being human, we intuitively believe that Crusoe can enter into an *inter-personal* relationship with Friday in ways he cannot with the chimp.

As far as the reasons for the special relationship between Crusoe and Friday are concerned, more than personhood seems to be involved. Chimpanzees, like most human beings, according to rather plausible criteria of personhood, have to be classified as persons.[43] Therefore, clearly, ascribing personhood to another individual is not sufficient for approaching that individual in interaction with a participant's attitude. Something else must be involved.

Stating in general terms what the extra quality of interaction that allows for adopting a participant's attitude might be is very complicated. Clearly, affection and utility interdependence do not suffice, nor are they necessary. We may feel very deep affection for our pets (in particular our dogs), but we would not say that we adopt a participant's attitude towards them when we interact with them. On the other hand, we may have no affection for our enemy and still interact with him as guided by the adoption of a participant's attitude.[44]

There is presumably no completely satisfying criterion that could distinguish our attitudes towards other sentient beings from a true participant's attitude towards another being. Yet, we all make a corresponding distinction when we put ourselves into the shoes of another individual to emulate that individual's thought processes or to "empathize" with the other *in full*. To see the world through the eyes of another human is possible in ways that are not available when emulating the internal point of view of non-human sentient beings. At least this is what economist like Buchanan implicitly seems to accept when they assume that social interaction gains a special quality whenever humans interact with each other. Such individuals view each other as independent centers of decision making whose actions are not predicted according to regularities in the sense of empirical laws but by putting themselves, at least in a way, into others' shoes. We look at the decisions of others then as a product of rational or intentional decision making rather than as a result of decision emergence.[45]

In sum, in Buchanan type political economy, others are seen as *doing* something rather than as entities to whom something happens.[46] Humans form a community whose nature cannot be understood adequately without factoring in the ability to adopt an internal point of view.

[43] This is meant in the non-speciesist sense of "person".

[44] Reading Sun Tsu or Clausewitz, each of us is introduced to thinking about strategic interaction in terms of an unsympathetic participant's attitude which tries to get into the other actors mind; see on "getting into an other mind" in the setting of modern warfare, rather impressively Handel (2000) or earlier works of Michael Handel as well.

[45] See for a bounded rationality account of decision emergence, Güth (2000).

[46] From psychological experiments, we know that intentions are sometimes only attached to our actions ex post; we know that, realistically speaking, decisions do emerge, but we still ascribe to other individuals the ability to author their actions as we claim that same ability for ourselves.

Buchanan does not only accept the preceding as part of our common experience, he insists on taking it into account in forming theories of social interaction. At least in the reading proposed here, Buchanan-type or, as we may also call it, "subjectivist classical political economy" can best be understood as "economics with a participant's attitude."[47] – The Kantian elements of the Buchanan approach are very distinct and as such may deserve some further attention.[48]

2.3.2.1.2. Three conditions for Buchanan's non-science of economics

The preceding characterizes the most fundamental level on which the participant's attitude does play a role in Buchanan's theory of social interaction. It is reflected in a constraint imposed on *theory formation*. Let us call this methodological rule, for convenience, the *adequacy condition*.

> 1. **Adequacy condition**: The choices of human beings cannot be adequately understood in the same (parametrical) way that natural events are predicted or explained as occurring according to (probabilistic) laws. They must be understood in terms that go beyond externalist, "natural" terms.

The first condition is methodologically normative in that it alludes to an "adequate" understanding. We *should* for methodological reasons frame the world as inhabited by persons who have capacities that go beyond phenomena that can be understood exclusively in externalist terms. It may be that due to this normative methodological twist Buchanan has sometimes referred to that *requirement of theory formation* as an element of his contractarianism.[49] However, though it is clearly among the ingredients of Buchanan-type normative contractarianism, this kind of an assumption is a constitutive element in the process of world making and as such different from the substantive normative assumptions and consequences of contractarianism. Stipulating how we can and should "make the world" in representing it for inspection should be separated from the outright normative and evaluative assumptions characterizing behavior in the world and, thereby, from contractarianism more narrowly conceived.[50]

[47] It is opposed to the classical system-oriented and in this sense "objective" classical political economy; see on this distinction as showing up in corresponding classical and neo-classical notions of equilibrium Walsh and Gram (1980)

[48] Though Buchanan may have been "speaking prose" without knowing it, he has been Kantian.

[49] There has been such a proliferation of uses of the term "contractarianism" that is almost devoid of meaning now. However, the core notion is respect for the autonomy of other individuals as independent centers of decision making. We may not impose our own views and norms on others without their agreement even if it is for what we think is their moral good unless the others have wronged us in some way or other.

[50] Buchanan would have been understood more clearly had he made the methodological move of building respect norms into the constitution of his favored type of economics so to say.

There are basically two conditions that characterize Buchanan-type contractarian normative economics:

> 2. **Normative condition**: The primary aim of applied economic theory is not to provide counsel for those trying to get their partisan way but rather to suggest appropriate agreement-creating procedures or institutions to those who intend to proceed only by mutual agreement (but have to make adjustments in view in particular of the transaction costs of reaching agreement).
>
> 3. **Evaluative condition**: Personal value judgments about society and its institutions should be formulated to shape (procedural, moral)*preferences* that favor agreement-seeking institutions (as may be proposed in the economic policy advice formulated according to the normative condition 2).[51]

The first condition, the adequacy condition, imposes a normative constraint on non-normative political economy; with the second one, the normative condition, a normative restriction is placed on normative political economy; the third, the evaluative condition, is about forming value judgments in political economy (and its welfare economics part). – For clarification, it is helpful and worthwhile (for other reasons as well) to discuss the three conditions one after the other.

Extended discussion of adequacy condition: Buchanan's critical position towards what he thinks are abuses of statistics and econometrics may serve as a springboard for our discussion of the adequacy condition. Like Kant, Buchanan accepts statistics to some extent. When Kant emphasizes that there are some behavioral phenomenological regularities among humans, regardless of the fact that they can also be seen as what he regards as "noumenal" beings guided by reason rather than laws of nature, Buchanan agrees. The Kantian example that we can indeed anticipate the range of the number of marriages in a given year while being unable to predict any specific marriage drives home the point.

The preceding common sense observation raises, however, the obvious philosophical question about how predictability on a general or statistical level can go along with idiosyncratic ("uncaused") choice making on the individual level. Of course, statistical predictability could be explained as resulting from the laws of nature operating in the background. Yet, if that were the case, the unpredictability of specific marriages would only indicate insufficient knowledge of the specifics of each case rather than some principal restrictions on causation under natural "probabilistic" laws.

As has been stated before, for the dual perspectives approach to be meaningful, such ignorance of the laws governing behavior may be sufficient (and

[51] If the evaluations are captured in a personal welfare function, the Harsanyi, Arrow, Sen approach in the evaluative interpretation (see part 3) emerges.

the claim that there are no such laws unnecessary). Even if we were to agree that some omniscient observer who is informed about all the minute details relevant to individual action could fully predict and explain the acts of an individual, *we* are not omniscient. Quite to the contrary, though we know that some natural laws may influence individual behavior, we either do not know which laws those are or if we know which ones, we are not able to combine them such as to yield a full explanation or an external behavioral law-based prediction.

Had we all the information about initial conditions and behavioral laws such that we could predict individual actions in entirely non-teleological ways, it would not make sense anymore to rely on teleology in reconstructing the action situations of individual actors. Looking at interaction with others with a participant's attitude would not be meaningful. That window to the world would be closed presumably not only for the factual but also for the normative variants of the participant's attitude according to which we are normatively obliged to respect the other. However, since we are sufficiently ignorant, the window remains open. It does make sense still and will make sense for the foreseeable future to adopt a participant's attitude towards social interaction if we are to some extent ignorant of the laws governing behavior.

The lack of knowledge in forming social theories is obvious. There are no behavioral laws that would come anywhere close to natural behavioral laws. The conclusion to be drawn from this may well be that there is no social science in the narrow sense of the term. Relying on the logic of the situation and adopting the internal point of view in understanding other individuals may have to serve as a substitute for true behavioral law-like knowledge for times to come.

Since this implies that certain teleological elements will not be eliminated from explanations and predictions, we may have doubts about whether the enterprise may be called science in the full sense of that term. After all, one of the hallmarks of modern science is that it got rid of all the teleological elements so prominent in the religious and mystical explanations of former times. Laws and initial conditions rather than purposeful action towards aims explain what we observe and predict what we will observe.

According to the objective attitude of modern science, the cart of human events is pushed from the rear rather than drawn from the front. Teleology is out on most levels. On the level of evolution of complete biological species as well as on the level of the development of societies and their histories, it has been discarded. So, why not eliminate it on the level of the individual, too? Why not avoid speaking of the author's aim (the "telos") when explaining an action?

One response to this is: Individual actors as opposed to species and unorganized collectivities of individuals *do* have aims, ends, or values that they pursue. On the individual level, it may be argued that teleology is in a sense clearly “existent.”[52] It is there as an individual disposition of thinking along such lines, of desiring, wishing, demanding etc. Aims as present motives do exist and can serve as initial conditions in law-like statements that relate the presence of such a motive with certain consequences. Why should it be non-scientific to deal with these facts?

Avoiding the exaggerations of behaviorist methodologies, it seems indeed that accepting aims, ends, or values as springs of human action in our theories of social interaction would be scientific in any plausible sense of that term. It can be done, too, with an objective attitude, treating aims, ends, or values as theoretical terms in the relevant theories and their law-like regularities.[53]

This being said, we can put the Buchanan approach to what he regards as adequate political economy on the intellectual map more precisely now. To see how, we need to focus merely on four criteria. The two dimensions of the teleological or non-teleological nature of the underlying processes (i.e. whether the entities in the relevant realm are seen as pursuing ends, having intentions etc. or not) form two of these criteria. The dimensions of the objective or participant’s attitude as guiding the process of theory formation itself (i.e. whether understanding the situation from the internal point of view of the acting entity is crucial to the enterprise or not) form another two. From this, we get the following four combinations of characteristics:

Underlying process Theory formation	Teleological	Non-Teleological
Participant’s Attitude	1	2
Objective Attitude	3	4

Table 2.1 Perspectives on the world

Modern so-called natural science would correspond to case 4 (whatever else may characterize it). The Buchanan approach to economics would fall into square 1. If we choose to call human endeavors characterized by the combination 1 “science” as well, then Buchanan-type economics would qualify as such; otherwise, we would be dealing with a rational non-science. As we shall see below, classical game theory falls into category 1 as well, whereas evolutionary and strictly behavioral game theory should be seen as falling into category 4 (or

[52] Within a more ascriptivist framework, we might also draw attention to the fact that seeing others as having aims is what really matters, and in this sense even organizations may legitimately be seen as having aims. They can be treated as responsible actors. That the lines between organizations of individuals and individuals may become rather blurred once we take modern preference and utility notions seriously will be shown below in discussing basic game theory.

[53] Using these theoretical terms in explanations and predictions would not require that we be able to understand the ends, aims, or values with a participant’s attitude. However, that leaves us with the question of how a kind of theory formation, in which the participant’s point of view would play a role, fits into the picture.

perhaps 3, depending on how we classify the cognitive psychology account of the mental representations of aims).

Regardless of whether we choose to classify them as science or not, some of the cells in table 2.1 stand for proto-typical activities. As has already been seen, activities fulfilling the main-diagonal combinations 1 and 4 qualify. An approach to economics belonging to the humanities – more traditionally understood – would fall into 1 and a natural science approach to it into 4. An approach based on cognitive psychology would fall into square 3 because there the aims, ends, or values of individuals are taken into account as initial conditions of law-like statements that relate their presence to certain consequences, but this is done from an objective point of view.

Thus, only cell 2 of the preceding table 2.1 remains. It is hard to imagine what kind of approaches could conceivably fall into this category. Most of us would regard it as pathological if somebody in explaining the fall of a stone were to focus on how the stone might feel but not ascribe intentions to it.[54] However, some elements of the mythical view of the world may come close to 2. This clearly would not be seen as science but not as crazy either. Moreover, we may perhaps refrain from ascribing aims, ends, or values to a worm, yet nevertheless empathize with it in some way or other. This again might come close to category 2. In any event, the more relevant distinctions are located elsewhere in the table.

Since theories of cognitive psychology have been located in cell 3 and, in the ways teleology is dealt with, are not categorically distinct from sciences falling into category 4, we should focus on the distinction between categories 1 and 3 when it comes to the issue of classifying Buchanan-type political economy as science or non-science. Theories of cognitive psychology take into account that human beings as a matter of fact behave in teleological ways. The objects in their realm are "teleological" entities in some sense. Yet, their teleological behavior guided in particular by means-ends considerations need not be understood from an "internal" point of view. Therefore, cognitive psychology theories need not be teleological themselves, nor is it necessary to adopt a participant's point of view when describing or explaining behavior in terms of its factual purposes.

As theoretical terms, "purpose," "aim," "end," "value," and the like may show up in behavioral laws. They all may refer to "phenomena" that are not directly observable. These theoretical terms are ascribed according to certain observable criteria. In the natural sciences of non-animate nature, entities like "force," which are not directly observable either, do play a crucial role as well. It is the whole point of advanced science that it uses laws that relate theoretical entities to each other and these entities only indirectly to observations. In the same vein, there may be (probabilistic) laws linking theoretical entities of psychology that refer to

[54] Recall that the Persian prince beating up the water and imagining that an entity like Neptun is involved is acting on behalf of a wrong theory of how the world works but is not pathological.

certain forms of purposeful behavior without any reference to “an empathy-based understanding” of what is at stake. The laws are formulated and applied from a purely objective point of view.

As far as human beings and other animals are concerned, the underlying ascription of purposes and some, possibly rudimentary, form of rationality seems reasonable itself. To ascribe purposes in the same way to, say, an acorn which is treated as desiring or aspiring to become an oak tree seems to be much less plausible. Moreover, we must acknowledge that nobody has yet come up with superior or at least acceptable explanations of the growth of oak trees relying on laws or law-like relationships formulated in such terms as the aims, ends, and values of acorns.[55]

If we accept teleological phenomena of the kind that exist in the world itself, cell 3 seems perfectly acceptable. The assumption that aims etc. exist seems to contribute to our understanding of the world if captured in theoretical terms that are part of our explanations and predictions. This, contrary to a still strong behaviorist streak in economics, clearly suggests that cell 3 activities may qualify as science. These activities can be performed with an objective attitude. **In sum**, no participant’s attitude is needed to apply cognitive psychology.

In view of the preceding, one could try to interpret Buchanan’s rejection of certain forms of empirical economics as a plea for, say, cognitivistic psychology as being superior to behaviorism.[56] Though such criticism of behaviorism would not be off the mark, it is clearly not Buchanan’s intention to make merely that point. Rejection of endeavors of type 4 as were popular in Marxist and Macro-Economic quarters of the discipline can also hardly be the chief aim of Buchanan’s criticism of economics as a “natural science.” Buchanan was also not lamenting the fact that genuine economic empirical laws of any generality seem to still be missing (institutional experimental economics being a possible exception here[57]). In view of the complexity of our world and the limits of our theoretical and cognitive abilities, there may be limits to what science of the types 3 and 4 can accomplish. Buchanan cannot legitimately and in fact would not criticize

[55] Nobody has come up with useful explanations based on law-like statements yet that ascribe purposes to collective entities like species in biology or whole societies in sociology. The problem is not that such explanations and predictions of behavior could not be formulated in objective terms; they could. However, as compared to explanations that do not make use of theoretical ascriptions of purposes, they have not (yet) led to superior explanations. As far as organizational entities with central direction etc. are concerned, things may be different though.

[56] I am thinking of Gary Becker, Milton Friedman or generally Chicago-type misplaced behaviorism here.

[57] Falsifying the homo oeconomicus model for the 10,000th time is not helpful. If homo oeconomicus behavior is modeled such as to involve an empirical claim, then it is gone for good anyway. The generalization ranging over a class of market institutions (more precisely certain forms of auctions) stating that they in general will work in a certain way if appropriate incentives are provided is a generalization of rather high empirical respectability. Note, however, that it is not ranging over human behavior but rather institutions.

economics for not transcending such limits. So, what exactly is Buchanan driving at with his insistence that a purely objectivist approach to human interaction is inadequate?

It seems to me that the only possible rational reconstruction of intuitions and views as those expressed paradigmatically and forcefully by Buchanan must refer to category 1 of the preceding table. Within the Buchanan setting, the participant's attitude of the factual condition 1 implies approaching another being as an "author of acts." The other individual is framed not merely as a part of nature from which certain events, or acts, emerge. Self and other are each conceived of as a "doer" of actions. And, this is not a kind of second thought of the theoretician but an integral element of the theory itself.

At the root of the Buchanan concept of economic theory as performed with a non-normative participant's attitude lies the ascription of the *same* rationality *symmetrically* ascribed to all individuals. By this assumption, which is partly *contrary to fact*, all individuals are *framed* as participants of an interaction or as members of a community of (in that sense) rational beings. They are "equals" in this sense. Since, as Buchanan is well aware, this ascription of symmetric and rather perfect rationality, which is typical of political economy, cannot be justified on factual or empirical grounds (individuals vary in their rational capacities that are, across the board, far from unbounded or ideal), he introduces it as a *constitutive* element of his particular version of theory formation.[58]

The teleological faculty to pursue aims, ends, or values is assumed to be the same for each and every individual. Though Buchanan is not very precise about it, making this assumption cannot merely be treated as a convenient approximation of the facts. All the participants of interaction must be seen as symmetrically equal by the *theory* if it is to be political economy in the Buchanan sense. Nevertheless, Buchanan does not want to introduce equal respect as a substantive norm. He does not appeal to fellow theoreticians to respect each other and treat each other as symmetrically rational and as equals in that sense and to treat all other individuals that way. He rather makes it part of a specific way of theorizing. One cannot participate in that way of world making without adopting a participant's attitude to all others. All are treated as if they were rational in the same way and deserving to be treated as independent centers of decision making according to the rules of his game of theorizing. You cannot play that specific game of political economy without accepting the fundamental "rule of recognition of the game".

In sum, it is impossible to play the game of Buchanan-type political economy without conceiving of individuals as rational equals in theory. This is required if the theory is to qualify as adequate political economy in Buchanan's sense.

[58] This is why I discussed all the preceding as concerning theory formation or the methodological thesis 1.

According to the preceding argument, Buchanan's premise is an empiricist version of a Kantian transcendental supposition. In that sense – and partly without noticing it – Buchanan clearly endorses a Kantian methodology. Quite in line with this empiricist (Strawsonian) Kantianism, Buchanan requires that, in what he regards as sound economic theory, all individuals are treated as actors who "make" choices (i.e. their choices do not "happen" to them).[59]

Within this framework, we understand the individual choices from the internal point of view of each participant of any interaction taken separately. Each such participant tries to understand the interaction by understanding how others try to do the same. In doing so, actors "explain" and "predict" in terms of "reasoning about reasoning" what happens in social interaction from their various participant's points of view. They form a community of rational beings engaged in the enterprise of exploring how and, in particular, under what kinds of institutions they should live in view of the rational pursuit of their own given aims, ends, or values if they do so under the side constraint of ascribing the same rationality symmetrically to each participant.

As we shall see, the preceding is also at the root of classical game theory. Therefore Buchanan had a very good reason for endorsing classical game theory (see Buchanan (2001), Buchanan et al. (2001)). However, before going into this, a few remarks on the much more straightforward normative and evaluative conditions may be in order.

Extended discussion of normative and evaluative condition: Buchanan implicitly insists on a kind of Kantian foundation not only of explanatory but also of normative economics. Economic counsel as based on economic theory *should* be such as to express norms of inter-personal respect.[60] Suggesting institutional arrangements that allow individuals to interact with each other respectfully rather than showing them how they can impose their will on others (either by manipulation and/or force) is the aim.

A hard-nosed economist without Kantian leanings would have to say here that agreement with another individual should be sought if this is instrumental for reaching the actor's given aims, ends, or values. (This kind of instrumental agreement-seeking would be within the limits of the Robbins enterprise). Whenever seeking agreement is not instrumental to the aims of the actor because he can (all things considered) reach his own aims more efficiently by simply imposing his will on others (exerting an externality on the other), then – according

[59] The aspects of the world that are beyond the making of individual choices comprise the rules of interaction.

[60] Here Buchanan like Strawson goes beyond the mere world making aspect and into some stronger view of implied norms. As in the case of Strawson, I am rather reluctant to follow if the two go that extra mile, so to say.

to this variant of economic means-ends rationality – he should do so.[61] The implicit assumptions of symmetrical abilities and rights may be normatively desirable from some point of view, but the actors themselves may well deviate from symmetry.

Economic theory, remaining silent on ends, must be willing to serve as an all-purpose machine like technical engineering. This machinery can be put to both good and bad uses (as evaluated from some normative stance or other), but as such it is not directed towards a constrained set of specific ends.

In contrast, it is constitutive for the Buchanan variant of "political economy" that it does not push the means-ends framework that far. If economics deals at all with maximization under constraints, then some of the constraints are of a very specific normative form. They are side constraints of mutual respect of a more or less Kantian nature. Yet, whereas Rutledge Vining in his essay on the state of economic science in America explicitly insisted that "true" economists subscribe to "the moral principle that no individual should treat another simply as means to an end" (Vining (1956), 18), Buchanan is not that explicitly Kantian. He hesitates to impose a Kantian imperative from the outside. Like other modern contractarians (with Kantian roots), he tries to get around this by referring to agreement. Accordingly, he insists that the political economist's proper policy advice should focus on the elicitation of agreement. The economist as a counselor should tell his addressees neither what they should do according to *his* values nor what they should do to get their way according to *their* values. He should tell them how to seek agreement, thinks Buchanan.

In sum, the proper economic counselor should advise citizens on how they themselves could agree on what should be done if they were themselves following broadly Kantian principles.[62] The advisor *restrains himself to counseling that is in line with the aims of agreement seekers*, but he acknowledges that the advisees in the last resort must themselves and as a matter of fact accept the advice (after all agreement is his basic normative premise).

Closely related to the normative condition but distinct from it is the evaluative condition. Compliance with the evaluative condition leads to a suggestion of how we should *think* and argue about normative alternatives in public discourse (or deliberation). We should do so in a way that incorporates ideals of agreement-seeking.

What exactly the evaluative condition aims at beyond this ideal needs some further clarification. It clearly does not boil down merely to the ideal of construing institutions that give mutual agreement the widest possible scope. It does not focus

[61] One could also say that showing respect for others does not serve as a constraint of the pursuit of own aims.

[62] In an institutional interpretation, this leads to forming procedural proposals in view of normative condition 2.

on the properties of the institutions and rules on which we form an opinion but rather on how we form the opinion itself. Understanding what is at stake here will be much easier at a later stage of the argument. Therefore, let it for the time being suffice to lump the normative and evaluative conditions together and make a kind of mental note that we will have to sort this out somewhat further down the road (in volume 2 in fact).[63]

For our present purposes of understanding the philosophical underpinnings of world making in economics or of economic modeling in general (rather than the special type of modeling relevant to normative economics and, in particular, normative welfare economics), the more interesting questions concern how we would reconstruct the adequacy condition more precisely. It may still seem to be rather unclear how an approach based on this premise is to be distinguished from an approach to social interaction based on cognitive psychology. Here the distinction between classical and other forms of game theory (e.g. behavioral game theory) is most helpful.[64] The tacit and often forgotten historical presuppositions of classical game theory may indeed serve as a further illustration and support for Buchanan's views. From a philosophy and economics point of view, they are of the greatest interest of and in themselves anyway.

2.3.2.2. Morgenstern & Co[65]

Participants in a game, as interpreted in classical game theory, know that there are others. They know that, and they know that others know that they know that and so on. The basic self-referential character of human knowledge was mentioned above but is obvious to us from everyday life experience as well. We know ourselves that we know, and we know of others that they know that we know that they know and so on. On whichever level such "common knowledge" applies, we are members of what may be called a "**knowledge community.**"[66]

Game theory has pushed common sense intuitions about knowledge in social interactive situations to its limit. In doing so, it has become a basically non-behavioral theory of "**reasoning about knowledge**" in a knowledge community. Classical game theorists are *not* dealing with reasoning in terms of cognitive psychology. They do not address the reasoning processes of boundedly (i.e.

[63] Condition three is located on the level on which the Rawlsian contractarian analysis also has to be located. Buchanan puts himself firmly in the Rawlsian camp here as well as on many other occasions. Following Rawls' emphasis on the central role of the sense of justice, Buchanan's endorsement of Rawls should be interpreted as an endorsement of a specific form of "value formation" liberalism or, if you will, "deliberational liberalism," which will be discussed later. At the same time one of the strongest criticisms of mixing evaluative liberalism with institutional or political liberalism is Buchanan (1975/1996).

[64] Buchanan is perfectly right in thinking of classical game theory as his close ally in these matters.

[65] This section should systematically be located here. It need not be read at this stage by a reader not yet familiar with elementary game theory. Such a reader may come back here later.

[66] See from a philosophical point of view Lewis (1969)

imperfectly) rational individuals. They rather try to find out what ideally rational individuals, who could reason without bounds or limits, would infer from their knowledge of the interaction, assuming that other symmetrically intelligent beings would command the same knowledge.[67]

Much of what is going on here can be illustrated more concretely by simple examples. For instance, the self-referential character of knowledge can lead to phenomena like self-defeating predictions or self-fulfilling prophecies. If it were predicted that everybody is going to travel on the first Saturday of the first holiday weekend in the summer, then perhaps most people would back out. Therefore, if the prediction is widely believed, it will be self-defeating. However, if people anticipate that it will be widely believed, they should ignore it. If they ignore the prediction, however, this in turn will provide a very good reason not to ignore it, since, if people do not respond to the prediction, it may well hold true.

Without going over the details of the well-known case of the self-fulfilling prediction of a run on a bank, we can simply note that in addition to self-defeating prophecies, there are also self-fulfilling ones. In both cases, knowledge of what may in the widest sense be classified as "theories" about the world influences the world itself. Among theories that can exert this influence, only some will be self-corroborating. To that effect, they must provide correct descriptions of what will happen after their own influence has been factored in.

A theory that has an influence on what is described by the theory itself can be valid over the long haul only if it does not have self-defeating properties. It must be "absorbable" by all those who are able to understand the theory without altering their behavior – (at least knowledge of behavior may not alter that behavior in the limit).[68] Modern economists, even if they are not primarily working as game theorists, imply such an absorption condition by assuming so-called rational expectations, i.e. the assumption that the world indeed is as the rational choice theory assumes it to be if rational choice theory is true and "observed" by all.

It was this line of thought that had fascinated Oskar Morgenstern for quite a while before he joined forces with John von Neumann in developing modern game theory. Even though the introduction of the technical concept of the Cournot-Nash equilibrium[69] into classical modern game theory is rightly attributed to John Nash (Nash (1951)), one should be aware of the concept of theory absorption in this context as well. For, neglecting it, we are hard put to see why the focus on equilibrium is meaningful at all (unless we would turn to adaptive modeling as in biological selection or learning theory).

[67] Again, see for an extreme version of this Fagin et al. (1995)).

[68] On theory absorption, see Morgenstern and Schwödiauer (1976), Dacey (1976, (1981), and with an eye on limited rationality, Güth and Kliemt (2000b)

[69] The informal discussion in the preceding chapter illustrates this equilibrium concept already and it will be further illustrated below.

Vice versa the absorption condition is basically an equilibrium notion and implies equilibrium play at least for games that do have definite solutions (see close to this Jacobsen (1996)). Only those theories that advise players to play equilibrium strategies can be self-corroborating or at least not self-defeating. Only these strategies can be common knowledge and be adopted by all individuals without a rational incentive to deviate from the theoretical recommendation. The argument is obvious:

> An **equilibrium of strategic plans** of n actors, n>1, is such that no actor can do better against the equilibrium plans of the n-1 other actors by planning otherwise. A theory which would recommend non-equilibrium plans would provide a recommendation such that the addressee of that recommendation would have an incentive not to take the advice. That individual in considering the situation would not be in "reflective equilibrium"70. Therefore, such a theory could not be commonly known and be commonly rationally followed. It is not in the ideal sense absorbable among fully rational individuals.

If a theory suggesting non-equilibrium behavior were assumed to be known by all concerned, it would not lead to a coherent setting, in which the rational actors know the theory in full. Conversely, however, the theory of rational play itself can become a kind of "signal". The common knowledge of the theory can successfully coordinate action.[71] If everybody knows that everybody knows that everybody knows the signal, nobody would have good reason to consider deviating from what the theory as signal may recommend as rational action. In full knowledge of the theory, all are in a "reflective equilibrium".

If the theory proposes a single strategy to every actor as her equilibrium strategy, then, since it is an equilibrium strategy that is recommended, nobody has an incentive to act otherwise. If, to repeat the argument, a non-equilibrium strategy were recommended to a fully rational actor, then that actor would have an incentive to act otherwise. We know in that sense that only theories that recommend equilibria can be fully absorbed into the reflective equilibria of all concerned.

Yet, in view of the multiplicity of equilibria, to recommend strategies as belonging to *some* equilibrium is not sufficient for recommending a specific choice. The ever-present equilibrium selection problem emerges on the level of theories as well. Several theories, even when selecting a single equilibrium in each and every case, could still recommend different equilibria. To solve that problem, we have two options. We either have to strengthen the rationality concept by

[70] On this in some detail see vol. 2, on reflective equilibrium shortly, Hahn (1998).

[71] Very much in the spirit of coordinating action in correlated equilibrium theory of the Aumann type, see Aumann (1987)

relying on some a priori criteria or other to such an extent that for all games of a class of games under scrutiny one and only one equilibrium would be selected as the single rational choice.[72] And, in that case, the theory of rational choice for that class of games and the choices recommended by the theory would support themselves within a reflective equilibrium. Or, we would have to accept that the meaning of rationality itself is to some extent fixed by a contingent convention. A convention as a prevailing practice singles out which one of several theories is to be applied in selecting the equilibrium. The common knowledge conditions of a convention – as conventionally defined according to Lewis (see Lewis (1969)) – then guarantee that nobody has an incentive to deviate from theoretically "signaled" "rational" play. However, *the factual prevalence in a practice* is decisive for fixing the normative content of rationality here.

Except for the aforementioned reliance on an established practice, the deliberative nature of the whole process should be obvious. We are talking about thought processes that could conceivably be going on among rational individuals in the idealized setting of a knowledge community of perfectly rational beings. These individuals all participate in the knowledge of the same theory. In fact, what is rational is explicated by the theory itself, and the equilibrium concept is applied to that explication – if only within the broader setting of searching for a wide reflective equilibrium.[73]

Much more could and presumably should be said here about theory absorption in a knowledge community; however, what has been said should suffice to embed classical game theory in a broader concept of reasoning about knowledge. Again, symmetry and other assumptions about the rationality of participants in games make the argument somewhat precarious from an empirical point of view. Among the several idealizations, that of an unlimited capacity to reason may seem almost outrageously unrealistic. However, this is not the only rather daring assumption. The ideal-type players do not only reason about the world, they anticipate in their reasoning that others will reason and anticipate in turn their own reasoning etc. Ideally rational players would, as "planners", also know about the existence of two worlds. They can anticipate that besides planning on how to play, there will be actual play. The "homo noumenon" of the Kantian rational world will anticipate that he is, in executing his plans, a "homo phaenomenon" who will conceivably deviate from the plans. And, the anticipation of the imperfect execution of plans by a rational planner may induce the planner to take into account imperfections in his perfect plans. Considering, in his ideally rational planning, an imperfect execution of plans will not necessarily lead him to

[72] See for a heroic and brilliant, though presumably not fully successful, effort to that effect Harsanyi and Selten (1988)

[73] See again volume 2 on reflective equilibrium; for a short account of reflective equilibrium at this stage of the argument, see Hahn Hahn (1998) for a more dynamic approach to rational deliberation see, Skyrms (1990).

endorse Murphy's law, but he may take into account "little trembles," small errors and the like.[74]

More on this and other ways to bridge the gap between the two worlds or perspectives sketched in this chapter will have to wait until later. First, the traditional rational choice perspective will be introduced in the next two chapters, which will focus on fairly standard elementary models of action (chapter 3) as well as on elementary models of interaction (chapter 4).

Though it will be shown that the seemingly innocuous concept of preferences is ambiguous concerning the dual world conception discussed in the present chapter, the traditional rational choice analysis should be clearly classified as "reasoning about knowledge" in view of "given preferences" and "other rules of the game of life." To the means of modeling aspects of this game I will turn next.

[74] As readers familiar with the trembling hand concepts introduced by Reinhard Selten into the technical game theoretic discussion will notice this is a somewhat unfamiliar philosophical underpinning for the concept in terms of the two worlds framing.

Part 2: Elementary concepts and models of rational choice analysis

After sailing some deeper philosophical waters in the previous part I, the next two chapters on philosophy and economics are rather straightforward. They present introductions of standard rational choice modeling techniques, though with a specific philosophical touch. They can be read as an introduction into elementary modeling for the "uninitiated" and as an introduction to specific philosophical and theoretical interpretations of the language of models for those who know the basics but are interested in an account of "what it all means." Whether the following indeed presents "what it all means," I am not in a position to judge. However, I will at least try to make some sense of rational choice in a coherent story.

3 Models of Action

According to the more conventional views of the matter, practical rationality must basically be understood in terms of consistency. Put simply, this means that if person r prefers alternative x to alternative y and y to alternative z – noted for short as x P_r y P_r z – then she should also prefer x to z. If she, nevertheless, prefers z to x, we would get x P_r y P_r z P_r x. This may be seen as a formal inconsistency implying x P_r x, which would mean that somebody would strictly prefer an alternative to itself.

It may also be seen as pragmatically incoherent. If a person r endowed with an alternative w whenever she prefers v to w would be willing to pay a little surcharge to switch to v she could be trapped into a pragmatically vicious circle. To see this, first assume that person r with preferences P_r initially possesses z. Somebody can take z and sell her y for a little surcharge since y P_r z. Then r possesses y, and, according to the same argument, y can be exchanged for x for a little extra charge, since x P_r y. Now possessing x, she gets back z for a little extra charge, while handing in x, because, by assumption z P_r x. Finally, r is in possession of z, while the trader holds x, and the whole process can start over again since y P_r z.

As in some of M.C. Escher's paintings, things descend all the time only to end up on the same level from which they started.[75] An individual showing such behavior with preference structures like these driving it is seen as violating the minimal standards of rationality. That rationality requires consistency is quite plausible, but in mental representations of the action situation, a minimally rational choice maker should not only be consistent, she should also be able to make the distinction between what is and what is not subject to her interventions.

In the subsequent presentation of the rational choice approach, due emphasis will be put on the second aspect of what may be called minimum rationality. From now on, **rational choice theory (RCT)** will be used to refer to substantive assumptions or theses about rational choice making and rational action in general. For example, if somebody claims that a rational individual, if presented with the choice between two alternatives that offer monetary gains, would always prefer the higher monetary gain to any lower one, then this is a substantive assumption about what rational individuals do. Yet, we must distinguish **rational choice modeling (RCM)** from substantive rational choice theory. RCM is a mere

[75] For other impressive examples see Hofstadter (1979).

language that allows us to express certain substantive assumptions about human action. This language comes with certain rules of interpretation, but it does, as we shall see, not imply a specific theory of rational behavior of *personal actors*.

The main advantage of using RCM is its precision and not any transportation of particular empirical content. RCM merely induces us to make the RCT assumptions on which we want to rely explicitly (see Güth and Kliemt (2007) on RCT and RCM in more detail). In particular, it forces us to model all constraints on choice explicitly because RCM as a language uses signs that, according to the rules, must be interpreted to signal unconstrained choice *unless* the constraint is explicitly modeled.

By choosing to demand that constraints be stated explicitly, nothing is said about the presence or absence of constraints. We can express quite different substantive views about the presence or the absence of constraints on opportunity seeking behavior in the language of RCM. There is, however, a traditional association of RCM with the classical substantive RCT of Thomas Hobbes to which I turn first.

3.1 The Hobbesian roots of rational choice

Economists tend to look at Adam Smith as the founding father of their discipline. This seems right with respect to institutional issues. However, the basic behavioral model of homo oeconomicus and its use as the basic building block of a general social theory must be attributed to Hobbes.[76] That does not mean that Hobbes restricted his behavioral theory to the homo oeconomicus model. Though using the model for a first bold assault on political and social theory, Hobbes himself built into his arguments certain precautionary elements. The rational choice makers he envisioned were not fully unconstrained and unbiased opportunists.[77] Together with his still medieval assumption that there is a natural duty to preserve one's own life – not only to do whatever seems to aid in reaching *any* given aim, end, or value – the Hobbesian view of the world does not fully coincide with the stereotype of the rational Hobbesian egoist. The precautions (constraints) in Hobbes' own account of social and political order amount to deviations from a strict homo oeconomicus explanation and the corresponding means-ends

[76] Of course, there were others like Macchiavelli or, for that matter, some thinkers of antiquity, in whose work we find bits and pieces of what should later come to be known as the economic approach to human behavior. However, it is only with Hobbes that the approach is systematically and generally applied across the board of human behavior.

[77] For instance, the idea that loss avoidance dominates the seeking of gains played a central role for Hobbes. This foreshadows Kahneman and Tversky, if you will, see Kahneman and Tversky (1984).

justification of human (social) behavior.[78] Yet, in later theory formation, the Hobbesian cautionary remarks did not prevail. Rather, a streamlined reading of the original texts dominated the discussion in early modern times and in the subsequent history of modern social thought.

Spinoza's strict **homo oeconomicus** view of the workings of law and social order expresses the conventional reading of Hobbes most succinctly and instructively (Spinoza (1670/1951), 203–204):

> Now it is a universal law of human nature that no one ever neglects anything which he judges to be good, except with the hope of gaining a greater good, or from the fear of a greater evil; nor does anyone endure an evil except for the sake of avoiding a greater evil, or gaining a greater good. That is, everyone will, of two goods, choose that which he thinks the greatest; and of two evils, that which he thinks the least. I say advisedly that which he thinks the greatest or the least, for it does not necessarily follow that he judges right. This law is so deeply implanted in the human mind that it ought to be counted among the eternal truths and axioms.
>
> As a necessary consequence of the principle just enunciated, no one can honestly forego the right which he has over all things, and in general no one will abide by his promises, unless under the fear of a greater evil, or the hope of a greater good…Hence though men make promises with all the appearances of good faith, and agree that they will keep to their engagement, no one can absolutely rely on another man's promise unless there is something behind it. Everyone has by nature a right to act deceitfully, and to break his compacts, unless he be restrained by the hope of some greater good, or the fear of some greater evil.

In good American parlance, we get an even more succinct version of the same substantive RCT by Mark Twain's *Huckleberry Finn* (reflecting on his own response to the slave chasers):

> Well, then, says I, what's the use of you learning to do right when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same? I was stuck. I couldn't answer that. So I reckoned I wouldn't bother no more about it, but after this always do whichever come handiest at the time.

[78] This is clearly expressed when Hobbes insists that somebody who received what was due to him according to a bilateral agreement does not go against the precepts of reason if he responds in kind. If Crusoe delivers his apples, then, says Hobbes, "it is not against reason" that Friday, though he has already received his apples would still deliver the promised oranges anyway. Hobbes deems it not unreasonable even though there are no additional future apples to be gained by that act; see Hobbes (1651/1968), §15, full citation below in 4.1.2.2 as well.

Even though Spinoza and Huckleberry seem as self-assured as many economists today of their theory of what is substantively rational, this theory of unconstrained rational choice is open to very serious doubts.[79] Above all, it is unclear how social order with the constraints it imposes on behavior is at all possible if all individuals act like "uncommitted" homines oeconomici all the time.

The so-called "**Hobbesian problem of social order**"[80] already emerges on the basis of what may be called "minimal rationality" (see section 3.3). As will be seen, even this minimal concept (as it may be termed in a certain sense) is too strong to allow for a convincing account of the possibility of social order. However, before we can turn to such issues, we must address another fundamental question, namely that of whether or not the standard economic explanations of human behavior in terms of given preferences make sense within the economic means-ends framework. Though most economists would claim that they obviously do, a closer philosophical and methodological look renders this tenet rather precarious.

3.2 "Given" preferences and explanations

There are at least two categorically distinct strands of rational choice modeling. On the one hand, rational choice making is understood from the internal point of view of the choice maker.[81] The explanation of choices relies on reasons for choice making. In that respect, it adopts the point of view of a participant of interaction. For instance, the profit maximizing behavior of a company is explained by the fact that the board members of the company try to maximize profits of the company since that increases their own earnings. They do what they do because they have certain subjectively conceived "reasons" for it. On the other hand, rational choice making is understood in terms of the "substantive" advantageousness of the choices made. In this account, there need not be a process of rational deliberation leading to what is also classified as rational choice making.

[79] When some years ago I introduced the Spinoza quotation as a target of my criticism of the homo oeconomicus model at a conference at the Humboldt University in Berlin, I found out that this was not far-fetched. The prominent German sociologist Karl-Dieter Opp as a co-panelist immediately asked me for the full reference since he, ironically, thought that Spinoza had express wonderfully how, according to his own rational choice perspective, the world lies. Opp may be a very consequent rational choice sociologist in this regard, but he is certainly not the only one who thinks along such lines even today.

[80] The social order problem (see Parsons 1968)) has always been on the minds of the most eminent social theorists like David Hume (see Hume (1739/1978), book III) and other British Moralists (see Raphael (1969)) – though not under its now popular name.

[81] For readers who read the chapters of this book not consecutively: The extended treatment in chapter 2 provides additional information on this, but it is not necessary to read chapter 2 before reading on.

Forms of behavior are deemed rational in view of the fact that they fulfill substantive, external criteria of rationality. They can be caused in any way, and in particular they need not be rooted in reasons internal to the mental processes of the choice maker to be so classified. For instance, the board of directors of a company is enthusiastic about the quality of the products of that company, for it abandoned profit seeking and adopted a "quality first" stance. Such are their reasons for action. These reasons are clearly contrary to conscious "profit-seeking." However, "objectively" speaking, their behavior turns out to be the profit-maximizing strategy in competition with other companies.[82]

To classify acts as rational in the second, substantive sense, no appropriate rational deliberations need to precede them. To classify acts as rational in the first sense, they need not be objectively successful. They must merely be triggered in the right manner subjectively. This leads, in a way, to theory formation from a participant's point of view. Later on, the objective window will be opened up as well, and then the two perspectives will be related to each other. Initially, the focus here is on that first subjective understanding of rational choice making.

3.2.1 Preferences as reasons for action or not

Normative as well as explanatory economics of the Robbins' "means to given ends approach" treats aims, ends, or values as exogenously determined. Starting from given aims, ends, or values in the eyes of most economists amounts to assuming preferences as "given". Constraints are "given," too. Methodologically, it seems natural to endorse the following norm: Choice making is to be explained in terms of rational behavior guided by (exogenously) "given" preferences that are pursued under exogenously "fixed" constraints. Rational behavior amounts to "**maximization under constraints,**" i.e. to seeking those feasible alternatives that lead to consequences that are ranked as highly as possible in the exogenously determined preference ranking.

To assess the plausibility of such an approach to explaining human planning and choice behavior, it should be recalled that what is "given" is in all likelihood not entirely explicit or transparent to the individual who is allegedly endowed with the "given." In fact, the individual herself might not exactly know how she should rank states of the world in view of her multidimensional and often complex evaluative attitudes based on her "given" aims, ends, or values. She might, and as a rule will, feel uneasy about trade-offs between several dimensions of value.

For instance, think of a person who intends to buy a used car. She considers cars in the price range, say, between $10,000 and $15,000. Some cars look nicer but are more expensive and are at the same time gas-guzzlers. Other cars are more

[82] This line of thought is, of course, quite in line with the proverbial "honesty is the best policy" phrase.

modestly priced but have dull looks and moderate gas consumption. Even if in the end a choice must be made, it is not obvious which. How, precisely, the choice making is related to the given aims, ends, or values of the choice maker is open since the relevant dimensions of value are so many in real life. This seems, in a way, trivial and commonsensical, which certainly should not count as a criticism. However, note also that besides the ends, aims, or values, which somehow may co-determine preferences and, therefore, clearly may show up in means to given ends explanations of behavior, the preferences need not play a role in the explanation of the behavior. The individuals will not inquire what the given preference orders are and then choose on that basis. Individual choice makers themselves, from their internal point of view, have not much use for preferences. They will *directly* rely on their aims, ends, or values in their cognitive processes. The preferences as allegedly "given" determinants of choice seem to be intervening variables at best. They are formed by the deliberation process based on the aims, ends, or values and are not given beforehand.

In sum, even though the premise that the aims, ends, or values of the choice maker are given may be true, it may still be misleading to assume that *therefore* preferences as rankings of evaluated states of the world are "given" as well. It may well be the case that not only choices but also the preferences underlying those choices are "made" (or must be made). A ranking among several choice alternatives or their consequences must possibly first be *construed* from the aims, ends, or values of the choice maker (if it is not construed from the choices in a so-called revealed-preference approach, which is, however, fully appropriate only in an externalist perspective, to which we turn later).

The preceding would not pose a serious problem if preference rankings were a context-invariant simple function of the given aims, ends, or values. However, more often than not preferences and choices will *not* at all be a *function* of the aims, ends, or values of the choice maker. If they were a function, then for any constellation of such aims, ends, or values, we could say what the preferences would be independent of the context. However, construing preferences, forming rankings according to several aims, ends, or values may and will as a rule be strongly context-dependent. Though some invariance across situations would be desirable, it may well be that the given aims, ends, or values will lead to rather different consequences depending on the context in which they prevail.[83] Situational aspects and human judgment come into play, and choice making becomes less "predictable."

In sum, due to the complexity of the human mental processes that "intervene," it will always occur that for the same constellation of aims, ends, or values of the choice maker, different preferences will emerge. But this lack of invariance across contexts will destroy the fundamental property of a functional

[83] It may also be true that there is no easy way to separate the context from aims, ends, or values.

relation. We cannot treat preferences as functions of some more fundamental variables unless we include context.

If we speak of “given preferences,” it cannot mean that preferences are given independently of the situation in which they allegedly guide planning, behavior, or choices. However, if preferences are not invariant across situations. if they rather must be construed from the given aims, ends, or values of actors by the actors themselves in the action context, would it not be adequate, even necessary, to model the process of construction along with forming the rational choice model of the action situation?[84] However, then the basic abstractions of rational choice theory or what gives them the analytical power to cut through the complex maze of mental processes seems to vanish in thin air.

Still, even if we had good reason to think of preferences as a function of both the given aims, ends, or values of the decision makers, on the one hand, and the decision situation, on the other hand, it seems doubtful whether the values of that function (i.e. the preference rankings) would be helpful in analyzing the mental processes of players. Why should the function or its values show up in those mental processes and how would knowledge of its values (the preference orders) help to make choices? At least the choice maker herself will – as a rule – not have use for preferences in her deliberations. However, if the choice maker does not rely on a representation of her preferences, it would seem to rule out that our theory of choice making can be based on preferences.

In sum, if we intend to form a theory of rational choice making that models what is on the minds of the choice makers and if preferences are not on their minds, it is hard to see how we can nevertheless stick to a preference-based rational choice approach.

Of course, it is not necessarily true that explanations of human behavior must be formulated in terms of mental processes in an internalist way. Explanations need not be based on the mental representations or models of the action situation as endorsed by the actors themselves. Nevertheless *standard* rational choice modeling, though often muddying the waters by alluding to forces other than mental processes, does make sense only as an idealized theory of *deliberation* of choice makers.

The rationality of participants of interaction as modeled in standard rational choice theory profoundly depends on cognitive processes antecedent to their choice making. Yet, then, only if preference rankings somehow show up in the deliberation process of individuals – and their representations of other individuals – would it make sense to explain planning or choice making behavior in terms of (given) preferences. If preference rankings did not show up in the mental processes of the decision makers, how could for instance a cognitive psychology

[84] This would make game modeling rather complicated.

explanation or, for that matter, a standard economic explanation in terms of strategic planning be based on preferences?

If we intend to explain matters in terms of what is on the actors' minds, then preferences must be on their minds. That preferences can hardly play such a role becomes particularly clear if we bring in the revealed preference concept so popular among economists. For the sake of specificity, assume that given the choice between x and y, person A repeatedly chooses y over x. Let us assume too that this is sufficient to conclude that A prefers y to x. In this case, preference orders or lists in which the higher ranking alternatives show up before or above the lower ranking ones would simply "stenographically" make note of the choices made and to be made. However, it is not assumed – this is the whole point of using revealed preferences – that such lists would show up in the deliberations of choice makers revealing those preferences.

If there were a role for preferences in an internalist account of choice making, they could not simply represent choices made. To reiterate, preference orders would have to be among the "causal factors" influencing choice making.[85] If the choice making entities are human actors and if we intend to explain their behavior in terms of what is *on their minds*, then the explanatory laws should be laws of cognitive psychology. However, within most cognitive psychology contexts, it seems a gross distortion to assume that in the mental processes of choice makers preference orders or their representations play a role.

In sum, human beings do not plan on the basis of preferences. In their deliberations and planning, they do not start from given preferences, neither of their own nor of other individuals' making (though they may be aware of the "given" aims, ends, or values of self and others).

Even economists as a rule do not deny that psychological processes do play a role in human planning and in determining the human choice making behavior consequent upon it. Economists know that deliberation is a mental process subject to laws including those explored in cognitive psychology. Many economists presumably are aware also that in the mental processes of human beings preference rankings as such do not play a role, and even those economists who are not aware of this may presumably be induced to acknowledge that almost no human individual would imagine her or his own preference ranking or that of other individuals and then to act upon that information.

To the extent that they conceive their own theories as models of rational *reasoning* that are somehow represented in the mental processes of the reasoning actors themselves, economists who insist on a preference-based account of human behavior are confronted with a dilemma: *Preferences must play a role in the deliberations or rational planning of choice makers according to economic*

[85] Of course, the reasons on the mind of the actors are causes, too.

modeling; yet, at the same time, preferences do not play a role in the deliberations of actual people.

The only way around the dilemma amounts to denying that economic choice behavior must be explained in terms of laws of the human psyche. Accordingly, economists must and do intend to explain human planning and choice making behavior in terms other than cognitive psychology. More often than not they think that relying on preferences as action representations from an external point of view can do that trick. Initially it might actually do so if they stuck strictly to externalist explanations and eliminated from economic explanations the knowledge, intentions and expectations of actors. Yet many of the economists who insist that externalist explanations will indeed do, at the same time, make rather extravagant internalist assumptions about the mental capacities and reasoning of choice makers.

In particular, choice makers are assumed to know all the preference orders of all the actors. From this, which goes practically unnoticed by the profession, a somewhat strange mixture of internalist and externalist perspectives emerges. Such a Janus-headed approach is full of internal tensions and incoherencies, and it prevents economists form moving closer to a reflective equilibrium on rational choice theories. However, economists have a reason for their rather stubborn behavior, which, for that matter, turns out to be not without plausibility within their specific way of world making.

3.2.2 Given preferences as shielding economics from psychology

Starting from given preferences serves the purpose of eliminating the need to use (cognitive) psychology as a basic explanatory theory of economic behavior. Relying on given preferences that *represent choices* without any references to the causal processes underlying the choice making, economists intend to explain human behavior and human planning without psychology. Using the all-purpose weapon of given preferences, they substitute psychological accounts of mental processes and behavior by "as if" accounts. Behavior is rational to the extent that it is "***as if***" it were maximizing behavior.

Though individuals do not consciously maximize, they behave "*as if maximizing a given objective function determined according to given preferences.*" They choose within given constraints the alternative that leads to overall results that are, according to the given preference order of the choice makers, ranked highest among the attainable results. Looking at the explanatory problem this way, economists believe that they can leave the cognitive processes in the dark. Using the preference order as a shorthand description of the individual, they try to uphold the traditional *explanatory* program of neo-classical economics[86] because they believe that their explanatory program will be lost unless they keep economics independent of psychology.[87]

Though I *do* believe that the approach is incoherent and I also believe that over the long haul it should be given up, the preference-based rational choice approach served economics very well. Shielding the discipline from empirical psychology, though untenable in the long run, has been and still is pragmatically a good idea within a rational choice approach to human behavior. I will in the following rely on a preference-based framework as well. The reason for sticking with classical decision and game theory is simply that such models are the best we have at present. It is good policy not to give up a well-developed body of theorizing as long as there is not yet anything that we could put in its place. Moreover, and even more importantly, as will become obvious, regardless of the latent incoherencies of a preference-based approach, such an approach fits very well with the program of reconstructing the search of a rational actor for a reflective equilibrium in action situations. For our philosophical rather than our empirical interests, a preference-based approach makes a lot of sense. The same holds true for us as reflecting and acting human beings. So, let us turn to internalist models of rational choice making, while keeping the preceding criticisms in mind as warnings however.[88]

3.3 Minimal rationality of a single actor

Regardless of the fact that rational choice theories, in particular those of economic origins, practically always have an externalist streak, they also have a tendency to lean towards adopting the internal point of view of the decision maker and her reasoning. Acknowledging Ken Binmore's perceptive analyses of these types of rational choice theories, I will subsequently often refer to them as "**eductive**" approaches (see Binmore (1987/88)). The notion of eductive decision making captures the fact that classical rational choice theory is a kind of social theorizing performed from a participant's internal rather than an objective, external point of view (see the previous chapter 2 for additional detail). The perspective is not that of an external observer of overt behavior who explains what he observes as the

[86] It deserves to be noted as an aside that economists are driven by the same concern as the traditional sociologists to keep the threat that their theories might be reduced to psychology at bay.

[87] The fundamental criticism of economics as idealist model Platonism is presented in Albert (1967). The criticism of someone who has brought the program to its extreme can be found in Selten (1990).

[88] The following may be used as a first introduction or as a re-introduction to those who have some previous knowledge of the field but are somewhat uncertain about "what it all means." Those who know all this just should jump to the next section.

outcome of causal processes under causal behavioral laws. It is rather that of an individual who perceives herself as exerting causal influences on the world. She reflects on the interventions that she *makes* rather than *predicts*. In deliberating decisions, she conceives of herself as a choice *maker* from whom causal chains originate and she conceives of herself as guided by reasons rather than as an entity subject to causal laws.

In sum, reasons rather than causes explain the decision behavior, at least to the decision maker herself.

3.3.1 The principles of intervention and opportunism

The most crucial assumption of (eductive or internalist) rational choice theory in terms of philosophy concerns the exercise of the human ability to

> 1) distinguish between what is and what is not a consequence of choices, which is referred to as the "principle of intervention," and
>
> 2) act in view of the future causal consequences of the actions to achieve what is regarded as an improvement of the situation, or the "principle of opportunism."

The principle of intervention is a principle of "world making" or of *construing mental representations of the world.* It requires that a rational individual *tries* to distinguish between what are and what are not causal consequences of possible alternatives of choice and, therefore, makes such distinctions in her model building. The principle of opportunism dictates that the interventions singled out by the principle of intervention be chosen in an opportunity-seeking way.[89]

When a person complying with the principle of intervention *forms a mental model* of an action situation, she will try to distinguish between the effects of her own actions and events that emerge independently. She imagines herself as an "*intervening*" actor in the strict sense in that she conceives of her own actions as "made" rather than caused.[90] Yet, saying that she judges herself to be an actor

[89] Admittedly, we could with some justification classify the principle of opportunism as a principle of intervention as well. Yet, I have now become too accustomed to the way the concepts are explicated to change terminology. It is important, however, to note that the two terms mean rather distinct things.

[90] On a higher level of analysis, the "choice maker" may well know that her choices are caused, but for the purposes of a mental representation of the action situation, she models herself as if she were the origin of an uncaused choice that she can make or not, but to which she cannot assign probabilities in the act of making it. As a philosophical aside, one might note here that at precisely this point a Kantian or perhaps better a Strawsonian participant's attitude meets with the causalistic Savage variant of decision theory as well as the basic assumptions of classical game theory, which has been formulated strictly from a participant's point of view rather than an

from whom an “uncaused” chain of effects originates when she makes her choices does not imply that she “judges right” (alluding to the Spinoza citation above, 3.1). The claim is that the (possibly erroneous) *self-image* of the rational actor in making a choice is such that she is *making* the choice rather than predicting it in and by her act.[91]

In sum, let us say that a person acts “***opportunistically rational***” if she, first, forms a mental model of the action situation in ways complying with the principle of intervention and, second, acts according to the principle of opportunism. The opportunistically rational actor chooses in view of causal consequences better rather than worse interventions (alternatives). She does so according to some standard of goodness, which may be related to anything that is relevant for the actor including unselfish motives, ideals etc.

For the sake of specificity, imagine a simple decision table, in which the columns name states of the world conceived as independently emerging (independent of the choice maker’s interventions), while the rows are the choice options conceived to be open to the individual. In the act of *making* a choice, the opportunistically rational choice maker *chooses* the row. She obviously does *not predict* its emergence but makes the corresponding choice by singling out the row.

From a technical point of view, rows are simply *functions* that map the states of the world into the results that are evaluated by the individual. The opportunistically rational individual chooses the function (i.e. that intervention or act) that maps the set of circumstances or events that *she frames as beyond her personal control* into a vector of consequences deemed most desirable by her in view of her other expectations.

Choices or actions	States of the world	
	Non-S	S
i	State of the world is determined by Non-S and i	State of the world is determined by S and i
u	State of the world is determined by Non-S and u	State of the world determined by S and u

Table 3.1: Choices and states of the world

In setting up the decision table 3.1 as the mental model of the action situation, the decision maker has to distinguish between states of the world that emerge independently of her own choices and states that emerge due to her causal influence on the world as expressed by choices. Circumstances that are influenced by her choices are grouped according to the choice alternatives in rows; thus, the

objective behavioral point of view; see on the participant’s vs. objective point of view Strawson (1962).

[91] To apologize beforehand, let me announce that this, to remind the reader, will be repeated several times later on.

rows, i, u, represent acts. Circumstances beyond those choices are grouped in columns, S, Non-S; thus, the columns represent act-independent expectations of states of affairs.

The results might be represented by very complex and detailed descriptions of the states of the world. The description of a state of the world may, for instance, comprise a characterization of the distribution of goods to all persons concerned. The distribution emerges from the combination of the act (the choice of the function) and the "state of nature" that has been realized as the argument of the function. If the model is well-specified in a way that complies with the principle of intervention, then the action-dependent conditional probabilities of all states $x \in \{S, Non-S\}$ fulfill p(x/u) = p(x/i) for all acts u, i.

Note that the independence of probabilities from acts as introduced here is a *modeling requirement* not a substantive thesis about how the world lies. It demands that the world be modeled such that actions must have the relevant properties of actions and states of the world must have the relevant properties of states of the world.

In this kind of model, things must be described that way. If we cannot find a partition of acts and events such that the relevant relations hold true, then we should not use the language of RCM. It is one of the requirements of rational choice modeling rather than rational choice theory, RCT, that a table like the preceding must be *read* in a specific way. According to the *semantic rules* of such modeling, the individual whose own view of the action situation is represented in the table thinks that the world lies in a specific way.

Empirically speaking, the commitments that go along with RCM as such are harmless. They only specify how to interpret a table if a table is set up as the model of the situation. Yet, as opposed to RCM, which leaves open any thesis about the nature of the actors depicted as choice makers, RCT may be much less harmless. RCT becomes so in particular if it is meant to state certain hypotheses about real persons as choice makers.

In RCT, opportunistic rationality is ascribed to economic man. This is unavoidable if we take seriously the assumptions underlying **homo oeconomicus**. Such an individual takes every act separately or by itself. He distinguishes between aspects on which he exerts a causal influence by his choices and other aspects of the situation. This separates each of his decisions from the past. For him, bygones are bygones, and with backward causation excluded, he decides sequentially (or incrementally) in each instance with respect to the future only.[92]

The immediate most relevant consequence of this specific homo oeconomicus variant of RCT is that the rational actor cannot meaningfully choose to *make* a series of choices in one act. He can only *plan* on a series of such choices, but he

[92] "Modular rationality" fits as well as a term here; for an attack on this incrementalism see McClennen (1990).

cannot make them in one choice-act. There is no act of which this would be the causal consequence (unless there is *in fact* what we will call later "commitment power" or the ability to "link" several separate acts). When actually going through the series of choices, he must make decisions sequentially, one at a time (or incrementally).

In sum, according to the principle of intervention, at each point in time the rational actor of standard RCT distinguishes between the causal effects of his choices and the aspects of the decision situation that are not causal effects of his choices and then goes for the subjective best.

It is important to distinguish clearly here between RCM and the specific RCT implied by the homo oeconomicus model. RCT can be expressed in models formed according to the semantic rules of RCM. These rules of interpretation merely spell out what the signs mean. Using the signs, we then know what is assumed about a rational choice maker according to the models proposed. If we choose to express some kind of thesis about how the world does in fact lie by means of RCM, then we implicitly commit to the view that we have succeeded in singling out *entities* to which we can meaningfully ascribe opportunistic rationality. However, the entities need not be persons. Therefore, to model something in terms of RCM does not amount to the view (RCT) that human persons are opportunistically rational in any conventional sense.

To put it slightly otherwise, only if we claim that the choice making entities of RCM are human persons, have we endorsed more or less the classical homo oeconomicus theory, according to which personal actors are in fact acting opportunistically rationally. However, the entities whose evaluations and choices are represented by preferences can be several persons forming a corporate actor as well as several sub-personal agents who are organized such as to form a person.[93]

For now, let us assume that the entities that make the choices according to the model formed in the language of RCM are human persons. In this case, it is assumed that they can instantaneously shift their behavioral gears when they are to make a choice. Since as choice makers they conceive of themselves as initiating a choice, they (subjectively) can always follow Huck or do what "come at the time." And, for the same reason, it is ruled out that they apply probabilities to their *own* acts.

As participants of an interaction, the persons need some guidance other than prediction, namely values or evaluations. To these and their representation, we turn next.

[93] We will make some rather extensive use of this latter possibility to model homo oeconomicus as a game played by several homunculi oeconomici "teaming up" or playing a game internal to homo oeconomicus; see below 6.

3.3.2 Consistent preferences and their representation by an index

3.3.2.1. Fire protection

Consider the following specific example of a so-called **game against nature**: the game of "fire protection." There are two states of the world, F(ire) and N(o fire), whose emergence is conceived by the model building choice maker as causally independent of his choices. It is a "small world" in the sense that nothing else, at least according to the model, is expected to occur in the future. Besides the states of the world, there are two options that can be chosen to exert some "value relevant" causal influence on the future.

To have a specific example assume from now on that in table 3.1 one option is that of becoming i(nsured). The other option is that of remaining u(ninsured).

Choices	States of the world	
	N with probability p	F with probability 1-p
i	v(N, i)	v(F, i)
u	v(N, u)	v(F, u)

Table 3.2: Fire protection

One of the options, u, amounts to doing nothing. It is an omission in the common sense meaning of the term. The other one, i, is an action in the more narrow sense of the term, implying "activity." Yet, this difference in interpretation does not matter for the purposes at hand (and, in general, does not matter in rational choice approaches focusing on consequences). The options i and u are symmetric insofar as both u and i may be chosen by a rational choice maker in anticipation of the consequences of making one of the alternative choices.[94]

To make things very simple, let us assume initially that states of the world as well as the options leading to those states are exclusively evaluated in *monetary* terms. Preferences among options are formed according to monetary expectations alone. In the example of fire protection, let us work on the premise that the *monetary values v*, which accrue if "option-independent" states emerge, are known. That is, if option-independent state X and action y are realized, then the net monetary value of the state emerging is known to be v(X, y). Let these values be represented by real numbers that comply with

$$v(N, u)>v(N, i)>v(F, i)>v(F, u).$$

[94] There is a normatively relevant difference between action and omission at least in everyday life, but that is of secondary importance here.

The states F and N of the world are expected to occur with conditional probabilities

p = p(F/i) = p(F/u) and p(N/i) = p(N/u) = 1 – p, respectively.

That the probabilities of fire under condition of being insured, F/i, or uninsured, F/u, are the same is neither trivial nor self-evident; instead, it is a substantial assumption that must be fulfilled if the model is to represent the true structure of the world. It makes it viable to represent a problem such that the options and the states of the world form independent dimensions. If no such partition can be found, the model is not applicable as a representation of the world. Analysis of the action situation must be pushed to that stage if we intend to use RCM to express ourselves. Conversely, once we write down a model in the tabular form used here and assume that the rules of interpretation of RCM apply, we must be of the conviction that the part or aspect of the world depicted by the table fulfills the requirement that conditional probabilities of states are action independent.

Alluding to the belief-desire-framework of modern action theory, we might say that p represents "beliefs" and, assuming that individuals strive for monetary gains, v represents "desires". The practical rational choice question to be answered on the basis of a model like *fire protection* is the following one: What should the choice-maker rationally choose given his beliefs and desires if the action situation can be depicted as in table 3.2?

Since the world is not fully controlled by our actions, practically everything we can do is only probabilistically linked with results. Choosing an option, we do not deterministically choose one and only one state of the world but rather a class of possible results and a probability distribution over those results. *In other words, as long as there are factors beyond our control, we always choose a **"lottery"** rather than a specific result.*[95]

In the case at hand, one lottery, i, can lead to the outcomes "N-and-i" or "F-and-i", the other to the outcomes "N-and-u" or "F-and-u."[96] Taking into account the (option-independent) probabilities p and 1-p and substituting "-and-" with "&", we get the two lotteries

i: = ("N&i", p; "F&i", 1-p) and u: = ("N&u", p; "F&u", 1-p).

Which of the lotteries is the better choice depends on individual preferences, i.e. the ranking order between the states emerging as evaluated by the individual. In

[95] We assume, however, that the choice of an action itself is deterministic rather than merely probabilistically linked to decisions and plans – for the time being not even allowing for slight mistakes.

[96] More precisely speaking, we would have to deal with the states of the world brought about by the joint causal influence of actions and states of the world. Yet, for the sake of simplicity, let us identify the results simply by the factors that uniquely determine them according to the assumptions made here.

the present case, the ranking order between states of the world is determined solely by the monetary payoff. However, this does not yet tell us how to evaluate lotteries in which these states of the world are realized only with some probability, p, 0<p<1.

According to one intuitive idea, we should simply form expected values in which we weigh the monetary payoffs with their probabilities and then calculate the sum (see on the history of probabilistic thought Gigerenzer and al. (1989)). Having done this, we could say that lotteries should be ranked according to their expected value.

In the simple case at hand, we would form

$$E(u) = v(F \,\&\, u)p(F) + v(N \,\&\, u)p(N)$$
$$E(i) = v(F \,\&\, i)p(F) + v(N \,\&\, i)p(N)$$

and then ask which one is larger. Proceeding this way, we no longer take into account how the probabilities are “distributed” by the lottery over events. All of the information is collapsed into a single number. Using numbers to represent the ranking of alternative lotteries is convenient since we can use the natural ranking of numbers to indicate the location of a lottery in our preference ranking. Yet, the expected monetary value may not truly represent our preferences in lotteries.

For instance, in the case of fire protection, the whole point of buying insurance, i.e. of choosing option i, consists in reducing the “spread” of monetary outcomes as occurring by chance. A sure loss, namely the insurance premium (and the reduction of the expected payoff by paying the premium) is often preferred to the higher spread of outcomes going along with the higher expected value (no premium paid). In short, *that expected monetary values do not represent our preferences among lotteries is the whole point of insurance.*

In view of the preceding, we need rank-indicating numbers including attitudes to risk. Once assigned, such a number “sticks” to the lottery like a price tag on a consumer good. With the appropriate “sticker,” we would always know whether a lottery is ranked higher or lower than another lottery by the choice maker. We would have an “ordinal” – i.e. just giving the rank in the order – representation of preference rankings of *lotteries.* This ranking could be formed as follows: The choice maker herself can in principle compare any lotteries l and l’ as a whole (or “holistically”). In the case of fire protection, she would just look at i and u and then say which is the one she prefers. Then ordinal numbers representing the ranking could be assigned accordingly; i.e. the highest to the highest ranking lottery, the second highest to the second highest lottery etc.

The expected monetary values of lotteries may play a role in ranking lotteries; the choice maker could take the expected values into account as *one* piece of information, but they would as a rule not be decisive for her ranking of i and u. She would make up her mind by looking at the whole picture, in particular the spread of risks and then after she has considered everything she would, “reveal” – as economists say – her preference between i and u by her choice. If she chooses i consistently when u is present as an option she thereby shows what she prefers. Whether such choice is what it “means” to prefer will be left open for the time being.

In sum, since we know from the insurance example, as well as from day to day experience, that expected monetary value will not suffice (or only in special cases), the function representing preferences under expectation formation cannot be a monetary evaluation function like the preceding function v. The formation of the function must include other dimensions of value and in particular incorporate *attitudes towards risk.*

Note again that a choice maker may make all comparisons between lotteries in the preceding sense of taking whole lotteries as objects of comparison *holistically*. Always looking at the full lotteries, she would order them. Nevertheless, it would be convenient if the ranking number of a lottery could be calculated from the ranking of prices of the lottery by forming an “expected value” of the numbers that represent the ranking of the prices. Let us refer to the function we are looking for as “v” (as before). However, now the construction of v is such that – beyond monetary evaluations – it can incorporate *all* “rank-relevant” considerations including attitudes towards risk.[97] With this aim in mind we can turn to the minimum required to put the basics of *standard utility theory* into perspective.[98]

3.3.2.2. Construing a measuring rod for value

Assume we want to derive value functions v that incorporate value dimensions other than monetary ones and, in particular, attitudes towards risk. We intend to do this such that we can use expected values for calculating the position of lotteries in our ranking order among various lotteries. The trick is to assign values by v_r such that expected value formation will not lead to distortions of “holistic” preference rankings that an agent r has among lotteries. The individual r whose preferences are to be represented by the expected values formed on the basis of v_r will come to the same conclusion if she evaluates the lotteries as a whole as she would by

[97] That such representation by v is possible and that such v exist can be proved if certain conditions are fulfilled. Because the precise proofs of representation theorems do not tell us much about what is going on, at least not much that would go beyond very simple intuitive constructive considerations, we will side-step that part of the construction. A beautifully simple and lucid construction can be found in Binmore (1992).

[98] The following may be used as a first introduction or as a re-introduction to those who have some previous knowledge of the field but are somewhat uncertain about “what it all means”. Those who know this all just should jump to the next section.

calculating their place in her ranking analytically by expected value formation on the basis of evaluations of the lottery prizes.

To see how this is accomplished, start again with the two options of i and u in the preceding example of table 3.2. If the assignment of v succeeds, then which of the two options is or should "rationally" be chosen is indicated by the larger of the expected values accruing to the two options. That is, answering

$$v_r (N, i)p+v_r (F, i)(1-p) >?< v_r (N, u)p+v_r (F, u)(1-p)$$

shows whether

("N&i", p; "F&i", 1-p) or ("N&u", p; "F&u", 1-p)

is preferred if the choice maker r were to make the choice "holistically" between i and u. The analytics involved here represent the preferences as revealed by choices of options.

We measure the "values" of prices by means of a *basic lottery ticket*, the so-called BRLT ("basic reference lottery ticket," see Raiffa (1973)), in which the probability of winning the better alternative can be varied continuously. A somewhat simplified construction is the following:

1. Assume that in a "small world," all of whose possible states you know, there are:

1.1. an optimal state of the world or a bliss point A, the best conceivable outcome, and

1.2. a devastating state of the world, Z, the worst conceivable outcome, further

1.3. a *generic lottery* L = (A, *p*; Z, 1- *p*) and a *class* of lotteries BRLT, all of the form L_{BRLT} = (A, p_{BRLT}; Z, 1- p_{BRLT}) with variable p = p_{BRLT}.

2. Assume that p_{BRLT} can be adjusted according to a random mechanism that yields any probabilities conceivable.

For each *p*, a separate lottery with winning probability *p* for "bliss price" A emerges. Note also that there is a *natural ordinal ranking* among the lotteries L = (A, p; Z, 1- p), (A, p*; Z, 1- p*), (A, p**; Z, 1- p**) etc, which is the same as the ordering of the p, p*, p** (a lottery with a higher probability of winning the better price is better). According to the probability p, p*, p** of winning the price A, we have as the natural ranking

$$(A, p; Z, 1- p) \text{ preferred to } (A, p^*; Z, 1- p^*) \Leftrightarrow p> p^*.$$

Imagine now you, r, own item h, and somebody offers you lotteries with alternative probabilities p*, p** for reaching the optimal result A. For any probability *p* offered, you must ask yourself the question

Do I or do I not prefer h to the lottery L = (A, p; Z, 1- p)?

Depending on how you answer that question, you will either be indifferent, or want to exchange your h for the lottery or you will not want to do so. You are indicating thereby your preferences and relating h to the natural ranking among lotteries according to the (higher) probability for the bliss price. This suggests the way to find the "value" of h as measured by BRLT:

3. To find the value of h in terms of lotteries, seek that probability $p_{h,r}$ that will make you, r, indifferent – you neither prefer the one nor the other – between h and (A, p; Z, 1- p). Once you have found that probability, we have:

r is indifferent between h and $L_{h,r}$ = (A, $p_{h,r}$; Z, 1- $p_{h,r}$).

Refer to $p_{h,r}$ as the **indifference probability**.

Now, the crucial but simple trick is to use indifference probabilities $p_{h,r}$ in BRLT as measures of value v_r of the alternative h. If the value $v_r(h)$ ascribed to any alternative h is determined by the indifference probability $p_{h,r}$ in the basic lottery ticket BRLT, we get $v_r(h) = f(p_{h,r})$.

For any alternative h, the indifference probability in BRLT serves as the measure of value as returned by applying the yardstick of BRLT.

4. We can fix for each individual r the values of A to be

v_r (A) = v((A, 1; Z, 0)) : = 1

v_r (Z): = v((A, 0; Z, 1)): = 0

Note that this fixing of extreme values is only a matter of *convenience*.

5. With the preceding "normalization" of the value function v_r for person r, we can for each h of a set of alternatives, from which the person r chooses, fix

$v_r(h)$: = $p_{h,r}$, the indifference value that r assigns to h in BRLT.

Note that with this definition and the convenient setting of the values of A, Z we have the desired property of expected value formation:

$p_{h,r} = v_r(h) = 1\ p_{h,r} + 0\ (1-p_{h,r}) = v(A)\ p_{h,r} + v(Z)\ (1-p_{h,r})$.

More generally speaking, it can easily be shown that the procedure – if certain general conditions are fulfilled – yields a function v_r that can be used to calculate the position of any lottery in the ranking order of an individual on the basis of the expected value of the values v of the prices determined according to the procedure. The holistic judgment of an individual who is rational (in a very elementary sense of consistency in her evaluations) must conform to the analytically determined judgments calculated by forming expected values from the values v_r and the probabilities p.

For any alternatives h and k as well as d and e, both evaluated according to the procedure described for an individual r before, we can find the ranking between the lotteries $l = (h, p_h; k, p_k)$ and $l' = (d, p_d; e, p_e)$ by using the indifference values in the BRLT.

To do so, we form

$$v_r(l) = v_r((h, p_h; k, p_k)) = v_r(h)\,p_h + v_r(k)\,p_k = p_{h,r}\,p_h + p_{k,r}\,p_k$$

$$v_r(l') = v_r((d, p_d; e, p_e)) = v_r(d)\,p_d + v_r(e)\,p_e = p_{d,r}\,p_d + p_{e,r}\,p_e$$

Now we can read off the position of the lottery in the holistic ranking by comparing the numerical values:

$$v_r(l) = v_r(h)\,p_h + v_r(k)\,p_k <?> v_r(l') = v_r(d)\,p_d + v_r(e)\,p_e$$

From this simple comparison, we get the answer to the question of whether one of the lotteries l, l' is strictly preferred or whether indifference between the two lotteries applies. So, once we have determined the values of all relevant prices by some BRLT for some individual r, we can describe the position of arbitrary lotteries of the prices in the preference order of the individual r by the expected value of the lotteries.

In sum, the holistic ordering of lotteries for an individual r can be analytically determined by using the indifference values from BRLT and then calculating expected values.

Note that v_r is an evaluation function relative to the aims, ends, or values of the specific person r. It is strictly *agent-relative*. In contrast, a lottery in monetary prices leaves this kind of agent-relativity out. If the monetary value is indicated, say, by "m, m'," we get in general:

$$v_r((m, p_m; m', p_{m'})) = v_r(m)\,p_m + v_r(m')\,p_{m'} \neq m\,p_m + m'\,p_{m'}.$$

In the case of fire protection, we had:

> i: = ("N&i", p; "F&i", 1–p) and u: = ("N&u", p; "F&u", 1–p). We used *v as (monetary) value*, which did *not* factor in the specific attitudes of agent r.

So, in general $v_r(i)$ $= v_r(\text{"N\&i"}, p; \text{"F\&i"}, 1-p))$

$= v_r(N\&i)\,p + v_r(F\&i)\,(1-p)$

$\neq v(N\&i)\,p + v(F\&i)(1-p)$

Likewise, $v_r(u)$ $= v_r((\text{"N\&u"}, p; \text{"F\&u"}, 1-p))$

$= v_r(N\&u)\,p + v_r(F\&u)\,(1-p)$

$\neq v(N\&u)\,p + v(F\&u)(1-p).$

Note also that it is unnecessary to have a best alternative A and a worst alternative Z as a basis for the measuring rod, the BRLT. To get the measurement working, it is only necessary that A is strictly preferred to Z by r. However, the way we have proceeded is sufficient to illustrate the basic point.

In sum, we can use a measuring rod for representing preferences that always measures risk and, thereby, has the attitude towards risk built in already.

Finally, any function v_r is only one of a whole class of functions that are as good as any other in representing individual preferences among lotteries. So-called positive affine transformations that multiply the measure by some positive number and add some constant can be applied *without distorting the ordering of lotteries and the expected value property of vr.* More specifically, for any v_r , we can use $a_r > 0$ and b_r to yield $v'_r := a_r v_r + b_r$. As a representation of individual preferences among lotteries v_r and v'_r are each as good as the other. If v_r is true to the formation of expected values, so is v'_r. The absolute values of v_r do not mean anything. It is also meaningless to express something in terms of multiples of scale values. If an alternative has twice the value of another one, we can easily find a scale on which this is not true (just subtract a constant b_r from v_r). Such a claim is, therefore, not invariant with respect to arbitrary choices of scale within the scope of admissible transformations. For all alternatives h and k as well as d and e, we have the following invariance though

$$\frac{v'_r(h)-v'_r(k)}{v'_r(d)-v'_r(e)}=\frac{(a_r v_r(h)+b_r)-(a_r v_r(k)+b_r)}{(a_r v_r(d)+b_r)-(a_r v_r(e)+b_r)}=\frac{v_r(h)-v_r(k)}{v_r(d)-v_r(e)}$$

In sum, the relative size of differences in values between alternatives is indeed invariant with respect to the specific form of representing the preferences among lotteries by a value function that has the *expected value property.*

Let us observe by way of a concluding remark that, of course, individuals may not only have **strict preferences** "P_r" over alternatives but rather may be indifferent between alternatives as well. Let us indicate the **indifference** of individual r by I_r and define **weak preferences** xR_ry as $xR_ry :\Leftrightarrow$"xP_ry or xI_ry." Stating that weak preferences xR_ry apply amounts to saying that an alternative x is at least as good as an alternative y. Obviously, we could just as well have defined $xI_ry :\Leftrightarrow$"xR_ry & yR_rx" and $xP_ry :\Leftrightarrow$"xR_ry & not (yR_rx)".

In the following, we assume that preferences among alternatives are such that all individuals r

can compare any two alternatives x and y according to xR_ry (completeness)

xR_rx (total reflexivity) holds good for all x and

for all x,y,z, it is true that xR_ry and yR_rz imply xR_rz (transitivity).

***In sum**, and for the following, it will be assumed throughout that the preference orders of individuals are complete, (totally) reflexive and transitive and that they can be represented by a utility index that has the expected value property.*

3.4 Understanding preference representations

One should be careful to note that according to the concepts of preference and belief *representation* employed here, it is *not* (!!) true that one alternative is preferred to another *because* it has a higher expected value. The whole point of the modern concept of "*representative* utility (cum probability)" is that the reasoning is just the other way around. Whenever certain representation axioms are fulfilled, we can assign functions v and p such that a preferred lottery has a higher expected value associated with it than a less preferred one. The functions v and p are chosen such that expected values can be used to *calculate* the position of a lottery in a preference ranking of lotteries. However, the position of the lottery in the preference ranking among lotteries is determined independently of ("before") calculating the expected value of the subjective value function v_r. The position is *not* attained because of the calculated value.

In sum, the expected value of the value function v_r is not among the reasons that are operative in fixing the preference order of an individual r.

Since we all tend to get confused about this rather trivial point once in a while by habits of our language, let us look at another illustration. Assume for the time being that:

$$v_r\,(N, i)p+v_r\,(F, i)(1-p) > v_r\,(N, u)p+v_r\,(F, u)(1-p).$$

Assume also for monetary values v that

$$v\,(N, i)p+v\,(F, i)(1-p) > v\,(N, u)p+v\,(F, u)(1-p).$$

If individual r were neutral about risking money while being exclusively interested in monetary gains, then the *reason* for preferring i to u would be that the *expected monetary* value of choosing lottery i is greater than the *expected monetary* value of choosing u. It would be true *then* (!) that i is preferred to u *because* of the greater monetary expectation of i.

If, however, v_r represents preferences among lotteries that emerge, "all things considered," then the expected value v_r does not and cannot meaningfully enter into the set of reasons guiding preference formation among lotteries. It *represents* the *result* of preference formation "all things considered." To use expected value formation on the basis of v_r as a consideration of what to prefer and what not to prefer would *then* amount to "double counting."

Avoiding double counting, it is hard to imagine how it could be rational to choose u if v_r represents preferences "all things considered." If the represented order among the lotteries (actions) i: = ("N&i", p; "F&i", 1–p) and u: = ("N&u", p; "F&u", 1–p) is the one that emerges "*all* things considered," there is no consideration left on behalf of which we could provide a *reason* for a "preference reversal." In particular, attitudes towards risk are already included in the representation $v_{r.}$. For, individual r's attitude towards risk has been measured or *factored in* when using BRLT (a *risky* lottery) as a measuring rod. The attitude towards risk cannot conceivably be used again without double counting to re-evaluate the ranking.

An actor can, of course, choose in deviation of the monetary expected value of the function v without double counting. Moreover, given an appropriate interpretation of the preference concept, r can *choose* against her preferences. As long as we do not interpret the functions v_r in behavioral terms, we can at least admit that actors can act otherwise than they "should" in view of their own preferences concerning lotteries. This would only be otherwise if we were to claim that the term "preference" means "will as a matter of fact be chosen." In that case, x P_r y or "x is preferred to y" would mean that "y would never be chosen by r in the presence of x" from a choice set that contains both x and y (implying that x would be chosen from the pair if no other alternatives are present). This would be a dispositional predicate based on a behavioral law characterizing the actions of r, and it would be in line with a purely externalist view on choice making.

It may be that such behavioral laws sometimes exist for some individuals r. However, it is at least problematic to simply assume that they exist. In addition, it is clear that the meaning of "preference" cannot be reduced to the behavioristic interpretation in many contexts in which we tend to rely on it. There is a non-behavioral, evaluative notion of preference. It sums up the results of evaluations or sometimes expresses them. Here preferences sum up our reasons for ranking rather than representing choices that are made on behalf of reasons.

Although we can, of course, define terms in all sorts of ways, that liberty does not apply if we intend to provide explications of theoretical terms that are reasonably similar to everyday uses of terms and can fulfill our theoretical purposes. In the case of the preference concept, it is simply wrong to claim that the pre-theoretical meaning of preference can be reduced to "preference as revealed in choice".[99] Moreover, *the expected value of a function vr that is formed "all things considered" is merely indicative of and not constitutive for preference formation of an individual r.*

Expected value formation can play a different role with respect to substantive rather than preference-representative functions v_r. As we have seen, we can consider the expected monetary value v as one of the value dimensions entering in

[99] See for a revealing criticism of revealed preference Sen (1973/1982).

our process of forming a preference among lotteries. A higher monetary value v is then one of the (possibly many) reasons to prefer some alternative to another one. It becomes co-responsible for the ranking. However, except for some endearingly old-fashioned economists, practically nobody will ever doubt that it can be rational to endorse values other than a desire for monetary gains. Therefore, preferences deviating from those formed by the aim of maximizing expected monetary value can be perfectly in line with rationality precepts. An individual who chooses a lower expected monetary value can do so rationally if there are things she values other than expected monetary gains.

In sum, according to the preceding line of argument, we would obviously be mistaken to associate rationality with maximizing an objective function like monetary rewards. It is not in itself any less rational to be interested in things other than money than it is to go for the money. Rationality does not seem to be related to the substantive content of the objective function an individual pursues. At least within the means-to-given-ends framework of traditional economic rational choice modeling, rationality is merely related to the ways the aims, ends, or values are pursued. Whatever the aims, ends, or values that induce individuals to prefer certain alternatives to others, consistency and representability of preferences that emerge are decisive, not the substantive content.[100]

Consistency and the absence of circular structures like the aforementioned money-pump form the minimum requirement of what may be called ***evaluative rationality***. However, evaluative rationality is not all that matters.

3.5 The minimal rationality of Ulysses and the principle of intervention

The so-called Ulysses problem is well-known (for instructive discussions from a modern rational choice point of view see Ainslee (2002), Ainslee (1992), Frank (1987), Frank (1988)). It emerges precisely because a rational actor can only *plan* to make future choices “now;” he cannot actually *make* them “now.” To put it slightly otherwise, he can plan on a strategy but not literally choose the sequence of choices forming that strategy in one act.

In qualitative terms, the problem is best represented from a first-person or I-perspective. Imagine yourself in the shoes of Homer’s Ulysses:

> I now prefer listening to the sirens over not hearing them sing.
>
> I now prefer not following the sirens over falling prey to their alluring songs.

[100] In this sense, “pushpin is as good as poetry.”

> Rationally planning not to follow the sirens in the future will not suffice since I will not then prefer to decide as planned now. (When I hear the sirens sing (in the future), my preference will be different from now.)
>
> How can I listen to the sirens and not follow them?

In the Homeric epos, Ulysses is not alone. There are rowers who can bind him to the mast, and there is a mast to which he can be bound. These additional aspects of the decision situation must be taken into account in the decision tree representing the decision problem Ulysses faces.

The following tree shows the sequence of choices that could conceivably be made by Ulysses at times 1, 2, 3 or by U1, U2, U3:

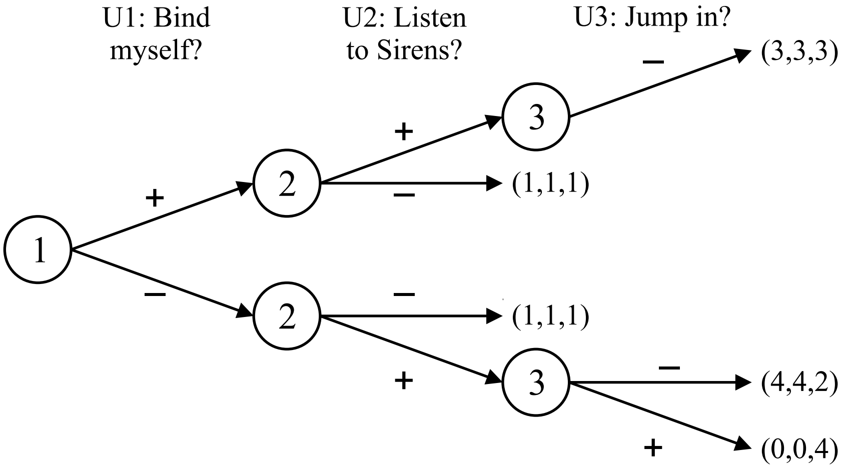

Figure 3.1: Tree of Ulysses' problem

Note that in this representation of the Ulysses problem, we have three evaluations of end results at each final node of the tree. Ulysses does have different preferences at different times. There is only one acting person. However, accepting the *principle of intervention*, that acting person is "forced" to form a model of the action situation that "separates" or "modularizes" him into at least three "agents," corresponding to U1, U2, U3.[101] These "agents" have the preferences of the person at the location of action at times 1, 2, 3. The preferences of the agents are represented by three preferences representing "utility" functions. Each end-node shows how the manifestations U1, U2, U3, the agent of Ulysses at

[101] Skyrms (1996) speaks of "modular rationality" in this context, see pp. 22–25 in particular. It explains more precisely what it means that the rational chooser envisions and then chooses what is "handiest at the time."

stage 1, 2, 3, respectively, evaluate that end-node as its first, second, and third entry.

For simplicity and initial plausibility, it is assumed that at each point in time the same decision maker is making the relevant decision. Nevertheless, in principle we could and in fact will distinguish between choice makers at each and every instance of choice (as in the example of "take it or leave it" below, 5.1.3). The decision makers at the nodes are all distinct entities, but we have collapsed those existing at the same time to one with identical preferences.

Keeping this in mind and being aware of the fact that at each node the past is beyond causal influence, we can analyze the tree. Since each decision is always made in view of the future only, we can look at the last decision nodes first. Ulysses at time 3, or U3, would evaluate the result emerging after (–, +, +) as the best outcome, putting this above (–, +, –). Therefore, we know that this decision maker will choose + after (–, +). Assuming that the decision tree is known to each of the decision-makers, the former decision-makers U1, U2 will understand this analysis and know what the last manifestation of Ulysses will do. The play of (–, +, –) can be left out in further considerations of actors who comply with the principle of intervention in their opportunistically rational choice making. Taking this into account, Ulysses at time 2, or U2, will, after +, at the upper instance of choice, have to compare choosing +, which yields 3 to him, as opposed to 1, which would accrue to him after choosing "–". At the lower instance of choice, U2 will opt for "–" after "–" because this gives him 1 instead of zero. The corresponding lower branches of the tree following "+" can be neglected for further consideration under the assumption of fully rational behavior. Doing this, Ulysses at time 1 – i.e. U1 – by choosing "+" will expect "3" and after choosing the alternative "–" as the initial move will expect from fully rational behavior of U2 and U3 to receive merely 1.

This ends the analysis of the game tree by means of so-called ***backward induction***. Backward induction exploits the fact that, according to the principle of intervention, at each stage of a game comprising several stages only the future matters. If there is a single node that does not have nodes that follow it, then we can simply look for the best result at that node. This will be chosen according to the future expectations at that node. Assuming that there is always a single best such solution – no indifference occurring – we can ignore all the other results at this last stage node and identify it as the single best result that can be reached at that last stage decision node. Looking at all last stage nodes and always substituting the single best result for the node, we can simplify the tree and can work our way backward to its root.

Rather than supporting the conventional view that Ulysses is irrational when jumping into the sea, the preceding model shows that when jumping into the sea, Ulysses can be rational *and* jump. His given preferences at the moment of choice making are such that in rational pursuit of his current given aims, ends, or values, he does the preferred thing at the time or instance of choice. That is not to say, though, that he should not try to exert a causal influence on the future to prevent certain events if he can. He should desire that he cannot make certain choices in the future anymore. Provided that he *can* in fact exert such an influence, he should use that *additional* option of choice. If there is a mast, a rational choice maker facing a Ulysses problem should choose to be bound to it given the desires the choice maker has expressed at the earlier point in time. However, *it is neither the desire to be committed nor the rational insight that it would be desirable to be committed that will commit the individual.*

In sum, endowed with the power of reason, rational beings understand that they *should* commit. They understand that they have good reason to hope to be committed, but that insight does not endow them with the power to commit.

Note that the conventional assumption of many rational choice theorists and, in particular, many economists that, being rational, we should have no problem with sequentially pursuing what is in our rational interest is mistaken.[102] If "commitment power" or the faculty to choose a course of action – containing several decisions – in one act were to emerge whenever rationality made it seem desirable to have that power, the world would be a different place. Yet, "masts" do not just appear simply because it would be "advantageous" to have them.

As we all know, the classical solution of the problem consists in Ulysses ordering his mates to tie him to the mast. Rational Ulysses at time 1 anticipates that the "untied" or "unbounded" rational decision maker at time 3 would jump into the sea. The rational decision maker at time 3 is going to make a choice that is disagreeable to the rational decision maker at time 1. The decision maker at time 1 has a first-mover advantage in that he can causally restrict the option sets of later choice makers or incarnations of his own person. It should be noted well that the first-mover advantage does not emerge because of the time structure as such. It depends on the fulfillment of certain factual conditions. The world must be of a certain kind if a rational first-mover is to exert a causal influence on later decision makers; more precisely in Ulysses' case, the mast, the ropes and others who are willing to bind him must as a matter of fact exist.

This being said, it should not be over-interpreted to imply anything about the world. Although in rational choice modeling it is often treated as a self-evident truth that persons cannot internally commit, this is in no way logically implied by rational choice modeling per se. As a language, RCM provides ways to represent internal commitments of a person as well as external ones, intrinsic as well as extrinsic motivation. It is rather an additional empirical assumption of a specific

[102] To speak here of problems of dynamic consistency is clearly in line with the desire to reduce rationality to consistency, but it is rather doubtful to refer to the aspect of rationality that is based on the principle of intervention using the same term as the one with which attention is drawn to features based on the principle of opportunism.

variant of rational choice theory, RCT, that there is no internal commitment technology that could and would serve for Ulysses the same functions as the mast. If an internal commitment mechanism existed and offered its options to the actor, then, at least conceivably, deciding to become internally committed could serve the same function as choosing to become bound to the mast. The option would have to exist as a *matter of fact* as part of the internal "behavioral technology" of Ulysses.

If an additional such option exists, then, of course, the option should be modeled as a branch of the tree (a part of the tree internal to Ulysses as a person). It could even conceivably be the case that we would know the decision tree that models the decision process taking place inside another person. Assume that e_j represents the internal decision technology of actor j. If j participates as a personal actor in an interaction represented by tree T, then we might write down the complete tree as (e_j-T) with e_j being put in wherever the player under consideration is deciding. Instead of a single decision node representing a personal player, we now have a whole tree in its place.[103] At this specific location, the actor j decides according to his internal decision technology e_j. He is generating some output depending on the input at that location.

The preceding discussion of a single personal actor was already an analysis in terms of several decision makers. Strictly speaking the situation was, at least in the way it was modeled, an inter-active one. It was, in fact, a "strategic game" among several "agents" of the same person. It should seem natural therefore to turn to minimum rationality in inter-active situations in the more conventional sense of the term in the next chapter.

[103] This is like entering a sub-routine in the older programming languages.

4 Models of interaction

The self-conception of a choice maker as a participant of interaction forms the starting point of this chapter. The concepts of choosing moves in a game and that of planning strategies are distinguished and related to the fundamental problem of the strategic commitment of opportunistically rational actors. Prisoner's dilemma-like games provide illustrations that exemplify the differences between choosing a strategy as a program, which requires commitment power, and planning a strategy as a sequence of incrementally rational moves, which does not require any such power.

Much of the following will in all likelihood seem familiar to many readers. Most of them will presumably have already encountered the prisoner's dilemma – which was presented in chapter 1 as the backside of the invisible hand, too – and the other paradigms of social interaction introduced here. However, the subsequent presentation is somewhat different from more conventional ones. Though it may be read as an introduction to game modeling by somebody who has never heard anything about it, the text is meant also as a "re-introduction," for it is intended to shed some fresh light on the "basic philosophical issues" of such modeling. The preceding discussion of the principles of intervention and opportunism as well as that of the meaning of the concept of "preference" and the utility representations of preferences should be "illuminating" subsequent discussions. The main points to be seen in a new light are, first, how basic conceptions of "causality" and "action" enter into the picture when we seem to do nothing but elementary game modeling of interactive decision situations and, second, what the concept of utility representing "given" preferences does and does not imply.

4.1 Prisoner's dilemma and some concepts of strategic analysis

4.1.1 The original story and its representation

The original story of the prisoner's dilemma is this: Two suspects are held in custody. The district attorney can provide sufficient proof to convict both of them of a minor crime, say, illegal possession of certain guns. This would put both suspects in jail for 1 year. The district attorney knows that the two suspects have also committed a major crime but cannot prove it without at least one of them confessing to the crime and providing additional information. To elicit a confession, the D.A. can offer the privilege of becoming the state's witness. Then, if one and only one of the two suspects decides to confess, he will become the privileged witness and go free without serving any time in jail. In that case, the other one will serve 10 years. Should, however, both decide to confess neither of them will become a privileged witness. In that case, both will serve 9 years (10 years reduced by 1 year for confessing to the crime).

The preceding verbal presentation of the problem is not very transparent. However, we get by with a little help from our friend, decision theoretic modeling. Use C_A for "A does not confess," C_B for "B does not confess," D_A for "A confesses," D_B for "B confesses," and assume that both actors conceive of the choices of the other one as causally independent from their own. Now, if A treats the choices of B, according to the principle of intervention, as independent of his own (like states of nature in the fire protection example), the outcomes for A would be as follows:

Prisoner B Prisoner A	C_B	D_B
C_A	1	10
D_A	0	9

Table 4.1: Prisoner A

If B treats the choices of A, according to the principle of intervention, as emerging independently of his own choices, the outcomes for B would be as follows:

Prisoner A Prisoner B	C_A	D_A
C_B	1	10
D_B	0	9

Table 4.2: Prisoner B

Both A and B choose functions that take the causally independent choices of the other as arguments into the collective results. Based on table 4.1. and table 4.2, we can form a compound table with the rows of table 4.2 as columns:

Prisoner B Prisoner A	C_B	D_B
C_A	1,1	10,0
D_A	0,10	9,9

Table 4.3: Game form of classical PD in strategic representation

In the cells of table 4.3, the number of years served by A shows up first and the number of years served by B second. In a two-by-two PD situation,[104] each of the two prisoners as a **player** can choose to either co-operate "with" the other player or to defect from what seems to be the common interest of keeping silent. That is why the choice to co-operate with the other actor is indicated by a capital C indexed with a subscript referring to the player who makes that choice. Analogously, the option of defection is symbolized by a capital D again with a subscript indicating the actor who can choose that option. (Both indexing corresponds to that in the Crusoe and Friday case discussed in chapter 1.)

Note carefully that up to this point nothing has been said about how the two actors evaluate the results, which are presented in material payoffs of the dimension "years served in prison." This is why table 4.3 has been characterized as a "game form." To transform this into a game, *assume that, all things considered, both prisoners always prefer a lesser sentence for themselves to a longer one without paying attention to what happens to the other prisoner.* We get value rankings v_A, v_B, or evaluation functions whose values are (at least) "**ordinal payoffs**"[105] representing the ranking of the emerging states of affairs from the point of view of A, B, respectively. The evaluation functions can be represented in relation to the choice making of the two prisoners by a table like the following one:

Actor B Actor A	C_B	D_B
C_A	3, 3	1, 4
D_A	4, 1	2, 2

Table 4.4: Classical generic PD in strategic form with ordinal payoffs

In the table, only "ordinal utility payoffs" show up. They indicate rankings with the proviso that higher numerical values indicate higher rankings of alternatives. A game in the full technical sense only emerges after the two actors

[104] A two-by-two game is a two person strategic interaction model in which each of the two actors has exactly two options from which to choose and the payoffs are in utility rather than in monetary or other real terms (like years served in prison). Whenever the payoffs are not utility terms, we speak of game forms.

[105] Ordinal payoffs just give the order of results, which are better or worse than other results without any measure of how much better those results may be.

have considered everything that could influence their evaluations (including the monetary or substantive payoffs but not only those). When considering everything, they, of course, factor in their attitudes towards risk. The full utility index as opposed to an ordinal one shows the ranking of alternatives including attitudes towards risk. Moreover, all the preceding, the table, the utility payoffs etc. must be known to both, and they must know that they each know it. If that is the case, the game as an object of what we will later call and more precisely characterize as "common knowledge" is well-defined as is the process of rational reasoning about it, too.

The term "**strategic form**" is used for tables like the preceding for reasons that will become obvious below (5.).[106] The higher the number, the higher the ranking of the result in the individual's preference order. In view of the rankings and the causal independence of choice options represented in the PD table, it seems quite obvious why, according to conventional wisdom, the only "solution of the game," namely a combination of choices that can be sustained by rational players in reflective equilibrium, amounts to a choice of (D_A, D_B).

By construction, what one player chooses is assumed *by the players themselves* not to (causally) affect the choice of the other (clearly and explicitly stated by relying on 4.1 and 4.2 as separate entities, initially). Regardless of what the other player does, it is for each always better to defect. Since the game is one-shot by construction, there is no future outside the game that could be causally influenced by choices made in the game. Therefore, accepting the principle of intervention and requiring minimal individual opportunism, the alternative that is better than the other one(s) *regardless* of what the co-player chooses should be the rational choice.

This argument should show up within the reflections of rational players in forming reasons for individual choice making. Putting oneself in the shoes of A, the argument is obvious: If the other prisoner, B, confesses, i.e. chooses D_B, it is better for A to confess, i.e. to choose D_A. If the other prisoner, B, does not confess by choosing C_B, it is even better for A to confess by choosing D_A. *Whatever* B does, the best choice for A is to confess. An exactly analogous reasoning holds true for B. He, too, should come to the conclusion that he is better off choosing to confess, regardless.

Therefore, if they are opportunistically rational, both actors will choose to confess and serve their 9 years in jail. However, had both of them chosen not to confess, both could have gotten away with 1 year of prison time each. Had they cooperated with each other rather than with the district attorney, both of them would have been strictly and considerably better off.

[106] It is, in fact, not a full-fledged strategic form since in the table only the individual ranking orders between results show up.

We say that such a combination of choices as (C_A, C_B) leads to a "**payoff dominant**" result, which in the present case is (3,3), as compared to the (D_A, D_B) result. The result is payoff dominant compared with the "**payoff dominated,**" which in the present case is (2, 2), a result brought about by the combination or "**profile**" of choices (D_A, D_B). The payoff dominant result is better for each of the players as compared with the dominated result. However, in reaching the payoff dominant result each of the prisoners would have had to violate another basic principle of rational choice, which is called the principle of "**undominatedness**" or "**non-dominatedness of choice alternatives**." That principle requires a player not to choose a "**strictly dominated choice alternative,**" that is, an alternative to which another exists that makes the chooser strictly better off *regardless*.

Payoff dominance in an interaction with at least two independent choice makers concerns effects beyond the control of any single actor (since none of them can choose a cell rather than merely a row or a column). The choice of alternatives – rows or columns including those with the undominatedness property – falls within the control of a single actor. As far as choice options are concerned C is for both actors strictly dominated by D, for the latter leads to better results regardless of the choices of the other actor (or whatever the other actor does).

In sum, according to their own model of the interaction, actors can choose functions C or D but not a specific value of these functions. If one of the functions leads to worse results for each of its arguments than another function, then that function (representing an option or act) should not be chosen.

The minimally rational actor who has construed the model of the situation in accordance with the principle of intervention understands the preceding. By construction, he thinks of himself as having full control over the choice of alternatives (functions) and no control at all over the states of the world or choices of the other actor (arguments of the functions). Opportunism dictates that he exert his causal control such that he will never choose a strictly dominated alternative.

In what follows, speaking of "(un-)dominatedness" relates to *choices* under the control of a single actor, while, in the case of "payoff dominant" and "payoff dominated" *results*, the qualifying attribute "payoff" will be used.

Undominatedness of the choices made by a single actor seems a rather trivial rationality requirement. Imagine that you are interested only in the size and price of an item and that you prefer lesser size to larger size and lower price to higher price (a mobile telephone for instance). If item X is smaller and cheaper than Y – and, thus, X is better along *both* dimensions of evaluation than Y – how could it be that you would nevertheless choose Y over X? That might happen if you are interested in something other than size and price. Yet, if you are, as assumed, exclusively interested in size and price and prefer smaller to larger as well as cheaper to more expensive items, it could not conceivably be reasonable to choose Y over X if Y is both larger and more expensive (i.e. worse along both dimensions) and *if you, according to your own model of the situation, can causally bring about result X.*

It seems clear that undominatedness of the alternatives actually to be chosen should be regarded as a rationality requirement in any “reflective equilibrium” about what individual rationality “means.” A person who plans to act accordingly would not be minimally rational in the sense of opportunism introduced before, whereas the realization of payoff dominated results by the choices of several *independent actors* is fully compatible with the minimal rationality of each of the actors. Quite to the contrary, as long as we require of rational choice that chosen options should be undominated, it is an implication of rationality that in games, like the classical prisoner’s dilemma game, the payoff dominated equilibrium must emerge as the unique outcome of rational choices.

In sum, minimal rationality requires that the chooser discriminates between what is among the causal consequences of his choice making and what not and requires that the chooser evaluate the choice alternatives in terms of the consequences causally brought about by his choices. Planning on what to do in games in which he *participates*, the minimally rational chooser cannot be in reflective equilibrium if he intends to choose alternatives that are dominated “all things considered” (or are dominated independently of anything that is beyond his causal influence as imagined by himself). The undominatedness of choices is a requirement of minimally rational planning and play if the principles of intervention and opportunism apply.

4.1.2 Some additional concepts

4.1.2.1. Concepts concerning incentives and evaluations

We have spoken of a “**game form**” if the results of the players’ choices in a game – as in table 4.3 – are *not* yet evaluated. There are no rankings that represent the evaluations of individuals *all things considered.* In the ordinal representation of the interaction in table 4.4 attitudes towards risk were not factored in. In that regard, it was a game form and not a game in the full sense of the term. Often ordinal representations are called games nevertheless. This is a little sloppy but not unjustified in that all things – except attitudes towards risk are included.

Before such ordinal evaluations are formed, years in prison could still be evaluated according to criteria other than time served by each prisoner himself. There would be no violation of minimal **rationality** if an individual in a social interaction that is of the *form* of the PD game were to, in his evaluations of results, take into account how his choices would affect the other player. Results are evaluated according to the aims, ends, or values of the actor *whatever these might be*. Therefore, *if* a prisoner has a strong “intrinsic” motivation to act co-operatively towards the companion of his deeds and is willing to pay the price of additional years in prison for having co-operated with his peer by refusing to confess, then the pursuit of such an aim would effect a different ranking among results and possibly justify a different choice.

Means-ends rationality as such does not preclude such pursuits as require considering the well-being of others. The concept of means-ends rationality does not imply that motives must be selfish or may not be other-regarding. It implies only that, *whatever the given aims, ends, or values* are, they should be pursued according to *standards of at least minimal rationality* by the rational actor. No matter what he considers to be relevant dimensions of value, *if* an alternative is dominated by another one according to a personal ranking of states of the world *after all things have been considered*, this dominated alternative can only be chosen in violation of the principle of opportunism (since the latter requires never choosing a dominated alternative). If the aims, ends, or values of an actor are such that he is interested only or at least pre-dominantly after all things have been considered (including how his own actions affect others) in minimizing the years he must serve in prison, then this assumption has certain consequences for what can and what cannot be deemed minimally rational. If he feels otherwise, it again has consequences for what may or may not be conceived of as rational.

If years served in prison by the choice maker himself are – as assumed here – the only evaluative concern of the two actors, then minimally rational behavior dictates bringing about combinations of choices to which another combination exists that could make *both* actors strictly better off simultaneously. Such a payoff dominated result to which another exists whose realization would make everybody better off is also described as (strictly) "**Pareto dominated.**" *A prisoner's dilemma interaction is a dilemma for each prisoner precisely because in this game Pareto dominated results emerge if actors do what is dictated by minimum rationality.*

Economists, like everybody else, tend to require that such situations to which another exists in which everybody concerned could be made strictly better off should be avoided. Accordingly, the "**weak Pareto norm**" or the "**weak Pareto requirement**" dictates that social results *should not* be Pareto dominated in the strong sense that everybody could be made *strictly* better off in at least one alternative social situation. In short, what emerges from our choices in social interaction should be such that we could not make everybody strictly better off by choosing otherwise.

Often the stronger requirement that results of social interaction be such that it is not possible to make at least one better off without making none worse off is accepted as well. Obviously, in this case, not everybody would have a motive to desire a change of the outcome, but none would have a motive to act against such a change while some desire it. The "**strong Pareto norm**" or the "**strong Pareto requirement**" suggests that social situations to which one other exists in which at least one actor could be made better off without making anybody less well off than before should be avoided. In short, gains or improvements for some without imposing losses or costs on others should be brought about. This norm could also be seen as expressing "*minimum beneficence*."

One may note that in the PD game there is – restricting ourselves to the two strategies for each[107] – only one Pareto dominated situation, namely the outcome of minimally rational play. A Pareto dominated situation is also called "**Pareto inefficient.**" In the PD case, the Pareto dominated outcome of minimally rational choices violates the weak as well as (by implication) the strong Pareto requirement. We refer to the underlying criteria as "**Pareto criteria**." If everybody can be made strictly better off by making a change, the "**weak Pareto criterion**" is *not* fulfilled. If one can be made strictly better off while none is less well off, the "**strong Pareto criterion**" is *not* fulfilled. The weak criterion implies a weaker normative requirement since everybody strictly prefers the other situation, while the strong Pareto criterion leads to a more demanding normative requirement. To put it slightly otherwise, the strong criterion classifies more situations as Pareto inefficient. It is, therefore, stronger in the sense that it more frequently suggests that states of the world be avoided.

As far as the normative suggestion that Pareto inefficient situations be avoided is concerned, one should be rather careful. After having considered all value relevant dimensions, the requirement not to use an alternative that is dominated within our own individual sphere of control is an implication of minimal rationality. Not to let Pareto dominated alternatives emerge is not an implication of minimal rationality. For, *nobody can choose the Pareto superior alternatives* (or, for that matter, other results of social *inter*-action) *single-handedly*. The norm that Pareto inefficient results should be avoided is not an implication of the norms of minimal rationality; it is a substantive (moral) norm (that goes beyond the rationality requirements of the principles of intervention and opportunism). That this is so can – if additional illustration is required at all – also be seen by considering the PD game again.

The prisoners' dilemma game shows that helping people to reach Pareto efficient outcomes for themselves may be socially undesirable if the interests of others affected are taken into account. For instance, it is presumably in the interest of the general public that the prisoners both confess and serve their time. So, we cannot assume that Pareto efficiency for each social interaction taken separately is desirable. Likewise, co-operation and the disposition to co-operate are *not* (!) intrinsically good. Individuals may co-operate for "bad" as well as for "good" purposes (as evaluated by some external standard of evaluation).

To give another example, we desire people to compete on a market. To induce them to compete, we try to expose them to PD-like situations. Co-operation by suppliers who, thereby, can act as if they were a single monopolist is seen – by

[107] Only the four so-called "pure strategy combinations" are taken into account.

outsiders – as the "collective bad" of collusion. Finally, all large-scale misdeeds require some form of co-operation and sacrifice by those who jointly commit those acts (see Arendt (1951)).

In sum, solving so-called collective action problems is not a good thing in itself; it rather depends on what the individuals aim at in their concerted actions.

We should also be aware that exploiting incentive structures like the prisoner's dilemma in the pursuit of collective goals like convicting guilty criminals has a down side. Presenting others with such choices is a very dangerous instrument since, in the prisoner's dilemma, confession is rational for any actor who is confronted with it. The incentive to confess is completely independent of whether the actor is in fact guilty or not. An innocent person has the same incentive to confess as a guilty person; in that regard, using the PD is like torture. Since confession as such does not imply the truth, the requirement that the suspects must be proven guilty by evidence other than the confession itself is obviously an indispensable safeguard. To what extent such a safeguard can be effective in view of confessions is an open question though.[108]

Even though the argument for choosing the two undominated strategies seems to be entirely compelling under conditions of minimal rationality, still another reason for singling out the emerging combination of choices as "solution" of the problem of rational play in the PD may be given.

The combination (D_A, D_B) is in fact the only one in which none of the players could do better by choosing the D-alternative *against the given, causally independent, choice of the other*. For instance, if the combination were (D_A, C_B), then it would be better for B to choose D_B. Likewise if the combination were (C_A, D_B), then player A would be better off with D_A leading to (D_A, D_B) if B's choice remained unaltered. Finally, if the situation were (C_A, C_B), then both would be better off if they unilaterally switched to the non-co-operative alternative. As long as the choice of the other is framed as causally independent, which according to the rules of RCM is implied by the tabular representation as such, any choice of a co-operative alternative C should be avoided.[109]

Players who make plans for playing a game like the PD can be unilaterally in *reflective equilibrium only* if they choose a plan that is a **best response** to an assumed behavior of the other player. If players who are interested only in

[108] See for realistic cases from recent American political history Muzzio (1982).

[109] If both players choose the dominant alternative of non-co-operation, D, this amounts to singling out a Pareto dominated result. In fact – confining attention to so-called pure strategies or choices that are made with certainty of that choice – it is the only Pareto inefficient result. The other three "pure" combinations of choices all lead to Pareto efficient results (i.e. neither can be made better off by switching to another alternative). However, none of the three Pareto efficient combinations can be rationally realized if the principle that strictly dominated alternatives should not be chosen is accepted. At the same time, both actors would be simultaneously better off should they both simultaneously violate the fundamental rationality precept not to choose a strictly dominated alternative that is within their full causal control.

reducing their own expected years in prison were to make plans beforehand and consider their own best responses (plans) against assumed behavior of the other player, respectively, they would be **simultaneously** in reflective equilibrium only if *both* of them would simultaneously plan on best responses. In reflective equilibrium, the response planned in reaction to the behavior of the other must be a best response to the best response behavior of the other etc. (If the plans of each were mutually known to both, neither of the players would, in pursuit of his own given aims, ends, or values, have a reason to plan otherwise in view of the "given" plan of the other.)

Plans for a game corresponding to such mutually compatible intra-personal reflective equilibria are called "**equilibrium plans**," and the combinations of choices are called "**equilibria**." Plans for a whole game are called "**strategies;**" therefore, "**strategic equilibria**" are simply combinations or **profiles** of "equilibrium plans."[110] Each of the plans is a "**strategy**" and each of the strategic plans that are part of a **strategic equilibrium (profile)** is called an "**equilibrium strategy**."[111]

We can with some plausibility relate instrumental or "means-ends-rationality" to the *global* or overall outcomes of action. If we do so, we might be tempted to say that the unique equilibrium in undominated strategies, the (D_A, D_B) combination or profile, is not the only rational outcome of a PD. At least, "since morally constrained agents seem to do better than rational agents – say by co-operating in social situations like the prisoner's dilemma – it is difficult to dismiss them as simply irrational." (Danielson (1998), 3). According to a rather wide-spread view, the emergence of *payoff* (!) dominated results must indicate some form of "irrationality," Some eminent theorists (though they normally avoid stating the thesis of a higher order rationality as bluntly as done here) try to associate rationality more closely with the "goodness of consequences" brought about by choices than with principles like that of never choosing a strictly dominated alternative.

4.1.2.2. Concepts concerning information

A classical prisoner's dilemma game with ordinal payoffs can be represented by a "**game tree**." This is the so-called "**extensive game representation**" of the PD. It shows what players can do when deciding on what to do and what they know when deciding on what to do. Since the PD is a game with so-called "**imperfect information**," there is a line between the two encircled capital letters B that indicate an occasion of choice making for player B. Player B cannot distinguish

[110] Profiles are lists representing the strategies of the players as fixed by them in a combination of their plans.

[111] The relationship to deliberation and reflection as leading to reflective equilibrium should be obvious. It will be taken up again in volume 2.

between these two occasions since she does not know how player A will chose when making her own choice. Of course, player A also does *not* know B's choice.

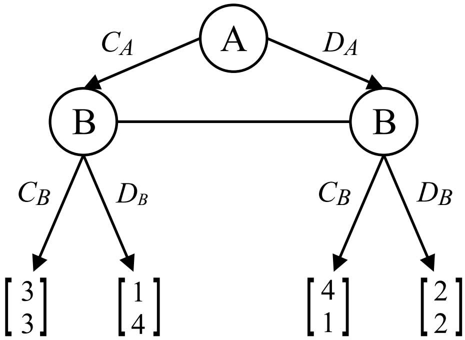
A
C_A
D_A
B
B
C_B
D_B
C_B
D_B
$\begin{bmatrix}3\\3\end{bmatrix}$
$\begin{bmatrix}1\\4\end{bmatrix}$
$\begin{bmatrix}4\\1\end{bmatrix}$
$\begin{bmatrix}2\\2\end{bmatrix}$

Figure 4.1: Classical PD in extensive form

The upper number in the ordered pair at each end-node indicates how the path leading to that end-node is ranked by the first moving actor, A. The lower figure indicates the ranking of that path by the second moving actor, B.

The information is imperfect in that the two actors do not know the choice made by their co-player when they have to make their own choice. Often this fact is explicated in terms of "simultaneous moves." Both players move simultaneously and independently of each other. Therefore, they have no way of knowing what choice the other is making when they make their own. It should be noted, however, that it is not simultaneity in time that matters but rather information flow. Even if one actor has moved a long time ago, the game is one of imperfect information as long as the other actor does not have any chance to know of this move.

Philosophically, it is important to note that, in modeling rational choice making in interactive situations from an internal point of view, it is conventionally assumed that information flow is all that matters. During a play of the game, no choice of one of the players can exert a direct causal influence on the choice making of the other one by ways other than information flow in the game. Moreover, the game (tree) is "**common knowledge**," i.e. each of the actors knows it and each knows that each knows it and knows that he knows that he knows etc.[112] Since the game tree is common knowledge, no "new" information concerning the tree or the rules of the game can emerge during a play of the game.

[112] See again Binmore (1992) for an excellent treatment, most useful for the purposes at hand, though slightly advanced.

To express the assumption that the game is common knowledge, the term "**complete information**" is used.[113] In a sense, all well-defined games are games of complete information. To put it in slightly other terms, if complete information does not prevail, then the game is not well-defined (at least not as an object of classical game theory). In such cases, RCM provides fictitious moves as modeling tools. By these moves, the lack of information and common knowledge is explicitly modeled, and then, heroically, it is assumed that the explicit model emerging from that trick is common knowledge.

Complete information, which refers to knowledge of the game tree, and perfect information, which refers to knowledge of what happens within a play of the game, i.e. when moving through the tree, must be distinguished. To clarify this, it is useful to approach it by discussing our specific example in a somewhat more formal and, at the same time, more elementary vein. Note first, that any set of instances of choice that are indistinguishable for the choice maker is called an "**information set**." For example, the option of choosing C_B after a choice of D_A by A is indistinguishably the same as choosing C_B after a choice of C_A by A in the case of figure 4.1. (That was the reason to connect the two indistinguishable nodes with a line.)

The encircled nodes refer to instances of choice making by the actor who is indicated in the circle; the branches represent choice options and, thus, choices that can be made at the node from which they originate. Plays of the game start, of course, at the root node and then end with results that are ranked by the players A and B according to the ranking indicators shown at the end of the branches of the tree.

Observe also, first, that the choices belonging to the same information set must always show the same set of alternatives; otherwise, they would not be indistinguishable (i.e. they could be distinguished by the differences in the sets of alternatives originating in them). Second, observe that the game representation in the case at hand is completely equivalent as far as strategic and information aspects are concerned with the representation that would emerge after substituting A by B and B each time by A.[114]

We can now slightly modify the information conditions of the situation. Let us assume that the second player, B, can distinguish between his choice making after a co-operative first move C_A from his choice making after the defection choice, D_A. Strictly speaking, B now has four rather than merely two distinguishable and, in that sense, different options. He can choose one of two alternatives after each of the preceding choices of A. It may be that the overt actions or bodily movements of, say, second-mover co-operation are the same

[113] See below for an illustration of the concept by means of a more specific example.

[114] In game experiments this may be different since there are positional order effects to be observed.

after first-mover cooperation as after first-mover defection.[115] However, regardless of being the same overt movements, the acts thereby performed are different.

If the second mover knows what the first did, we say that he has **perfect information** about what the first mover did. Accordingly, a game is a "**game of perfect information**" if *all* information sets are singletons. It is a "**game of imperfect information**" if at least one non-singleton information set exists.

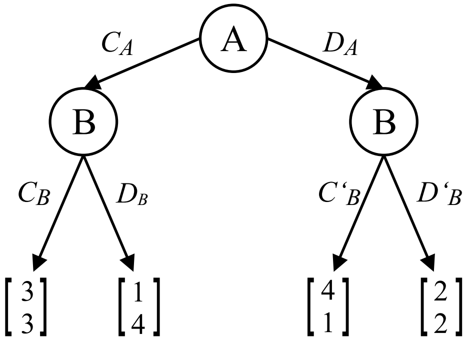
A
C_A
D_A
B
B
C_B
D_B
C'_B
D'_B
$\begin{bmatrix}3\\3\end{bmatrix}$
$\begin{bmatrix}1\\4\end{bmatrix}$
$\begin{bmatrix}4\\1\end{bmatrix}$
$\begin{bmatrix}2\\2\end{bmatrix}$

Figure 4.2: Perfect information PD

In the preceding graph, we distinguish between C_B and C_B' as well as D_B and D_B'. These options are distinguishable by B according to their position in the sequence of acts. From a common sense point of view, the distinction between acts according to their place in a sequence of acts is quite clear. To co-operate after co-operation has a different subjective or symbolic meaning for the actor than to co-operate after defection.

To illustrate, imagine that you promise to deliver your apples to my door tomorrow, and I promise to deliver my oranges to your door a month hence.[116] A month hence my overt actions of delivering the oranges – pushing the cart etc. – are the same whether you have delivered the apples or not. However, if I am bringing those oranges after you did indeed deliver your apples, I am *fairly retributing* to your execution of your part of the bargain. There is no uncertainty whether you will execute your promise. If I deliver the oranges even though I know with certainty that you did not keep your promise, I am not retributing in kind. I am holding out my other cheek; I am committing a "supererogatory" (a "saintly") act that goes beyond moral duty rather than fulfilling my moral duty (see on supererogation for instance Heyd (1982)).

[115] As we shall see immediately below, the *meaning* of the two choices may be completely different. As far as "social meaning" is concerned, it is different to cooperate after defection from cooperating after cooperation. The first act is generally seen as more or less saintly, while the second seems merely fair.

[116] For further illustration the reader might want to go back to chapter 1.

It may be noted here that Hobbes' view that the certainty of receiving or not receiving what has been promised may be interpreted as leading to different evaluations of the branches of the tree and, therefore, to a different ordering of results under this consideration. Not to deliver, which according to Hobbes, is always justifiable as long as uncertainty and the dominance in material or substantive payoffs prevails, is not anymore justified in case of certainty of a first mover's execution of a promise:

> "For the question is not of promises mutuall, where there is no security of performance on either side; as when there is no Civill Power erected over the parties promising; for such promises are no Covenants: But either *where one of the parties has performed already*; or where there is a Power to make him performe; there is the question whether it be against reason, that is, against the benefit of the other to performe, or not. And I say it is not against reason." (italics added) Hobbes (1651/1968) chap. 15, 204

To "reciprocate in kind" is not as absurd as using C'_B after D_A. Nevertheless, it is not in line with future directed opportunity taking behavior. Still, Hobbes, quite contrary to the Spinozist logic of future directed rational choice, argues that a second mover would *not* act against reason if using C_B after C_A. I believe that Hobbes felt that the "Logic of Leviathan"[117] in its Spinozist reading is too radical. Hobbes saw that some restrictions of the approach were needed. Choice cannot be all forward-looking, opportunistic, and, thus, concerned only with what is "handiest at the time" if we are to understand social order adequately.

Yet, the good common sense of Hobbes notwithstanding, a tension remains between the Hobbesian "natural obligation" to preserve one's life as best as possible and fulfilling a "promise" as second mover.

We might say that, according to Hobbes, in a PD-like game form with perfect information, the ordinal ranking, all things considered (including moral considerations), of unilateral non-performance after the other has performed and is known to have performed is or should be subjectively ranked lower than performance according to the promise. If such re-evaluation of objective results by the second mover did not only prevail but was also known to the first mover, then achieving Pareto efficient outcomes would be guaranteed and the original problem would be gone.[118]

We will come back to this central point quite extensively in the further discussion of extensive game representations (in section 5.1.). For the time being, we will focus on the more conventional tabular representations of choice making in interactive situations.

[117] Of course, there has been a book of that title, see Gauthier (1969))

[118] See more on commitments and their role in particular in trust problems and also the AG-Game below.

4.2 Some elementary games in strategic form

Although games are not the only models of moral science, they are chief examples of such models. Neglecting the sequence of moves for the time being and looking at the strategic plans only, we can form simple tabular representations of interactions that show how combinations of plans map into the payoffs expected to result from the execution of the plans.

In the simple case of the PD, the table shows the following functional assignments "f " of lists to the strategy combinations:[119]

$$(3, 3) = f(C_A, C_B),\ (4, 1) = f(D_A, C_B),\ (1, 4) = f(C_A, D_B),\ (2, 2) = f(D_A, D_B)$$

We call such representations of interactions **games (or game forms) in strategic form**.

Certain elementary two-by-two games in strategic form are of particular importance since they are paradigms of the most elementary kinds of social interaction. Often we know intuitively what these games are about from our own experience. However, it is helpful to have simple and precise models of them in our conceptual tool box. Such elementary models of moral science form the elements of which much of the world of the social theorist is made. This, rather than the fact that they show up in real life in pure form – which they in all likelihood do not – is the reason why they are introduced subsequently for those who have not been exposed to this kind of elementary game theory. (Others might want to go straight to chapter 5.)

4.2.1 Two-by-two game paradigms

Let us start with a "**prisoner's dilemma**" game form in *monetary terms* rather than in years in prison.[120] Consider for the sake of specificity the following table of results or outcomes of action choices of two individuals:

player 2 player 1	C_2	D_2
C_1	\$ 75, \$ 75	\$ 0, \$ 100
D_1	\$ 100, \$ 0	\$ 25, \$ 25

Table 4.5: PD game form in monetary payoffs

[119] Of course, f may itself be a tuple of functions like the v_r, by which the players r evaluate the results of play according to the strategic plans showing up in the strategy combination.

[120] In the appetizer section, it was fittingly in beer and steaks, in between for the vegetarians, in apples and oranges, and now in money.

If we assume that the rules of interpretation of standard game theoretic models apply, then this table must be interpreted as construed according to the principle of intervention. If read appropriately, it reflects not only monetary payoff expectations under alternative constellations; it also expresses the view of the players that their moves are causally independent. To re-iterate, this "reading of the table" is so due to the rules of the *language* of representation (i.e. RCM). It is not due to facts (of the represented part) of the world.[121] If the facts do not correspond, then the model is misspecified.

It is not true that there are deeper factual, empirical or metaphysical, reasons why causal independence must prevail. The actions of the two players could conceivably be related in a (possibly very complicated way) by some common cause.[122] If the table were to be read as exclusively representing the payoff interdependence between the actions of the two actors, then it would even be possible that the act of the one could causally bring about that of the other actor. However, the table *as used in eductive game analysis* does *not* merely form a representation of the (monetary) payoff functions of the two players; it is a stylized account of social interaction from the participant's point of view rather than an objective or external point of view. As such, it implicitly expresses certain assumptions about causal relations, too.

If we reject the assumption of causal independence of actions, then we should not use the tabular representation of classical game theory, or if we want to use such tables, we should embed them in a theory other than classical game theory such as to allow for another interpretation. Yet, if the semantics of RCM are used to read the table, then this is how we must interpret it.

In sum, if tables like the preceding are used in standard game theory to depict how actors reason about social interactions, then they do not only ascribe knowledge of the payoff interdependence to the actors; the tables also ascribe to the actors the conviction that their own actions are causally independent of the actions of the other actor. According to RCM, actors, who as participants commonly know such tabular representations, also know that both are construing the interaction from an internal point of view and that they imagine themselves in the act of choice making as independent origins of a causal chain.

Under this standard interpretation, it does not make sense to speculate about causal dependencies from the actors' internal point of view. However, independent of the specific meaning of our modeling devices, substantive issues about whether there can or will be certain dependencies between action choices

[121] If we formulate models from a classical game theoretic point of view, it cannot be avoided that the signs we use mean certain things. And, within a classical perspective, the tabular representation means that we ascribe to the players the view that the choices of columns and rows, respectively, do not directly affect each other.

[122] I am indebted to Max Albert and Ron Heiner for insisting on that conceptual point though I do not agree with Ron's more far-reaching conclusions from this.

are open. They must be decided on grounds other than the rules of some language. If we come to the conclusion that, for instance, the acts of others are contingent on our own, then we must try to represent that fact by other language tools than provided by the simple bi-matrix of our tabular representation.[123]

However, for the time being the simple tools of matrix representations in RCM suffice. Using them we can look at paradigm examples other than the PD.

4.2.1.1. Assurance game

The following game is called the "**assurance game**," or "AG game" for short, since the Pareto efficient equilibrium result – and it has only one Pareto efficient outcome at all, which happens to be an equilibrium – emerges if actors feel assured that the corresponding equilibrium strategy is played by their co-player:

player 2 player 1	C_2	D_2
C_1	4, 4	1, 3
D_1	3, 1	2, 2

Table 4.6: Assurance Game (AG)

This game emerges if in the ordering of results of the PD game form the top ranking unilateral deviation from co-operation and the mutual co-operation result switch places. Both players then rank mutual co-operation higher than unilateral defection or exploitation.

The name of the game derives from the fact that, as compared to the PD game, there are two equilibria now and players need to be assured that the option leading to the one rather than to the other is chosen by their co-player. To put it slightly otherwise, the situation is such that if each were assured that their co-player would play the C option, each would have an incentive to play the C option as well. Expecting the other to play C is a reason to play C and, after doing so, there is no reason to regret what has been done (i.e. this is an equilibrium).

One might want to say here that in the AG game, it is almost trivial that the Pareto-efficient equilibrium will be the outcome of individually rational choices. However, it should be noted that the expectation that the other is going to play C is the *only* reason to play C oneself. Conversely, if the other is expected to play D, this is a good reason to play D. As long as both are assured that the other will play C, both will reach the Pareto superior of the two equilibria of the game. However, if they are not assured of this, the *only* reason to choose one or the other

[123] As we have seen before, the very notion of a strategy as a plan is "loaded" with assumptions about causality and, therefore, requires that we very carefully observe the principle of intervention and its implications. For the time being, we start with examples in which there are only moves, and, thus, the problem of not being able to choose a whole strategy as a move (unless there is a way to causally effect such a choice as a commitment) is avoided.

alternative is the expectation that the other player will choose the corresponding one.

How could assurance game rankings emerge? Think of the preceding PD game form of table 4.5. People who are, for example, confronted with monetary payoffs conforming to the ranking order of such a PD game may often think that they "should" co-operate. They would in fact rank mutual co-operation above unilateral defection since they evaluate results not only in monetary terms. They are intrinsically motivated to bring about the Pareto efficient solution dominating the equilibrium solution in the monetary PD. They (re-)evaluate the results in terms other than the money they personally receive. They think so highly of mutual co-operation that they rank it above unilateral deviation (the monetary temptation to the contrary notwithstanding). At the same time, they resent being exploited unilaterally and, therefore, would prefer to deviate themselves if their co-player deviates. What they need is some assurance that they are both not going to give in to the monetary temptations of the interaction represented in monetary terms by the PD game form. So, their "true" ranking order in the monetary PD-like structure has "all things considered" become that of an AG game.

The subjective re-evaluation of the objective monetary terms renders the (C, C) outcome an equilibrium in evaluation (with a monetary payoff of $75 for both) but by no means the only rational outcome of the AG game. The latter would happen, though, if either of the two players were able to pre-commit. If one is bound to C, and this is known to the other, then this commitment makes C the only rational response for the other. Of course, both commitment and knowledge of that commitment must prevail to assure this result.

4.2.1.2. Chicken

The game "**chicken**" can also be derived from an interaction that can be represented in monetary terms as a PD game form by a single change of position of results in the rank orderings of the two players. Now it is the lowest and the second lowest ranking result that change places. The worst situation is now actually mutual defection.

player 2 player 1	C_2	D_2
C_1	3, 3	2, 4
D_1	4, 2	1, 1

Table 4.7: Chicken Game (CG)

One story used to introduce the game is that of two cars racing towards each other. The driver of the car who swerves out of the way before the other swerves is a “chicken”.[124] If both drivers swerve at the same time, neither is a chicken. Yet, both drivers would like to be the one holding out and making the other a chicken – if there were not the risk of a real “big bang.”

It is also instructive to compare CG and PD with respect to certain other interpretations. The PD is a kind of rough model for a “cold war.” It models an arms race in which both players would gain if neither invested in armament. At the same time, it is better to insure oneself against the armament of the other by one’s own defection from disarmament policies. As compared to that, “chicken” models “hot war.” It models the “rather red than dead” preference that was alluded to in cold war times when many people were expressing the view that to give in unilaterally to the Russians was still better than to die in defending one’s turf.

In the last resort, it is only the expectation of what the other will do that can justify a choice in chicken. Again, if either of the two players were able to pre-commit and credibly communicate the commitment, one of the two specific results on the off-diagonal would be determined. The player who could pre-commit would go for the action leading to his most preferred result leaving the other no choice but to accept the ultimatum and to give in. If one of the two actors were bound to D, it would render C the only rational response for the other one. The corresponding equilibrium would emerge. One actor would receive the result with the rank number 2 and the other the one with the rank number 4. At least both would avoid the worst possible outcome with certainty.

4.2.1.3. Battle of the sexes

The “**battle of the sexes**” game, or BS[125] game for short, is also a game with two equilibria. There is no natural way to relate this game directly to a one-shot PD game form. In this game, the coordinative aspects are even stronger than in the AG or the CG game.

player 2 player 1	B_2	K_2
B_1	1, 2	0, 0
K_1	0, 0	2, 1

Table 4.8: Battle of the Sexes (BS)

In BS, to co-ordinate on one of the equilibria is always better than no co-ordination, regardless of which is chosen. Think of a man and a woman who discuss where they want to spend the evening together over the phone. For both of

[124] Very nicely illustrated in the part of the Walt Disney movie “Herbie” where the spectators of the cars racing towards each other can raise card-boards with the word “chicken” if one swerves unilaterally

[125] No pun intended. However, the game is at least crab grass in the game of life of the game theorist.

them, the first priority is that they spend the evening together. Secondarily, the female would prefer to see the boxing match while the male would rather go to the cinema; however, for both the most important thing is to go together. From the phone conversation, complete information of the situation emerges as presented, but the phone is disconnected before they make a decision.

Where should they go in the evening? Each of them understands in his or her planning that the only reason why she or he should plan on either alternative is that the other does indeed plan to choose that way. However, if, beyond common knowledge of the game, they do not have any additional knowledge about what the other actor intends to do, they have no reason to prefer one act over the other. Their orderings of the end results and their common knowledge of that ordering do not provide a reason for action to them.

Again, commitment power could solve the problem. If commitment power were bestowed on exactly one of the two and if the commitment could be communicated beforehand, the other would know what to do and one of the diagonal results would emerge.

4.2.1.4. Pure coordination

Pure co-ordination or the “**right or left game**,” the RL game, emerges if there is no conflict about which of the equilibria is to be chosen.

player 2 player 1	R_2	L_2
R_1	1, 1	0, 0
L_1	0, 0	1, 1

Table 4.9: Right or Left (RL)

Think of drivers on an island where there have never been any cars before. There are two cars and two drivers. They need to decide on whether they are going to establish driving on the left or right. What the best choice is completely depends on what the other driver does. Neither cares about *which* side they drive on, as long as it is the same side that the other uses. The problem is one of co-ordination such that “unfriendly” encounters on the same side of the street are avoided. There is no conflict of interest but still a problem of common knowledge here.[126] Again, a unilateral commitment and communication to use one side of the street would solve the problem.

4.2.1.5. Pure conflict: Matching pennies

“**Matching pennies**” emerges if two people play a game of doing just that. The rules are quite simple. At some signal, both players simultaneously drop a penny

[126] See on this, in rather elaborate ways, Lewis (1969) and Rubinstein (1989).

on the table. If the pennies match, i.e. are both heads or both tails, then player 1 loses his penny and player 2 gets both, thus wining an additional penny. If the pennies do not match, player 1 wins an additional penny while player 2 loses his. Quite obviously, the two individuals are not only matching pennies, they are matched against each other such that the one can only win what the other loses. The situation is a game of pure conflict in which no co-operative co-ordination of strategies is possible. There is no equilibrium in so-called “pure” strategies here.

player 2 player 1	K2	Z2
K1	–1, 1	1, –1
Z1	1, –1	–1, 1

Table 4.10: Matching pennies

Yet, there is an equilibrium in so-called “**mixed strategies**”. There is more to be said about the very concept of a mixed strategy than is suitable at present. However, the reader may simply think of each of the two players choosing a probability distribution over the two moves. In that case, so much would be obvious: The reason for the behavior of a player must be the expectation of what the other would do. Now, if one of the two expects the other one to play one of the alternatives with higher probabilities, then he should completely switch to playing the alternative with the then higher expectation all the time. If that were the case, another one expecting this should switch as well etc. The expectations can lead to plans that in turn do not imply a reason to change the plan only if the choices of each of the actors are unforeseeable and lead to the same expectation of gain or loss no matter what.

Under the parameters presented, that can only be the case if each of the players assigns exactly the same probability to both of the option choices of the other one. They could not do better against the behavior of each other than they do if they both choose each of their options with probability one half and expect the other one to do so with the same probability.

Two things should be noted here: 1. It is sufficient that each of the actors expects the other one to choose with equal probability each of the alternatives. 2. If the actors did not choose their alternatives but rather threw a fair coin to get equal probability, then they would in fact play a game with three rather than two choices: put head on, put tails on, throw the coin and then act accordingly. Furthermore, in a complete model of the game, all three alternatives would have to show up. In this sense, the mixing of strategies is possible only in the eye of the beholder; otherwise, the game would have a different nature and would have to be modeled differently.

4.2.2 N-person paradigms

Though many encounters of our social life are two-by-two in some sense, they are often not in others. There are more than two actions from which to choose, and there may – as we have seen already – be a sequence of moves. Another essential generalization emerges if there are more than two actors. In that case, a simple matrix representation might not do, yet there are ways to get around that problem and to represent the interaction of many in matrix form as well. It is possible in particular if the game is "symmetric" in the sense that for every individual r from a set of N+1 individuals, the game played with the N other players looks exactly the same. We must then just let one arbitrary, though specific, individual play as "representative agent" for all others.

4.2.2.1. N-person PD

Consider the following two-by-N table for a game played among N+1, N+1>2, individuals in which the N+1-th individual plays as a row player and the others serve as column players. It is only shown what the payoff to the one representative individual will be since all others symmetrically have to decide as she does (the presentation follows Buchanan (1965)).

Number of other players cooperating Player N+1	0 other Cooperators	About N/2 other Cooperators	N other Cooperators
C_{N+1}	a	m	y
D_{N+1}	b	n	z

Table 5.11: N+1-person prisoner's dilemma game form, N+1-PD

Assume that the interaction is characterized by monetary payoffs that fulfill z>y>n>m>b>a for each and every individual. Moreover, let us also assume that for all numbers of other cooperators the C is always smaller than the D payoff for the player who is playing as row. Assume also that the individuals are only interested in gaining a monetary payoff as represented by $\{z,y,n,m,b,a\}$ that is as high as possible.

Observe that all would be better off in a state of universal co-operation – a state in which everybody would receive y – than in a state of universal non-co-operation in which everybody would receive b since b<y. Assume also, that for any number of other co-operators in-between, i.e. 1 to N-1 others cooperating, it is always better for each player not to co-operate herself rather than to co-operate. Therefore, non-co-operative choices form a dominant strategy for each and every individual. Since it is not assumed that actions of any individual can influence the choices of any of the others, it is obvious that all individuals should plan on choosing the defection alternative. The outcome of this is a dominant strategy

equilibrium that is Pareto dominated. In the game, all will rationally try to be "**free-riders**" on the efforts of the others; at the same time, all would be better off should nobody take a free-ride. Still free-riding is a dominant strategy.

The story to be told here cannot be told better than by David Hume and so, instead of telling any story myself, I let the master speak (Hume (1739/1978), book 3, chap. 7):

> Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each others mind; and each must perceive, that the immediate consequence of his failing in his part, is, the abandoning the whole project. But 'tis very difficult, and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expence, and wou'd lay the whole burden on others.

As we all understand this is the paradigm situation of a so-called "**collective good problem**" in a large group.[127] It is well known to us from our experience with such matters as environmental protection and the like. Even if all intend to do good, they all would have the suspicion that they might become "suckers" in the "lawlessness" of others. They may desire some kind of "assurance" here to get to a state of affairs in which almost all co-operate.

There may be saints and heroes (see Urmson (1958)) in our world who, in situations like the N-PD, might want to co-operate regardless. There may be people who on rare occasion may be willing to co-operate even if large numbers of other potentially free-riding individuals are involved while their own acts are as insignificant for the collectively perceived result as a single grain of sand on a heap. However, in particular if we are participating in such interactions again and again, most of us will expect that co-operation will deteriorate. As shown in the field as well as in experiments it will wash out over the long haul.[128]

4.2.2.2. N-person volunteer's dilemma

Consider a game in which each of N+1 individuals can provide a collective good all on her own at fixed cost c>0. One individual is sufficient. If none does what is required, then the result is bad for all, for it yields "0". If one volunteers, this individual expects U–c, while all others, if they can let her do the job alone, would get U. Assume U>U–c>0. It would indeed be pure waste to have more than one individual spend c, yet if there is no way to determine who is going to be the

[127] For the time being, it should suffice to point out some excellent literature, Olson (1965), Taylor (1976), Taylor and Ward (1982), Taylor (1987), de Jasay (1995) and in an applied way Ostrom (1990).

[128] This is one of the most robust findings of experimental economics and already well documented in Davis and Holt (1993).

volunteer, for example, by lot, which would amount to adding an additional option to the interaction, then each might speculate on somebody else providing the collective good.

Number of other players cooperating Player N+1	0 other Cooperators	About N/2 other Cooperators	N other Cooperators
C_{N+1}	U–c	U–c	U–c
D_{N+1}	0	U	U

Table 4.12: Volunteer's dilemma

In real life such situations are not as rare as one might think. For instance, one person may be sufficient for getting the information out that the emperor has no clothes on, but one must indeed speak out. There may even be situations like the famous case of Kitty Genovese in which many individuals witness a crime that they could prevent but knowing that there are others who could do so as well, each and every individual speculates that somebody else will incur the risk or cost etc.[129]

A more amusing story may be just as well suited. For penguins, it is risky to jump into the water since a leopard seal may be waiting there for a juicy penguin "burger." So, one of the penguins must literally test the waters in the morning. Penguins have to eat and, therefore, have to get into the water eventually. On their own, they would prefer to jump in, even at the risk of being eaten. However, it is better if somebody else volunteers. So, all the penguins become very polite to each other, indicating as true gentlemanly fashion to each other: "after you." Since volunteers are in scarce supply, they shove politely along the edge of the ice until one of them accidentally falls in. All watch to see whether or not that "fallen comrade" will reappear. If the penguin has not "fallen into a trap," all are happy to jump in after him; if so, they will wait a bit longer or try to get in at a somewhat different spot.

Now, we are not necessarily interested in the behavior of penguins. The film "Happy Feet" notwithstanding, it is not too easy for us to identify with them directly; however, we should note that the example may give rise to a few interesting speculations concerning our own species. The penguins may push each other physically, and humans may do so by some kind of group pressure. They can focus their applause on the volunteers, and the free-rider flipped over is indeed the zealot (see Coleman (1988)). There may also be a whisper in us that makes us prone to volunteer.

[129] See on the crime story Frank (1988) and on volunteer's dilemma more generally Brennan and Lomasky (1984), Diekmann (1985)).

In this context, it may be worthwhile to note that a penguin might have an incentive to jump in if it were to his advantage that individuals from the other sex observe whether or not he is volunteering. If survival of such volunteering actions is differentially related to good judgment – the wise male penguin jumps in only if he has a fair chance to get out living – and physical fitness – the better and faster swimmers among the penguins can afford some jumps that would be lethal for others – then females with a penchant for heroes may make a better choice of mating partners and, thereby, have better and more progeny. In turn, it may reward the male penguins to be chosen that way by females with the right capacities (see on this Zahavi and Zahavi (1997), Zahavi (1975)).

Of course, it would be foolish to generalize too quickly from the animal kingdom to the world of humans, but remarks like the preceding are in order to warn even at the present stage of the argument those who tend to think that only selfish behavior could have been selected for in evolution. An evaluation function that is much more complex may have survived as well.

In sum, we should not infer too swiftly from dominance as measured in substantive payoffs what will be the case in subjective terms of evaluation. Moreover, in nature, it is not clear that only the subjectively selfish can survive, and it may be no different when it comes to the human species.

5 Plan, Play and the Limits of Rational Choice

When introducing rationality, we can either emphasize aspects that are objectively advantageous or subjectively appear so. Doing the first we tend to “explain” behavior in terms of its contribution to objective success of those showing that behavior. However, this objective notion of rationality applies not only to human individuals but also to other animals and to supra-individual entities like firms. Though drawing attention to an important feature of life it does not capture anything specific for members of our species. If we intend to weave our conceptual net such as to catch what is characteristic about the opportunity seeking behavior of individuals of our species, then we should emphasize the subjective side of human choice making as accessible from a participant’s point of view. Accordingly I believe that the focus should be on the advanced human ability to discriminate between what is and what is not a direct causal effect of an individual’s actions as perceived on the basis of “mental models” of the situation. This is what humans can do distinctively better than other animals (which like us all have a “survival related interest” that their situation be “improved”).

If we set up our conceptual framework this way, something can be a violation of rationality and, nevertheless, serve our interest (e.g. if both prisoner’s co-operate by choosing a dominated strategy in a PD they serve their interests). To me the fact that we are conceptually forced to admit contradictions between rationality and interest seems advantageous. If the concept of rationality is an independent item in our conceptual net, then it is in fact desirable for it not to be too closely associated with the concept of interest and that of serving our own good. After all, the main reason for using separate concepts is that they refer to separate things. Therefore, “rational and against interest” as well as “rational and in line with interest” should be allowed for.

When it comes to the specific aspects of being human, the focus should be on the principle of intervention rather than on fulfilling interests. The ability to form models of the action situation that comply with the principle of intervention is the distinctively human faculty and therefore, should be emphasized in idealizations of human rationality. This is exactly what we will do next.

5.1 Plan vs. play

5.1.1 Basic concepts

"**Strategies**" are *complete* plans or lists of intended actions for a game. They specify an action-plan for *each* contingency that might arise in any play of a game under consideration. As opposed to strategies, "**moves**" refer to acts that can be chosen. Moves can actually be executed, not just planned. In other words, "moves" are acts of choice that can be causally effected by the player in an instance of choice.

In the original tabular representation of a PD, the difference between a strategy and a move does not show up. The failure to make the distinction has led to serious confusions. The sequential PD-like game with perfect information of figure 4.2 may be used to illustrate the relevant issues. For convenience, the figure is reprinted here again:

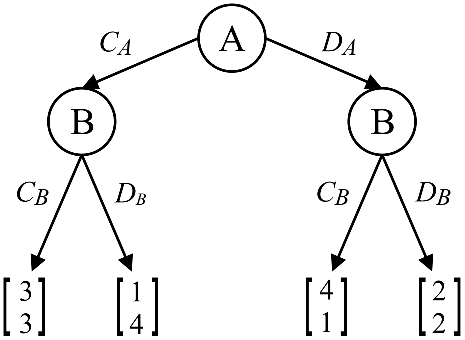
A
C_A
D_A
B
B
C_B
D_B
C_B
D_B
$\begin{bmatrix}3\\3\end{bmatrix}$
$\begin{bmatrix}1\\4\end{bmatrix}$
$\begin{bmatrix}4\\1\end{bmatrix}$
$\begin{bmatrix}2\\2\end{bmatrix}$

Figure 5.1: PD-like game in which perfect information prevails

For the sake of simplicity, we will not symbolically distinguish anymore between co-operation after defection and co-operation after co-operation (i.e. leave out the "primes" showing up in 4.2), even though that would be more precise.

In the prisoner's dilemma-like game with perfect information presented in figure 5.1, all information sets are singletons. To form a *strategy*, the player B must make a *plan* for each singleton set in which he may be called to move. That is, the plan is made for *both* situations (both singleton sets) in which he *may* have to *move*. To form a strategy for the game, B must specify what he *plans* to do (how to move) after C_A as well as after D_A. Though he may never be called upon to act according to one part of his plan since the play of the game may take the other course, he needs to make a plan for both (all) contingencies. This is required by the concept of a strategy as a *complete plan for any contingency that may arise in the game*.

In sum, and somewhat more technically speaking, a strategy specifies one "move" or choice for any information set of a planner/player. It is a function from the set of information sets into the set of options. The function assigns to each information set the options that are available at that information set.

In considering (full) strategies, one should be careful to see that *planning* to make a choice and *making* that choice – executing the plan – are totally different things. The planning is, so to say, "in the mind," while the actual choice *making* is "in the world."

In sum, a plan to move is not a move.[130]

To the extent that game theory is performed from a participant's point of view *in terms of reasoning*, it is all about planning rather than acting (making moves). To plan a choice in the future, say for tomorrow, merely amounts to now forming the *intention* to make that choice tomorrow; it does not amount to making that choice now. To plan to move is different from actually moving. RCM, if properly understood, forces us to make that distinction explicitly.

If plans of action cannot only be used for forming intentions but rather be chosen in the proper sense of that term, then, according to the principle of intervention, these additional options of exerting a causal influence must be taken into account explicitly in any RCM model as separate choices.

Behind the preceding requirement is a tacit requirement that may be named the **explicitness condition**. It states that what is not explicitly modeled does not exist according to the semantics of RCM. Moves that are not explicitly represented in the game model are, thus, assumed not to exist for the purposes of the model so formulated. Since this rule of *interpretation* does not rule out modeling explicitly whatever we want to be included in a model, it forms no strong restriction (if at all) on what can be expressed within a model in terms of RCM. As long as we introduce our assumptions explicitly, RCM (as opposed to specific versions of RCT) does not keep us from doing so.

In sum, the condition of "explicitness" requires that everything that is relevant to planning the behavior to be performed in an interaction is actually "on paper." If it plays a role, then it must show up in the tree. If it is not explicitly stated, then it is assumed to not exist by the rules of the language of game trees.

The explicitness presumption is a consequence of speaking in terms of RCM and not of any hypotheses concerning the real world. It is one of the two

[130] The dual world conceptions are again mirrored here. For instance, in Kantian terms, the "noumenal" plan and the "phenomenal" action; see above chapter 2. If the reader feels that I am beating the plan vs. move distinction to death here he might be assured that this beating is performed for a good reason: there has been too much confusion on this issue.

characteristic assumptions of so-called **non-co-operative** modeling: Explicitness and common knowledge of what is made explicit.[131]

The second assumption of non-co-operative game modeling from an internal point of view has been mentioned already. It concerns what the choice makers or players know. If a tree is meant to be a representation of a game in extensive form, it is presupposed that not only the external analyst knows the tree as presented; *the actors themselves know that tree.* They not only have it on their minds, they also assume that each actor has it on her or his mind. Moreover, in order to get classical game theoretic analysis going, it is necessary that each knows that each knows the tree and knows that each knows it and so on.[132]

As an object of reasoning, a non-co-operative game is not well-defined without the explicitness and the common knowledge assumption. Only with a common object of reasoning – with complete information – can we hope to characterize what the process of planning in terms of reasoning about knowledge among the players might be from their various participants' points of view.[133] Once a well-defined object of analysis exists, we can start to analyze what is on the players' minds when they reflect on the social interaction represented as a mental object or a game.

5.1.2 Strategic planning illustrated

To understand the character of game theoretic analysis as a discussion of planning how to play, go back to the PD-like game of the preceding extremely simple example. An actor in the role of a second mover in such a prisoner's dilemma-like game with perfect information (as in figure 5.1) can make four plans, each plan comprising a move for each of the two information sets that can be reached in the PD-like game with perfect information.

Planning for the extensive game tree is not the same as choosing (moving) at instances in the tree. Planning is part of reasoning about and *not* part of action in the tree. Confusion emerges if plan and play are not distinguished as carefully as they should.

For instance, a tabular representation of a PD game. This "strategic form" of a PD is often interpreted as representing actions by rows and columns. In this reading, columns and rows of a game matrix refer to *actions*. As long as each

[131] As opposed to non-co-operative, co-operative modeling allows for restrictions on opportunistically rational choice making that are not explicitly modeled but are understood as part of the rules of interpretation of the model.

[132] The common knowledge assumption must apply on some level of analysis after all that players do *not* commonly know has been modeled as imperfect information about some "*fictitious* move of nature." (Not going into the Harsanyi device of modeling such ignorance, see Harsanyi (1967-8), we must let it rest with that for the time being.)

[133] See on reasoning about knowledge in general again Fagin et al. (1995)

player can, as in the original PD, only move once, no confusion between a move and a sequence of moves can arise. However, now imagine an actor in the role of a second mover in a prisoner's dilemma-like game with perfect information. He has two moves that he can make. Accordingly, he can form several strategies or full plans.

To see what is involved, assume that the player has the option of actually *choosing* the behavioral *program.* He does not merely make a plan but rather chooses a compulsory choreography for a sequence of actions to be performed by himself like a programmer does for the computer when writing a computer program. The program determines the actions of the actor who chooses the behavioral program – like the program of the computer – for each contingency.

The assumption that the actor can choose that way may possibly be correct. Yet, if it is, the following two-by-four (rather than the former two-by-two) table forms an adequate representation of the strategic situation (again leaving out the primes which indicate a move after a first-mover defection).

Actor B Actor A	C_B/C_A C_B/D_A	C_B/C_A D_B/D_A	D_B/C_A C_B/D_A	D_B/C_A D_B/D_A
C_A	3, 3	3, 3	1, 4	1, 4
D_A	4, 1	2, 2	4, 1	2, 2

Table 5.1: Strategic form of the PD with perfect information if strategies can be chosen

This table of the **strategic form** of the prisoner's dilemma variant with perfect information emerges only if we assume that strategies can be chosen and not merely be planned.[134] It contains four rather than two *choice* options for the second mover.

In sum, a strategy is not "in" the game; it is "about" the game. The term strategy is used to describe how somebody *plans* to play in the game. Strategies are sequences of planned actions rather than executed moves. We can, possibly, choose to form one *plan* rather than another for a game, *but we cannot choose the execution of the plan in one single act.* A strategy for the game is not among the moves present in the original game.

For example, in the original PD-like game with perfect information as presented in the extensive form of figure 5.1, the strategies of the second mover do not show up.[135] When the four strategies (rather than two moves) of the second mover show up in table 5.1, we must read that table either as being a representation of strategic planning in figure 5.1, or the table 5.1 presents a game different from that in figure 5.1. To choose a strategy in one act rather than merely planning to make several choices consecutively presupposes that the option to

[134] i.e., if we read the table as representing choices rather than plans.

[135] This is also one of the reasons for the rather clumsy formulation of a "PD-like" game.

make *such* choices as show up as the – 4 rather than 2 – columns in table 5.1 exists.

If we lack such additional options in a game like that of figure 5.1, we are in the same situation as Ulysses on a boat without a mast and without companions to tie him to it. As in the case of Ulysses, *as a matter of fact* there may well be options that amount to something akin to the choice of a strategy. However, if there are such options, then (according to the principle of intervention and the explicitness condition) these options must be modeled explicitly *as moves*. To get a fully specified rational choice model of the interaction situation, the options must show up as part of the "**rules of the game**," which comprise everything that is beyond the causal influence of the choice making of the players in a play of the game.[136]

If the preceding table 5.1 is interpreted as a game model rather than a "model of the mental model" of planners considering plans for figure 5.1, it is instructive to actually draw the game tree. The extensive game representation of the preceding strategic form representation in table 5.1 would be the following one:

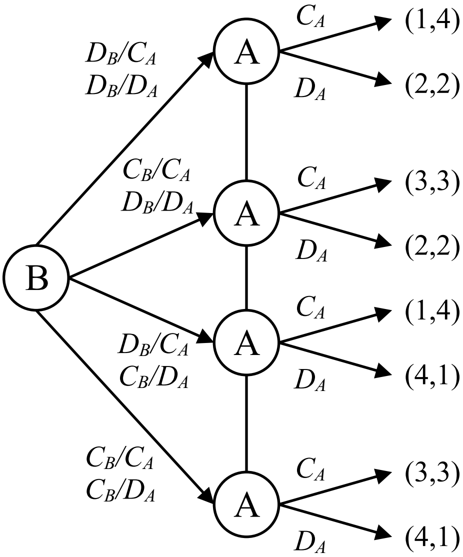

Figure 5.2: PD with power to commit to strategies as programs

Due to the presence of the "**commitment option**," the plans become fixed as behavioral programs for the relevant choices of B in the game of figure 5.1. After

[136] Note that to the extent that preferences are treated as "given," they are part of the rules of the game.

the information about A's choice becomes known, B's response is "programmed." The choreography for the game of figure 5.1 has been fixed by choices described in 5.2, which in turn is equivalent to table 5.1 but *not* to figure 5.1.

If player B could not only make his strategy choices beforehand but could also make them perfectly known under conditions of perfect information, the game would again be different. In this case, the co-player A would actually know what the commitments of player B are. The **commitment power** of player B, his ability to choose a strategy rather than merely to plan on the execution of a strategic plan, would change rational play, and this would become sufficient for avoiding the Pareto dominated result under conditions of opportunistically rational choice making.

In the following figure, the end-nodes that are to be chosen by player A if he comes to move and if informed about the commitments of B are indicated by an asterix "*". The end-node indicated by "←"would be reached if a rational choice maker B made a commitment in anticipation of the rational choices of A.

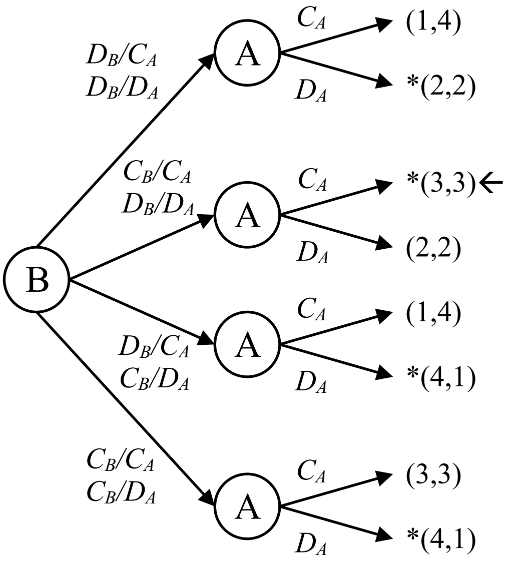

Figure 5.3: Commitment with perfect information in a PD setting

As in the previous discussion of Ulysses, the solution of the game emerges by going through the tree backwards. Player A, when he comes to move, will know how B programmed his choices. B made what may be called a "**constitutional decision**" for the original PD-like game. Assuming that he had the option of programming himself the programs express his constitutional commitments to make certain choices in the course of the game depicted in figure 5.1. This constitutional decision will have happened before the game of figure 5.1 is actually played by the two players.

In the PD-like game of figure 5.1, as opposed to the constitutional game of figure 5.3 in which the PD-like game is embedded, B is a kind of automaton. B is not making choices anymore but simply executing a choreography of overt acts that B has fixed "one level up," so to say, on the commitment stage described in figure 5.3 or before the PD-like game of figure 5.1 is actually played. Player A anticipates how B is programmed when making his own first move in the PD-like game. This is something that player B as a planner knows. B in planning on which commitment she should choose will anticipate the anticipation and knowledge of A. She knows that A will have perfect information about her commitment,[137] and this will induce her, B, to become committed such that the best response of A to the information about her commitment will be an initial co-operative move in the tree of the PD-like game of figure 5.1.

Finally, consider again the case of the original prisoners' dilemma game with imperfect information. If actors – before that game is played – can choose a behavioral disposition to cooperate and if the commitment to cooperate can be chosen such that it is "binding" *if and only if* the other actor is committed likewise, then the Pareto dominated result can be avoided by rational choice makers. If players *cannot* make their commitments contingent on each other, the PD problem will not go away but rather resurface at the commitment stage. Not to commit, remains a dominant strategy.

In sum, the discussion of the distinctions between figures 5.2 and 5.3 is in a nutshell the contribution of game theory to the philosophy of constitutional political economy. It illustrates the far reaching consequences of the rather elementary *distinction between making a plan and making choices* (according to a plan) and how, if present, the faculty to constitutionally commit can change games.

Planning to choose is not tantamount to choosing, and, a strategy can be chosen only if there is such an opportunity to choose or to become committed.[138] Actors who have the additional ability to commit are reaching better results than other actors. External commitment options like the mast of Ulysses or a contract institution may provide such means to reach given aims, ends, or values.

Yet, there may also be internal commitment options (giving rise or at least intimately related to what traditionally has been called virtues).[139] Commitments

[137] See also Gauthier (1986)

[138] If the reader wonders what this is all good for, let me say this: If RCM is used and interpreted carefully to express all options to constitutionally commit and if the explicitness condition in reading game representations is taken seriously the strategic insights of game theorists like Schelling and Selten become accessible and the confusions of philosophers like Gauthier transparent.

[139] See on this Baurmann (2002).

that are internal to the personal actor can be modeled in RCM, too, if we split the personal actor into several agents. This we will sketch next.

5.1.3 Internal Commitments expressed by RCM

Classical RCT, rational choice theory, looks at decision makers as opportunistic persons, who choose whatever comes handiest at the time. In this sense, RCT and the homo oeconomicus model as its most prominent variant have very uneasy relationships to internal commitments and rule-following behavior of personal actors. RCM, rational choice modeling, does not suffer from the same problem. Contrary to a widely shared misconception that does not separate RCM sufficiently from RCT, commitments can be expressed most precisely in the language provided by RCM.

What is at stake may be best understood by looking at a simple example. This is called “take it or leave it.” We will in fact look at three representations of that game. Doing so will make clear, too, what three different concepts of representing interactions mean. The first presentation is in **personal player strategic form** (table 5.2), the second in **personal player extensive form** (figure 5.4), while the third splits the personal players, A and B, of the extensive game representation into agents and, thereby, models internal commitment power explicitly in an **extensive form agent representation** (figure 5.5).

To get to the presentations let us start with the tabular representation of the game. It becomes clear from the context which of the different actions of different choice makers are represented by +, –, respectively.[140]

B A	+	–
+	1,0	–2,–2
–	0,1	0,1

Table 5.2: “Take it or leave it” in personal player strategic form

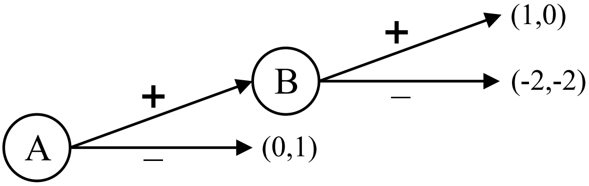

Figure 5.4: “Take it or leave it” in personal player extensive form

[140] Leaving out the subscripts that show who is choosing the relevant options, +, – will avoid notational clutter without spreading confusion, I hope.

In the strategic form of "take it or leave it," it is obvious that there are two equilibria. However, as the second presentation in extensive form shows, only one of those equilibria is a plausible one – at least if players are assumed to be guided by the principles of intervention and opportunism. Once player B has to make a choice, she knows that she would hurt herself should she choose the option "–". If she ever came to move, the future causal consequences of her choice dictate that she choose "+". The principle of intervention rules out the equilibrium (–, –).

A forward-looking rational actor will have to behave as described before. Once she comes to move, rationality dictates her choosing "+". She might, however, try to threaten the other player A by announcing her intention to play "–".

If, for instance, the two players were to sit together in a room before actually playing the game, the second moving player B might inform the first mover that she will choose "–" should the first mover make her move at all as a second mover. She may seriously plan to respond to "–" by choosing "–" herself.

However, if the planning and communication stage were over and the game was being played, she would be in the same position as Ulysses after falling under the spell of the sirens without being tied to the mast. Like Ulysses who will rationally jump into the water if not tied to the mast, the second mover in the "take it or leave it" game will rationally avoid executing her threat. The first mover in the "take it or leave it" game will anticipate that choice of the second mover. Provided that the rationality of the players is common knowledge[141] and provided that choosing "–" is what rationality dictates, knowledge of B's rationality will induce the first mover, A, not to comply with any threat of "–".

In sum, the second mover is a victim of her own rationality and of the fact that she is known to be rational. Without pre-game commitment power, her pre-game threats become nothing but "cheap talk," so to speak.

Conceivably, a second mover could respond to this insight in two ways. On the one hand, she might try to appear "irrational;"[142] on the other hand, she might try to find some commitment device. In the first case, it must be possible to deceive the co-player in order to induce some uncertainty. If such a means of deception exists, then it would have to be modeled explicitly in the tree however.[143] Yet, we will not take that route here.[144] Since the game tree would

[141] In the case at hand, it suffices that A knows that B is rational. Higher order knowledge is not necessary to reach this result.

[142] This would be a case of "motivated irrationality."

[143] Arguably, such an uncertainty, namely that the Americans might respond "irrationally," kept the Russians from marching to the river Rhine and seizing all of Germany in the early years of the Cold War.

[144] We will not introduce a fictitious move of nature and a distribution of player types, see Harsanyi (1967–1968).

change in any event, we can just as well focus on cases in which deception is impossible.

As in the case of Ulysses, for true commitments to occur, it is necessary that commitment devices exist. However, contrary to the case of Ulysses, they may be internal to the choice making entity. Whether people can internally commit or not is a factual question. Whether actors, like human persons, can conceivably communicate their internal commitments such that common knowledge of the part of the tree that is internal to the actor and perhaps even perfect information may be assumed to apply is an open factual question.

Assume that these factual conditions are met in the case of "take it or leave it." Then, a tree of the following kind might emerge:[145]

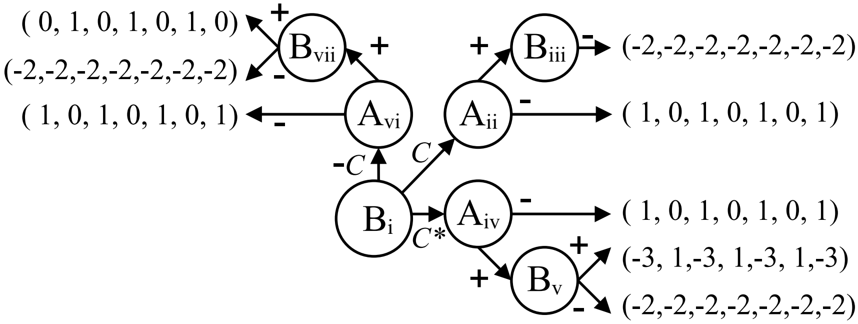

Figure 5.5: "Take it or leave it" in extensive form agent representation

The agents of players A and B are indexed with Roman numerals indicating the position of the payoff for that agent in the payoff vector. That payoff applies to the node with the respective index and shows preferences all things considered *at that* node. In the initial move, agent i of player B can make one of three choices. First, the agent can choose the move -c, which stands for remaining internally uncommitted. This leads to the original game as a **sub-game** of the larger one.[146] Second, the agent can choose c* to become relatively committed by changing his preferences. Third, with the choice c, the agent might be able to bring about a state of mind in which the later agent of B would not perceive the option of acquiescence as a possibility any longer. The rest of the tree is more or less self-explanatory.

Now, no claim is being made here that internal "masts," which are functionally equivalent to Ulysses' "external" mast, do in fact exist. However, the preceding should suffice to demonstrate how flexible the language of non-co-

[145] This and the subsequent discussion are taken from Güth and Kliemt (2007).

[146] A sub-game is a node of a game tree with all subsequent nodes and no connections to information sets in other parts of the tree.

operative game theory is. Meeting the requirements of the explicitness condition forces us to say what we assume about the abilities of rational actors. At the same time RCM provides the means to express almost any assumption about strategic abilities of choice making entities.

In sum, the splitting of personal players into agents provides the means to express all substantive assumptions about internal constitutional commitments of a personal actor explicitly in the language of non-co-operative rational choice modeling.

One should note carefully here that the preceding is all about rational choice modeling, RCM. The empirical issue of whether or not certain assumptions of RCT apply in real games is a completely different one. Moreover, whether it is a good idea to make up a world by RCM is another question. Yet, it speaks for RCM that in particular those who reject the RCT view of completely uncommitted opportunistically rational actors should want to express explicitly their deviating assumptions,[147] and, as in particular the difference between the personal actor and the agent representation shows, that is precisely what RCM allows for.

5.2 Trust and Commitment

5.2.1 Economizing on trust

Imagine that you are asked to lend ten thousand dollars to one of your business partners. In return, she promises to pay back eleven thousand dollars next year. Your preference for present as compared to future income, i.e. your "discount rate" is low enough and the next best investment not as good as the one on offer. Yet, the fact that you know her to be rational may make you think twice. With the words of Hobbes (see Hobbes (1651/1968), chap. 14, 196):

> "For he that performeth first, has no assurance the other will performe after; because the bonds of words are too weak to bridle mens ambition, avarice, anger, and other Passions, without the feare of some coercive Power; which in the condition of meer Nature, where all men are equall, and judges of the justness of their own acts cannot possibly be supposed. And therefore he which performeth first, does but betray himselfe ..."

In the real world, the reasons for the act of lending may range from a rational expectation to get the money back to blind faith without any rational expectation formation. Within the world of economics and classical rational choice theory, the

[147] A polar case would emerge if one used a language according to which all acts are committed to certain norms or rules and the exceptions of opportunistically rational, uncommitted choice must be specified explicitly.

set of possible reasons is, however, narrowed down considerably. An economic account of the act of lending must present it as a type of behavior that is "as if" maximizing a utility index. Slightly modifying Robertson's (see Robertson (1956)) famous characterization of economics as "**economizing on love**" as the scarcest resource, we could also say that in economics all efforts have to be directed at inventing mechanisms that "economize on trustworthiness." The economic solution of trust problems amounts to eliminating the necessity and usefulness of trust by rendering the act of rewarding a first mover's trust a utility maximizing choice of the second mover.

In the example, we must see to it that by some mechanism or other, the rules of the game are changed such that the borrower will have no better alternative than to pay the money back in the future. Foreseeing this, the lender does not need to trust to lend the money but will rather rationally bet on getting his money back. Hobbes own answer is exactly of this kind. It is based on the introduction of the state as a kind of external referee that operates as an enforcer of rights, promises, and contracts. In this role, the state can define property rights for individuals in their private capacities and make possible a process of contract enforcement such that it is in the rational interest of individuals to behave well and to fulfill their contractual promises.[148]

Obviously, the Hobbesian state economizes on the need for trustworthiness and, consequently, on the necessity for showing trust. The Hobbesian enforcer ideally creates a world in which no trust is needed, only rationality. Contracts will be enforced, and one can trust the expectations based thereon. (That it requires quite a bit of trust to hand over ourselves to "mighty Leviathan" is another matter though).[149]

According to the preceding view of the world, the state makes viable ways of co-operation and mutually beneficial exchange – in the widest sense of that term – that would be non-viable without the state as a guarantor of rights and contracts. In the Spinozist world with a Hobbesian Leviathan, individuals either do not need to trust, since they can rationally expect that contracts will be performed, or if trust is needed, they would do better to refrain from all transactions. However, besides a state-sponsored process of contract enforcement, there is room for other mechanisms that can potentially serve as substitutes of either superior quality or lesser costs or both. There are non-statist alternatives to external contract enforcement and commitment. To what extent these nowadays popular alternative strategies of endogenous order creation are fully compatible with RCT and to what

[148] From this point of view, it is obvious that the market – at least any large market – is not an anarchical device of social co-ordination. Though what is going on on the market is quasi-anarchical, it is not taking place in anarchy properly so called.

[149] On such a Spinozist account, the efficient results emerge precisely because the state guarantees individual spheres and empowers individuals to make decisions of and on their own (see for a criticism of such statist views perceptively de Jasay (1997), de Jasay (1995)).

extent they depend on building some commitment power into the rules of the games real people play remains to be seen though.

5.2.2 Internal commitments in trust games

The trust game that will be used subsequently to illustrate commitment problems in their most simple form is derived from the PD-like perfect information game of figure 5.1. Removing the play (D_A, C_B) as irrelevant for our present concerns, the following general form of the trust game in ordinal utility payoffs emerges:

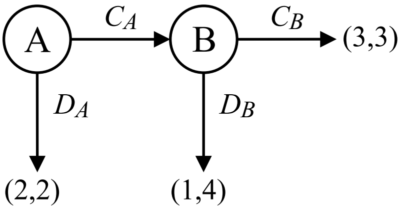

Figure 5.6: Trust game in ordinal utility payoffs

Let us assume for the sake of specificity that the game emergent in utilities, which represent attitudes to risk, is the following one:

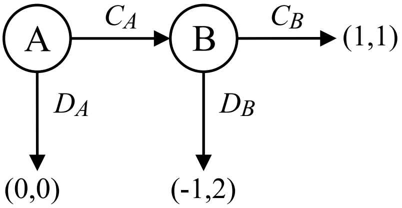

Figure 5.7: Simple trust game

In this game, the utility index shows that the second mover prefers to move down rather than forward should she at all be called upon to move. This is what she should and will do *all things considered.* Though the game is conventionally referred to as the "trust game" or "game of trust," there is no room for trust in such a game. If the utilities are common knowledge, the first mover knows that the second mover will, all things considered, move down. The first mover should, therefore, refrain from choosing to move forward and rather be satisfied with a zero payoff. However, both would be better off could they refrain from complying with the principle of intervention in perceiving the world and from making each choice separately in an opportunity taking way.

If the second mover can choose between playing the "trust game" or another game in which she is committed, then she may want to choose that other modified game. The modified game may provide higher payoffs for her (and also the other actor). What is at stake here is shown in the next figure:

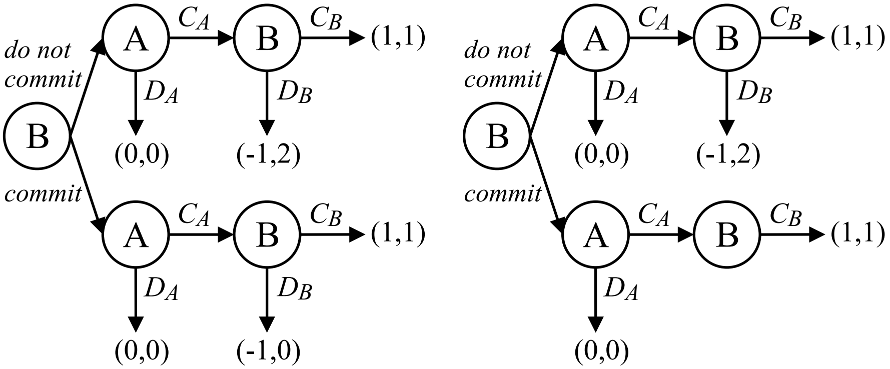

Figure 5.8: Games that may emerge if actor B is able to commit

The left side of figure 5.8 shows a relative commitment. It is assumed here that the actor who is to play as a second mover in a decision prior to the game itself can deliberately commit by changing the "utility" order (or rather the underlying preferences). If she can do that, she should rationally do it. For, if common knowledge of the game tree prevails, actor A will have an incentive to move forward when he comes to move after actor B has chosen to commit. The solution of the game is obviously one in which we have commitment followed by two acts of cooperation C_A, C_B, respectively. On the right side of the figure an absolute commitment to co-operate after the other actor A has moved first in a co-operative way can be chosen. Here the option to exploit the first mover is cut off completely and, again, the solution will be commitment followed by two acts of cooperation C_A, C_B.[150] Without a commitment – relative or absolute – by B the actor A would foresee that she, should he move forward rather than down initially, would move down. She then would end up with "2" in that case and he with "–1".

In sum, there can be relative and absolute commitments concerning a trust game. The language of RCM allows us to express these commitments if the ability to commit exists. Moreover, it does not rule out that the commitments might be internal to the personal actor.

[150] The reader may want to go to the section on strategic planning 5.1.2 and to reflect on how the commitment options might be modeled there explicitly.

5.2.3 External Commitments to repeated trust problems

One way to create "trustworthiness" derives from what has been described as "the discipline of continuous dealings". If on each round of play of an ongoing trust interaction there is another round, then the "shadow of the future" (as Axelrod, 1984, has named it) will be present throughout. Each of the actors, according to the principle of intervention will factor in the effects of her own choices on the future choices of the other actor. Each knows that each knows that there is another round, and this may provide a reason for trusting in the co-operative future moves of a co-player.

The so-called "centipede game" emerges if this type of argument is applied to trust problems. The centipede is a model of finite repetitions of the simple trust game played 50 times over. Each round of play adds another two legs to the animal. Since it is clumsy to depict a centipede, the following discussion of repeated trust interactions is restricted to the decipede case. This restriction does not eliminate anything substantial from the picture but leads to simpler trees.

The next graph shows the game form of the simple trust interaction in monetary or substantive payoffs played five times over. The generic game form of the kth round of play is also shown. Setting k = 1, we get the incremental substantive payoff on any round of play while for any k rounds of play the accumulated substantive payoff is presented.

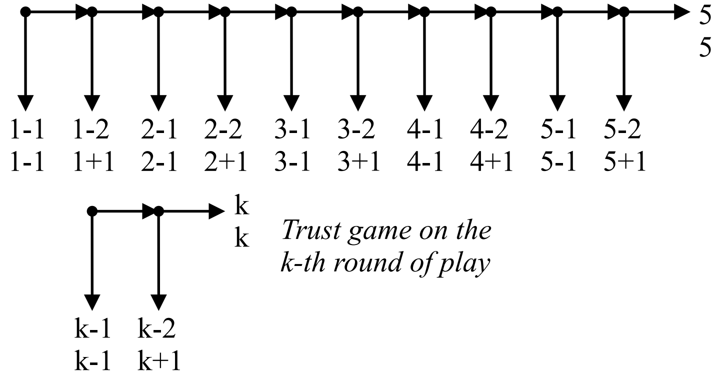

Figure 5.9: The decipede game (form) and constituent trust game (form)

In the one-shot trust interaction, only first mover trust is involved. In the repeated interaction, the second-mover can and must show trust as well if the interaction is to reach another round of play. If the second-mover does not move down immediately, she must trust that the first mover of the next round will not move down immediately as well. More generally speaking, a player who moves forward must always trust that the next moving player will trust and not immediately choose to move down. In each round of play, a player who does not play down hopes for another round. In that sense, we get an ongoing trust relationship.

As long as there is always a future round, each of the players might also intend to reward the good conduct of the other player in the future and at the same time expect the other player to hope for such rewards. However, when the last round of play is reached, the relationship cannot be ongoing any further. There are no future directed reasons for trust on a last round of play. So, if the end is known, no trust should be shown in the hope for future trust from the other player. The last mover who knows that she is moving last should move down on the last round of play. The second to last mover should know this and, therefore, not move forward but rather move down before the last mover can. In view of that, the third to last mover should clearly prevent the play from reaching the last two movers and move down herself in the third to last interaction. Anticipating that, the fourth to last mover should move down immediately and so on.

In sum, according to the preceding "**backward induction**" argument, the first mover in any interaction of the form of the decipede (centipede, millipede etc) should immediately move down and, thereby, stop the interaction before it begins.[151]

More generally speaking, for any finite natural number k of repetitions of the constituent interaction, the definite backward induction solution of the 2k-pede interaction for players only interested in maximizing the substantive payoff is to play down immediately. The vector of substantive payoffs of this solution play is (0, 0). At the same time, for any finite k, violating the principles of maximizing substantive or material gains incrementally could lead to accumulated payoffs of (k, k)>(0,0). If k becomes arbitrarily large, the possible gains forgone by following the backward induction logic become arbitrarily large as well.

The self-interest of the players would prescribe that they both should not obey the "dictates of rationality." This raises obvious problems: If the dictates of rationality so strongly contradict self-interest, can we still speak of rationality in its full sense? Does not backward induction put rationality backwards, so to say, in that it insists on principles like non-dominatedness as a guidance of choice making rather than on the objective success of choice making? On the other hand, if somebody violates the principle of backward induction and becomes, thereby, more successful in objective terms, can she do so *without* violating principles of rationality or *without* some irrationality?

[151] In experiments with real players, this is not observed but end-game effects are well-documented, as for instance in Selten and Stöcker (1983)). However, here we are talking about the thought processes of fully rational beings.

Many people would say that a person who acts according to the principle of backward induction is at best a “rational fool” (alluding to Sen’s article of the same title Sen (1982/1976)). Conversely, if some act violates the principle of backward induction and, thereby, serves their interests better than an act in accordance with that principle, they – it is argued – are not fools but serve their common interests. According to some views, they are showing “higher order” rationality by going against backward induction.[152]

This controversy has been going on for a long time. However, taking seriously the principles of intervention, of opportunism and the concept of a (all things considered) utility representation, the arguments for backward induction are strong.

5.2.4 Repeated trust games and backward induction defended

Assume that in figure 5.9 after all things have been considered by the players, the monetary payoffs have been transformed into a utility representation of exactly the same magnitudes, depending on k the same way the substantive payoffs do etc. Then assume that the result of considerations including the form of the tree are common knowledge and lead to a game as represented by figure 5.9 (figure 5.9 now in *utility* terms). This leads to a decipede *game* (not only a game form) in the full sense of that term and the rules of RCM and non-co-operative game modeling do fully apply.

Note that any path dependence of payoffs as applying to later rounds of play by assumption has been factored in when all things were considered.[153] Moreover, according to the RCM rules of reading (interpreting) a game tree, the result of considering all things must be common knowledge, for, otherwise, the game tree would be different.

Assume that the two players who have to play this decipede game can communicate before playing the game. They might discuss that they both would be better off if they co-operate. Assume also that talk is cheap not only with respect to effort but that it is also ineffectual in the sense that it does not bring about a change of rules (including preferences).[154] The two talk to each other before the game is played. After this, they are, say, brought to separate quarters to actually play the game over the internet, knowing that they will not meet again afterwards.

[152] The so-called chain store paradox (see Selten (1978)) is avoided.

[153] Otherwise, not all would have been considered.

[154] If it were otherwise, then according to the explicitness condition of RCM the modifying options would have to show up explicitly in the game tree – just like the commitment options in the preceding trust game.

Under such conditions, there is *no* “shadow of the future” beyond the last round of play and, once they are in their separate rooms, prior communication is a bygone fact. If communication did not alter the evaluation of results, that is *if the game is still the same* as in figure 5.9, backward induction should kick in and dictate to play down immediately if, as assumed, the utilities represent the evaluation of results *all* things considered.

But what if one of the players in the play of the decipede finds herself in a position to move forward, say, in round 2 of the game? Shouldn’t she then conclude that her co-player does not understand backward induction or has intentions other than what the rationale of backward induction would dictate? Couldn’t she have learned something about the other that could’ve induced her to believe that the other will move forward again? If so, would or should she not – as real players commonly do – choose to continue the game as well? Should we not anticipate that the backward induction argument is self-refuting since in order to formulate the argument it has to be assumed on all rounds of play after the initial move that the thesis underlying the argument has been violated?

The rather simple answer to the implied “refutation” of the backward induction argument is that according to RCM and non-co-operative modeling, *a player in a given game model can only learn what the rules of the game allow her to learn.* Strategic thinking as understood in classical game theory allows us as external observers and as players to think through the game completely. The possibility of genuine surprises or bits of information that have not been anticipated *when setting up the model* does *not* exist.[155] Any strategy as a plan for the whole game contains responses to all conceivable contingencies that might emerge in the game. The strategic plan is formed under the presumption that the model to which the strategy applies already contains all possible information states etc.

In sum, in a properly specified game, the planning strategist does not know before play which of the contingencies of play will emerge, yet she knows in advance which contingencies may possibly arise and what she would learn from the fact that they have arisen.

Applying this to the decipede game, this implies that the player on round 2 of a play of the game cannot learn something about the strategic intentions of her co-player that is not already anticipated in the game model and her analysis of it before the play commences. It must be possible to anticipate in the strategic plan all responses to what may transpire in the play of the game.[156]

If the player, when considering her plan for all contingencies that might emerge according to the game tree – i.e. when considering her strategy – were to

[155] The world is in the sense “small” that the model anticipates by assumption all relevant possibilities.

[156] All that might transpire must be already captured by the model to which the strategy refers.

assume that there could be a genuine surprise, i.e. a possibility not anticipated as a possible state of information in the game tree, then this would lead to the suspicion that the game model for which she is going to formulate her strategy is mis-specified. For, there would not be a clearly defined object of common knowledge. If the model is as specified in the game tree and if the game tree, in line with the explicitness condition of non-co-operative modeling allows only for the states of information and commitment as depicted in the decipede game, then the formulation of the tree contains everything relevant.

By finding herself at the second node, she cannot learn something about the intentions or rationality of the other player that has not already been factored in when setting up the tree, preferences and future options from the second node on. If she finds herself at a node on the second round of actual play and if the preferences at decision nodes are satiated in representing *all* aspects relevant to decision making at *that* node, then she should still come to the conclusion that backward induction is right. Moreover, one should not forget that she does not analyze the tree when she is actually playing. While planning, she is not going through the interaction represented by the tree as a non-co-operative game.

In sum, "given" the utilities that were formed to represent preferences when reaching the relevant *node to which they apply*, we must recognize the backward induction argument as valid.

Since arguments about backward induction can fill whole libraries it may seem almost arrogant to deal with the problem "the short way". Therefore, though I believe that the preceding is conclusive, let me try to add to the argument.

5.2.5 Repeated trust games, backward induction re-considered

We would presumably be less reluctant to accept the conclusions of backward induction if the two personal players were teams of players. In the decipede, we would have five separate persons in each team. To assume that the last player in each team would have good reason to move down does not seem outrageously irrational to most of us. Neither is it absurd that one of the players in a team would move down out of self-interest.

In the real world, there may be a common good for the teams.[157] Members of teams may take into account the interest of other members of their team. There may even be material payoffs that could be accumulated in a common pool for all team members. Such accumulated substantive payoffs might form "side-payments" to be distributed among team members after the game ends. All this is possible.

[157] See for an overview of some of the ways and the references to the relevant more technical literature here, Brennan and Kliemt (1994).

If we wrote down the decipede game in so-called agent form – as introduced in the Ulysses problem above and pushed to some extreme in the “take it or leave it” example – then this would correspond to the game with two teams, each composed of five one-time personal choice makers. The preference orders would represent how the agents would evaluate their two options at each instance of choice making *all things considered.*

That according to the assumptions underlying preference representations “all things” are considered in the preference representation (that the preferences are “satiated”) makes it viable that each decision can be analyzed according to the preferences relevant for *that* decision. Due to this interpretation of “representative utility,” decisions can be analyzed completely independently from each other. Once the utilities are written down, the model is specified for each decision with the utilities representing the preferences relevant *at that* decision node *all things considered.*

In view of the preceding, the question to be asked is really: If the *ten* (!) preference orders implicitly assumed to apply in the decipede game were constant and were as we have assumed them to be, would backward induction still seem absurd? If the values in figure 5.9 could be strictly interpreted as utilities representing satiated preferences, would it be absurd to assume that the decisions would be made in the way described here? If backward induction seems absurd to so many, is this not due to or at least related to mixing up utilities and substantive payoffs?

I do indeed believe that part of the confusion arises from mixing up substantive and utility payoffs. To see what is involved, let us go back to monetary payoffs for a moment. Assume for the sake of specificity that each personal player functioning as an agent of a team receives either zero dollars, one dollar, two dollars, or loses one dollar to another player as a result of playing the additional trust game. We interpret the trust game of the previous section (see figure 5.7) now as a game form in substantive payoffs, say in “$”:

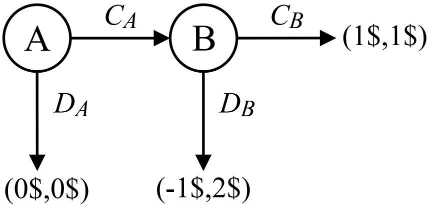

Figure 5.10: Simple one-shot trust game form with monetary payoffs

Note that we have no difficulty imagining that this game form *in monetary payoffs* is played again and again with *identical monetary* payoffs. That the monetary payoffs remain constant regardless of repetition is a rather innocuous assumption. Assuming that the monetary payoffs of incremental gains or losses of $0, $1, $2, $–1 dollars induce corresponding preferences that can be represented by the same numerals – all things considered – we arrive at the result that rational personal players (as well as agents) who would not trust could do so without violating plausible behavioral assumptions. It is not absurd not to show trust in a one-shot game.

Now, turn to the repeated game in dollars as played by a team. Assume that in each period exactly one personal player exists. After playing he gets out of the game. *As modeled*, each of the payers will receive only the incremental payoff. There is no accumulation of k since each personal player gets paid off according to the monetary payoffs of the simple trust interaction. Assuming that the monetary or dollar payoffs translate into utility payoffs of the same magnitude it is still not absurd that each of these players does not show trust.

It may be that personal players do have some regard for the other players in their team or in the other team. However, that would imply only that their preferences would deviate from the order indicated exclusively by the natural order of monetary payoffs. A re-evaluation would have taken place. This would only lead to the conclusion that the game is not as specified. It cannot be represented by a tree that builds on the repetition of the *same* constituent game on each round of play (which is different from the repetition of the same game *form* with the same monetary or substantive payoffs on each round of play). I have no doubt that the assumption of an identical repetition, though rather innocuous as far as substantive monetary payoffs are concerned, is almost absurd if the payoffs are in utility terms. However, if this assumption is made then it seems to me rather plausible that the fact that a trust game is embedded in a sequence of such games will lead to identical solutions for each of the players who on behalf of their team play the identical one-shot game.

I cannot see any reason why the preceding line of argument should change if the team is formed by agents of a personal player rather than by personal players. If the utilities are the same for each of the agents the solution should remain the same. And, the latter is true by assumption of identical repetition.

In sum, *if* the preferences at the nodes of the sequence of decisions are such that they can be represented by utility indices as given in the presentation of the game tree, then these preferences apply at the respective nodes all things considered. If not, then not.

In case of a team of separate agents who are all personal players in the common sense meaning of a person, identical utilities may seem intuitively more plausible. However, if we acknowledge that at each decision node a separate utility function must be operative – as representing the preferences at this node – it is, in principle, not otherwise in the case of a single personal player coming to move at separate instances of choice. For a personal player it is, of course, more likely that the preferences change at separate instances of choice making. Vice versa, it is *empirically less likely that in the case of the repetition of a trust interaction the preferences of a single personal player over the material, monetary, or objective payoffs remain unaffected by the history of the interaction if she comes to move several times.*

The preferences relevant at different nodes of the tree of a repeated game may differ between otherwise identical stage game *forms* (in substantive payoffs) as a function of the round of play. As a matter of fact, it may be exceedingly unlikely that a personal player would ever be totally unaffected in her preferences concerning monetary payoffs in a sequence of repetitions of the same *monetary game form.* When going through such a sequence of plays she would change her preferences concerning monetary outcomes and the path leading to the outcomes. Yet, then the fact that the preferences for monetary payoffs will change for the personal player should show up in the preference representation, and the identical repetition assumption would be gone. Moreover, should somebody claim that there is no way to repeat an identical game – proverbially nobody can enter the same river twice – then let us acknowledge that fact of psychology and model the alterations of the rules of the game explicitly rather than fool around with the rules of interpretation of non-co-operative game theoretic modeling.

In sum, that the rules of a basic game and in particular the payoffs are changed by the repetition of a game form structure with identical objective (monetary) payoffs may be unavoidable. However, if that is so, we need to model the modified rules and not to modify our concept of rationality or our technique/language of rational choice modeling.[158]

Somebody might still object to the assumption that persons can be split into agents. However, given the notion of utility as representing preferences at a specific choice node *all things considered*, we can hardly avoid relying on the so-called "**agent form**" representation of interactions. The agent form is implied by the notion of a preference representation by utilities *all* things considered in combination with the view that the utilities applying at a node represent *everything* that is relevant for the choice at that node. *If* utility is representative of preferences "*all things considered*," then *the position of a decision in the sequence of decisions is already factored in (considered) by the representation of the preferences operative at the node where the decision is to be made.*

A "signal" that comes, so to say, from the past can influence a decision-maker who decides in view of the future *causally* only by leading the decision maker to one specific node rather than to other nodes. Yet, within the given structure as set up in the language of RCM all choices are assumed to be made in a forward-

[158] One could imagine that the monetary payoffs in a trust game form structure would be modified after each round of play. The temptations to exploit and the losses from exploitation would be increased on each round of play such as to compensate for the influence that repetition might otherwise have on preferences.

looking manner. They are drawn from the future or in formal terms always by the "rest" or remainder of the tree and not pushed by the past or by what happened at preceding nodes.

Taking the principle of intervention seriously and approaching social interaction from a participant's point of view the preceding seems to follow. However, some *tough questions* remain:

1. The personal player or the knowing and thinking entity does not show up in the game tree anymore. Yet, at least the rational choice maker as a mere agent of a personal player can hardly be the one who thinks through the game. So, who is it then who thinks the game through strategically and as a participant of what? Sticking out the consequences, have we not shown that eductive theory is in the end absurd (though for a reason other than the conventionally cited one)?

2. Can we assume that at later nodes all that matters for future directed choice is always represented explicitly in the rules of the game rather than being in some path-dependent way on the "mind" of the decision making entities (as an input from past play) and at the same time avoid the conclusion that there cannot be any general hypotheses about solving games?

Economists are fond of the assumption of given and satiated preferences that are revealed by choices. They tend to defend these twin assumptions since they seem to allow them to treat the human mind and human reasoning as a kind of black box. Whatever may be going on in the box is represented in the economic game model by the stenographic device of the utility function. In particular no analyses of cognitive processes and reasoning in terms of cognitive psychology are required.

It is intuitively obvious that a more or less behaviorist view like this does not fit well with the aim of analyzing games in terms of reasoning about knowledge. Taking the concept of satiated preferences to its extreme the utilities must include everything that would be relevant for choice making. To be all-inclusive in that sense, utilities would have to represent preferences that emerge only due to reasoning about the game itself. The reasoning about knowledge would have to enter the formation of preferences at each decision node. It would be among the reasons for preferring alternatives and in that sense could not be included in what is treated as "given" in the reasoning process. The utilities would be needed to solve the game but would be fully determined as representing preferences all things considered only after the game has been analyzed on the basis of the allegedly all-inclusive preferences.

Such absurdities can only be avoided if preferences are treated as non-satiated and as something that is not completely revealed by the choices made. We must look at the reasons for preferring. We cannot treat these reasons as irrelevant. We must open the black box and have to give up the rather crude behaviorism that underlies the economists' "preference" for the revealed preference concept.

In sum, the combination of revealed preference and reasoning about knowledge that underlies classical game theoretic reasoning seems incoherent and in need of some "repairs".

However, the rather strong reason for combining the seemingly incoherent concepts of satiated preferences – that can be represented by utility functions – and an eductive approach to games in terms of reasoning about knowledge should be taken into account as well. For, unless we rely on some concept that is at least akin to satiated preferences classical game theoretic analysis seems impossible. We need preferences that can be determined independently of the specific game context to which they apply. We need preferences that can be represented by utilities independently of reasoning about the game. Otherwise we might come dangerously close to saying that each case is different from all others in social life and no analysis would be possible.

What is at stake here can be illustrated again by turning to the decipede game and the concepts of separability and agent-based models of social interaction. So let us turn to these to explore the limits of rational choice analyses somewhat further. In doing so I will start with the second of the "tough questions" and then turn to the first.

5.3 Rational choice analysis at the limit

5.3.1 Listen Folks

In game theoretic modeling, it is assumed that an analysis of complex structures by parts is possible. That this is indeed possible is the main point of the whole effort developing analytical tools. For extensive game representations, the following separability condition is crucial in making analysis viable (see on separabitlity also McClennen (1998)). Let T be a decision or game tree:

1. Consider a sub-tree T/s that emerges after all nodes T_s preceding node s are cut off while s and all its subsequent nodes remain, then the solution of the game represented by the sub-tree is the same as the solution for the sub-tree while it was still embedded in the larger game tree.

2. Assume that tree T/s considered as a separate game has solution h. Assume that T/s becomes embedded in a larger game T as a sub-game of T. When playing the larger game T, if s is reached and the sub-game T/s

> ensues, the solution of T/s is still h; i.e. h applies to T/s taken separately and to T/s as embedded in T = (T_s, T/s).

The consequence of taking separability seriously is that whatever comes from the past, so to say, must be anticipated in the rules of the game as represented in the game tree. *The play of a* game T_s before T/s *cannot influence what is strategically rational in T/s, only the rules of the game T/s itself are relevant for this.*

Assuming separability seems almost a corollary of accepting the principle of intervention. According to the principle the actor perceives of the situation such that at s only the future T/s matters. However, it is a very strong notion. With separability even the so-called **Folk Theorem** of the theory of indefinitely repeated games becomes precarious.

The Folk Theorem got its name because it was folk wisdom among game theorists long before it was formally proved (for an early overview, see Aumann (1981)). Put simply, the Folk Theorem says that in a repeated game, in particular in a repeated prisoner's dilemma game, a wealth of strategies that condition play on previous play can be in equilibrium. The strategies are such that against the planned "punishments and rewards" specified by each strategy a deviation to another plan would not pay for any of the participants.

For instance, if each of the participants in an indefinitely repeated prisoner's dilemma game planned on co-operating as long as no deviation on any round of play ever occurred and planned to deviate indefinitely if the first deviation occurs, then all might plan on this in equilibrium. Playing according to plan, all would co-operate at the beginning and would go on to do so until the first deviation occurred. But nobody would in equilibrium plan on a defective move. Deviating once from the plan to a defection move on one round of play would bring all participants down to the non-cooperative result on all the future rounds of play if all stuck to their plans otherwise. Therefore, any single deviation from this strategy to include a single additional defection in it would not be worthwhile. Conversely, if all others stick to their grim strategies to include one additional co-operative move, would not bring any of those who are planning to defect indefinitely around to respond with co-operation once the co-operation broke down. Therefore, against a set of grim strategies a less grim one would not be an improvement.

The unraveling of the equilibrium among conditionally co-operative plans of the kind sketched before does not occur because there is no last round of play in which it would be better to plan on an exploitation move. In games with no end, backward induction arguments can be avoided. As we may add here, though it will not be demonstrated, the Folk Theorem also shows that with less grim and more complicated contingent or conditional strategies for the repeated game, which not only specify indefinite defection in the case of a single deviation but complicated conditional reactions, almost any rational "payoff constellation" and thereby any "average" payoff can be realized in equilibrium.

Yet, the Folk Theorem does not apply if we take separability seriously and assume that there is a definitive solution to all sub-games (see on this Güth, Leininger and Stephan, 1991). To see why this is so, imagine that you are dealing with a simple two-by-two PD repeated indefinitely. Note that the constituent game is assumed to be the same on each round of repetition.[159] If at any node s of the game tree T you cut off finitely many preceding nodes, the remaining game tree T/s is structurally identical to the full tree T.[160] After all, the same basic game is repeated over and over again. Because the basic game remains identical, the indefinite repetition of this identical basic game will be still an *indefinite* repetition of the "same thing" even if finitely many initial games have been cut off. Yet, if this is so, then the solution of the structurally identical game emerging should remain the same.

Making later play contingent on previous play will not make sense if separability is assumed to apply to the preceding case. After all, the remaining sequence always looks the same and must, therefore, be solved in the same way "unconditionally." If according to the principle of intervention only the future matters, then, if the future always looks the same, the future-directed rational choices should be the same regardless of how the first decision node leading into that future was reached. In the preceding case of cutting of the first s or r initial rounds of play, regardless of $s \neq r$ the games T/r and T/s should have the same solution h. More specifically, whether a co-player in an infinitely often repeated PD did or did not co-operate *before* should not affect the decisions of an opportunistic choice maker who looks at the future only. Making one's own co-operation contingent on the past co-operation of a co-player is ruled out by strict future directedness and the insight that the future in an indefinitely repeated identical constituent game always looks the same.

The preceding does not yet reveal what the solution should be. In view of the principle of intervention, it tells us, though, that it should always be the same for structurally identical sub-games regardless of the past. However, it seems that the only plausible remaining candidate for a solution strategy for the indefinitely repeated PD is the so-called ALL-D, the play of the dominant strategy D of the constituent PD game on every round of play. This follows if we accept backward

[159] As stated before, if everything is explicitly modeled, including the preferences that are relevant at any *local* decision node, then it is exceedingly unlikely that in social reality there would ever be repeated games of the kind of the decipede game. The constituent game will not remain constant since preferences change or information conditions must be assumed to diverge from perfect information etc. But the question of whether or not such games ever exist in reality is an issue completely different from the question of how a game tree once it is written down should be interpreted.

[160] Think of the natural numbers: it does not matter whether you cut off finitely many first elements of an infinite series since infinitely many remain.

induction (and therefore that all finite game trees have ALL-D as their solution) and if we assume that the solution of the infinite case should be approximated by the solution of the finite. In this case, the ALL-D argument carries over to the "infinipede," i.e. the basic trust game repeated not only for five rounds but indefinitely.[161] Here, too, the solution should be "**sub-game consistent**" in the sense of solving identical sub-games – or sub-trees in the preceding sense – in identical ways (leading to ALL-D).

The assumption of *identical repetition* of a game in the full sense including utility payoffs is the culprit. If we look at the solution of a game as a function of its rules, which seems the whole point of the exercise of analyzing games, and if we use the principle of intervention to cut off the past, then – together with the explicitness condition – it seems hard to avoid the conclusion that solutions of games that are all structurally identical with respect to the future should not vary. Even if we did not subscribe to the view that the solution must be a singleton, the solution sets would have to be the same for structurally identical sub-games. All variation according to past play would be ruled out for the simple reason that the rules for the future game are all that can matter.

In sum, the history dependence of the conditional strategies of the Folk Theorem logic is ruled out if we take the principle of intervention seriously and apply it to games assumed to be the result of repeating an identical base-game indefinitely.

In the limit the future directedness of human rationality seems to have problematic implications. The more subjectively rational individuals are the higher the objective payoffs forgone. Playing All-D indefinitely in a PD in which co-operative results would lead to Pareto superior objective (and subjective) payoffs puts the fully rational at a disadvantage as compared to individuals who are merely "boundedly rational". Yet there are further limits not only to RCT but also to RCM.

5.3.2 Invariant payoffs?

Consider two decipede games, decipede-1 and decipede-2, that are completely identical in move, information, and payoff structures. Assume that the payoffs of the decipedes are given in monetary terms as monetary-decipede-1 and monetary-decipede-2 and that the preferences over the end-nodes of the games can be represented by utilities that are numerically identical with the monetary values, leading to utility-decipede-1 and utility-decipede-2. Being identical, both, utility-

[161] That the infinite may have very different properties from the finite in game theory is shown in Rubinstein (1989). However, even if we factor that in, the requirement that structurally identical games should be solved in identical ways would still exist. It should be obvious that this together with the impossibility to condition on the past makes it impossible to get to the normal Folk Theorem.

decipede-1 and utility-decipede-2, should have identical solutions. Take now the game *form* that emerges by going on with monetary-decipede-2 after the last node of monetary-decipede-1. Refer to the embedded decipedes based on monetary-decipede-2 as monetary-decipede-3 or utility-decipede-3 respectively. That is, we refer to the former T/s with different names when embedded in T and when considered as standing alone.

We have *first*

monetary-decipede-1 and monetary-decipede-2,

from which we form the pair

(monetary-decipede-1, monetary-decipede-3)

Being embedded in the sequence does not affect monetary payoffs. Therefore:

monetary-decipede-3 = monetary-decipede-2 = monetary-decipede-1.

We have *second*

utility-decipede-1 and utility-decipede-2 based respectively on identical

monetary-decipede-1 and monetary-decipede-2 and therefore taken separately we should have

utility-decipede-1 = utility-decipede-2.

However,

(utility-decipede-1, utility-decipede-3) based on

(monetary-decipede-1, monetary-decipede-3) may well lead to

utility-decipede-1≠ utility-decipede-3 regardless of

monetary-decipede-3 = monetary-decipede-2.

There is no reason why it should be difficult to offer the same monetary payoffs in the second sequence of five repetitions as in the first sequence. However, the invariance assumption may be problematic in the case of utilities. Preferences “all things considered” may be influenced by the history preceding an interaction. Therefore it may well be that the rules of utility-decipede-3 which derive from monetary-decipede-2 as *embedded* in (monetary-decipede-1, monetary-decipede-3) differ from utility-decipede-2 standing alone. In line with the principle of intervention, going through the history of decipede-1 may have causal effects on the evaluations that are showing up in the later sub-tree of the pair forming the full tree. This must be taken into account when formulating the game model for later stage games. However, once the effects are completely taken into account in the model, only the future matters and backward induction kicks in *since the whole analysis of influences that come from the past has already been embodied in the rules of the tree, in particular into the utility payoffs.*

As the combination argument for decipedes shows, we cannot take for granted that a game structure standing alone will have the same solution as an embedded one. The game form may be the same, but the utilities may differ depending on the preceding game tree. It is in fact often highly implausible that a preceding history would leave a game tree unaltered in *subjective* terms event though it remains unaltered in objective terms (i.e. as far as move structure, information partition, and material payoffs are concerned). All the work must be put into the formulation of the game model in the first place, and this work must be put in for each interaction situation anew unless special circumstances make it very unlikely that context matters. There is no way to take over the results of former analyses from other contexts without further ado if we take subjectivism seriously. Assuming separability for the model in utility terms does not help much. Utilities based on the same monetary payoffs are not invariant with respect to context. The monetary payoffs may be identical as in decipede-1 and decipede-2, but the location in a larger tree may affect preferences and, thereby, utilities.

Invariance of preferences among consequences is highly unlikely if context changes. The situation would only be different if we insisted on a type of modeling that gives up on the assumption that payoffs are representative of preferences "all things considered". However, if preferences are non-satiated the choice making itself cannot be "predicted" any longer by "payoffs" and the common knowledge of the game tree. Substantive models of cognitive processes guiding human choices or leading to them would be necessary.

Let us finally return again to the first of the two aforementioned "tough questions", namely that it is unclear how we could attribute cognitive processes based on preferences to sub-personal (or super-personal) agents. An appropriate response to this is related to the preceding answer to the second problem. If we intended to model interactions from a participant's point of view but use game form models with monetary payoffs, we would be "up to our necks" in psychology. We would have to model how the participants deliberate in making up their minds when faced with certain game forms. There would be no preferences "given," they would have to be made up. The main advantage of preferences, that they sum things up "all things considered," would be gone, and the main advantage of moving towards agent-based models in RCM would evaporate into thin air, too.

In sum, relying on substantive payoffs and using (cognitive) psychology, the cognitive processes can no longer be concealed by the veil of allegedly given preferences. We cannot assume that utilities have been formed "all things considered." Being in a different setting altogether, there is no way to do game theoretic analyses in the traditional "logic of situation" sense.

In view of the preceding response to the two “tough questions” for rational choice modeling it may well be that we eventually will have to give up RCT and perhaps even RCM. The research program that provided many of the best insights of philosophy and economics may be at the brink of becoming a degenerative research program. However, until something better comes up, it is presumably wise to stick to what we have and to use it as well as we can. In doing so, it is useful to consider approaches that let preferences change systematically with some measure of substantive success. This will lead to kinds of models which are not fully in line with RCT but can be expressed within a slightly enlarged vocabulary of RCM. The underlying view of the world is based on the premise that both the subjective planning of forward-looking choice making and the objective relative successes of the past do matter. In particular we get a systematic perspective on how the rules of the games in a series of identical game forms may change in a history dependent manner. Going on with the example of the simple trust game, it is easy to illustrate what is involved in principle.[162]

[162] For details, see Berninghaus et al. (2003), Güth and Kliemt (1994), Güth and Kliemt (1998), Güth et al. (1999), Güth and Kliemt (2000a), and originally Güth and Yaari (1992).

6 Two perspectives in one[163]

If a research program gets into trouble this may lead to progress. This progress results from repairs that not only fix the old problem but add new insights. In this spirit economists who became aware of the limits of the "eductive" rational choice approach have turned to adaptive and evolutionary arguments. For instance, confronted with anomalies of individual choice making they pointed out that, though the individuals are not rational choice makers, the results of their choice making are as if originating from rational individual choices.

Nobody has ever explained how outside an – I believe inadequate – instrumentalistic view of science the "as if" argument can be a defense of the rational choice model – or, for that matter, of homo oeconomicus – in *explanatory* scientific argument. However, the ad hoc defense has led to a very fruitful discussion of evolutionary models in economics and philosophy. In particular these models can be used to build bridges between what is *objectively* and what is *subjectively* good for actors.

An early model involving both the subjective and the objective level of rational choice analysis can be accredited to Armen Alchian.[164] It is written in the spirit of Darwinian evolutionary theory.[165] Even if Alchian's argument does not hold as much water as once believed, we must address it since it inspired so many later arguments and still can serve as an inspiration for present discussions.[166]

6.1 Alchian's paradigm

Alchian suggests that we imagine a stylized market on which entities, called firms, compete with each other (see Alchian (1950)). Each firm is pursuing a fixed

[163] Though the influence of Werner Güth could be felt throughout in the preceding this chapter is clearly joint work even if Werner does this time not act explicitly as co-author.

[164] There were, of course, other such arguments in a Social Darwinist spirit. Although I will neglect that tradition here, see for instance with respect to the American case Hofstadter (1969). See also Sumner (1914). For readers of German or Spanish, it may be useful too to look at my own, Kliemt (1985), Kliemt (1986b).

[165] Until quite recently I tended to believe that the Alchian model was impervious to fundamental internal criticism. However, as Steven Durlauf and in particular Vernon Smith have made clear to me, there is much more to be said on the matter than Alchian and many of his later followers were aware; see on this Smith (2008), Radner (1998).

[166] See for a standard economic account Nelson and Winter (1982).

behavioral program (it endorses a strategy as a choreography and not merely as a plan). Assuming discrete time, after each period of interaction, profits and losses are calculated. Those firms that gain above average – assuming that profits are positive on average – will also gain in market share while the shares of those whose profits are below average shrink.[167] Without going into the details of the evolutionary process, it seems obvious that under suitable competitive conditions only programs that are doing relatively better than others in terms of objective profits will survive.[168] Those that are objectively less successful than average will have lower shares while those that are better than average will spread. Obviously, if this goes on indefinitely only the relatively best ones will survive.

Which of the programs will in fact succeed may depend on the institutional framework of the market and on the initial population composition as well as some other contingent factors. Particular outcomes cannot be predicted; they emerge. However, as in particular Friedrich August von Hayek – under the unacknowledged influence of older Social Darwinist thought – has emphasized time and again, the pattern and kind of the selection process can be predicted under a broad set of circumstances (see for instance, Hayek (1972)).[169] And the "pattern prediction" seems remarkably robust in certain aspects.[170]

The subjective side of cognitive processes, of motivations and reasons for action as such does not determine how market institutions "evaluate" objective success.[171] For instance, whether the person who operates on the maxim "quality first" does so consciously and strategically in view of the long-term reputation effects[172] or whether she does so out of a non-reflected commitment to the "rules of the trade," whether it is altruism or egoism that is driving her, in the end the selection process will evaluate alternative forms of behavior according to its "*objective* standards."[173]

[167] Of course, nowadays we also think of genetic algorithms in such contexts, see Holland (1975)

[168] To put it slightly otherwise, only population compositions in which no local deviation can increase profits against the given behavior of others can be evolutionarily stable.

[169] On pattern emergence, see Schelling (1978) See also the fine discussion of the latter in Sugden (2002), also Flache and Hegselmann (1998), and from a biological point of view in the same spirit Eigen and Winkler (1975).

[170] As experimental markets show, markets may clear and reach equilibrium quite independently in particular of information conditions, see for instance, Smith (2000), Kagel and Roth (1995).

[171] Market institutions are so robust that even with "zero intelligence traders," markets clear, see Gode and Sunder (1993).

[172] See on reputation the fine anthology Klein (1997).

[173] To provide a specific example, in German car companies for many years the engineers dominated the policy decisions of the companies. There were those who aspired to build good cars according to their engineering standards, yet their aspirations did not lead to profit for the company. This seemed a recipe for disaster in the eyes of economists who had a keen eye on short-term profits. However, to go beyond what the market required in a short- term perspective may have been to a large extent the driving force behind the success of those companies whose policies were led by "economically incompetent engineers."

In sum, in a selective, competitive environment, individuals who subjectively intend to do otherwise than to maximize an "objective function" may in an evolutionary sense be "objectively" more successful than "opportunists".

Nobody has seen the force of behavioral adaptation more clearly than Joseph Alois Schumpeter. Even subjectively maximizing behavior is – as far as it is successful – often an adaptation rather than the result of forward looking rational calculation. Here is what Schumpeter says (Schumpeter (1959), 80):

> The assumption that conduct is prompt and rational is in all cases a fiction. But it proves sufficiently near to reality, if things have time to hammer logic into men. Where this has happened, and within the limits it has happened, one may rest content with this fiction and build theories upon it … and we can depend upon it that the peasant sells his calf just as cunningly and egoistically as the stock exchange member his portfolio of shares. But this holds good only where precedents without number have formed conduct through decades and, in fundamentals, through hundreds and thousands of years, and have eliminated unadapted behavior. Outside of these limits our fiction loses its closeness to reality.

We may try to use some tools of RCM to describe the process operative "to hammer logic into men" somewhat more clearly. When doing so, it is necessary to make an effort not to throw the baby out with the bath water though. That is, even if we include adaptive forces and their evolutionary modeling, we should still take into account that humans have foresight and understanding. They are drawn by the expected future and are not merely driven by the past; they are pulled *and* pushed.

6.2 Evolution of individual rule-following behavior

Assuming that rule-following behavior does exist, it must be understood how that behavior *can* prevail if opportunism is possible, too. In particular, why is it so that rule-bound behavior can survive on the individual level if institutional rules provide a competitive environment in which uncommitted behavior may co-exist and be directly advantageous?

In terms of economic modeling, the basic challenge is to give an account of how individual rule-following behavior and the disposition to forego opportunities in particular cases can survive in view of the fact that – except for pure co-ordination rules – deviation from the rules remains advantageous as compared with a commitment to rules. Why do those individuals who are uncommitted not "out-compete" those who are committed to foregoing opportunities? If seizing opportunities is – when the opportunity arises – almost by definition advantageous with respect to the very opportunity, why don't opportunists drive out the rule-followers? Such are the questions that must be answered.

Since it seems to me that there is no way to demonstrate generally that the abstract disposition to obey rules can be evolutionarily stable, I will focus on one paradigm. Relying on ideas that originated basically with Werner Güth, I will sketch how a general disposition to abide by rules can conceivably be evolutionarily stable.[174] Doing so, I do not aspire to show that having the disposition to abide by rules is in itself a "good thing." I will only illustrate that such a broadly Kantian disposition can conceivably meet the ultimate test of surviving even in the large numbers setting of a Great Society. It can survive as a disposition in circumstances in which the disposition to decide cases separately on their own merits will be a competitor for evolutionary success.

As we will see, there must always be a niche for those disposed to exercise case-by-case discretion (moral or other). Yet, under suitable conditions of knowledge processing in society, the "moral" virtue of a general commitment to abide by rules can survive and even dominate in a population.

6.2.1 The intuitive explanation of evolutionary stability of rule following

If a person is at all able to commit to rule following behavior, she may make good use of her individual commitments. They may help her to pursue her long-term interests in view of opportunities to which her own future agents might give in otherwise (see for an intuitively plausible statement of this argument Frank (1987), Frank (1988)).[175] For instance, the ability to follow dietary rules may be helpful in view of the many Ulysses problems that we face. Here it is protection against break down of the will that is provided by commitment power (see Ainslee (2002)).

Yet, there is another more important aspect of the ability to commit. *If and in so far as other individuals can recognize an individual's commitment to abide by ('moral') rules*, such non-opportunistic individuals should be in high demand as transaction partners. It is this demand for their partnership that may give them the competitive edge over uncommitted individuals.

Although not being able to seize an opportunity when it is actually offered is *always dis*advantageous on the occasion of the specific opportunity, it can nevertheless be advantageous to be committed to rules if co-operation under such

[174] Note that this *general* disposition to abide by rules independently of their content is one level up or one level removed from the disposition to abide by specific rules.

[175] With respect to rational self-management, see Schelling (1984).

a restriction is still better than no co-operation at all.[176] As is triumphantly obvious from the prisoner's dilemma, those who can engage co-operation in rule-bound ways will have the competitive edge over those who cannot as long as they can prevent unilateral defection or the risk thereof.[177] However, to point out that, for example, in a two-by-two prisoner's dilemma, mutual co-operation is better than bi-lateral non-co-operation does per se not eliminate the dominance of the defection strategy and the "temptation" of unilateral defection.

The crucial step in the argument is that commitment to rule-following behavior (amounting to individual constitutional constraints on opportunism) can be recognized by others[178]. For those individuals who are stuck with the in-period disadvantage of being unable to exploit opportunities, greater opportunities of co-operation may open up. This happens because they are sought after by other forward-looking rational choice makers – for strategic reasons or because they are committed to interact only with the committed – as partners and, thereby, receive higher objective payoffs than when left out.

In sum, the general disposition to show rule-following behavior can be "good for" the individual in both the subjective and the objective sense. To have the generalized disposition to abide by the rules (in the presence of "temptations" to exploit others unilaterally) can be advantageous for the individual because and to the extent that others can discriminate between those who have the disposition to follow rules and those who do not.

The preceding intuitive argument can be presented in somewhat more precise terms, and it is to a very brief sketch of exactly this that I next turn.

6.2.2 Evolutionary stability of individual rule-following in a simple model

As far as the basic problem of social order is concerned, the crucial commitment is the commitment to observing a rule of executing explicit or implicit promises to play by the rules of established practices even if and in so far as this does not have any *direct* causal future consequences for the rule-abiding actor. I will refer to this commitment as **trustworthiness**. As is well-known, the general prevalence of such a disposition in society can be most conducive to the welfare of individuals in that society. Societies that are characterized by a high degree of trustworthiness

[176] A point impressively made and illustrated in Baurmann (2002).

[177] In view of the fact that exchange as well as contract have a prisoner's dilemma structure (see on this (Hardin (1982), Kliemt (1986a)), this observation is of general relevance for co-operation in (the great) society and its moral and legal super-structure. See also in an evolutionary biology spirit Ofek (2001).

[178] On personal constitutional commitments, see also Vanberg and Buchanan (1988).

tend to be richer and more generally speaking "better working" on almost all accounts than those in which case-by-case opportunism is more prevalent.[179]

It would, so to say, be "convenient" if individuals would behave in trustworthy ways. Yet, how trustworthiness can be maintained if it is superior to be untrustworthy in particular instances is still an open question. This fundamental question is not answered by pointing out the general advantage of living in a society of trustworthy individuals. Social good in this sense does not directly translate into what is good for individuals. We must still explain how those crucial individual restrictions on unilaterally exploitative behavior – without which the moral and legal order of a Great Society could not conceivably work – can actually be maintained.[180]

The core of the problem of trust (worthiness) can be demonstrated by the game represented by the tree of figure 6.1. The first moving individual i plays the role of the trustor, and the second moving individual j the role of the trustee. The objective (substantive or material) payoffs are in the order $0<s<1$ for the first mover and $0<r<1$ for the second mover.[181] When showing trust, T, the player i makes herself vulnerable to exploitation, for by choosing E, the trustee may bring her down to 0 although she could guarantee herself a payoff of s. If the trustor trusts, she is hoping to receive the co-operative payoff of 1 accruing after R. The trustee will receive nothing if no trust is shown while his objective payoff will be r after fairly rewarding trust or $1>r$ if exploiting it.

For our present concerns, the crucial point is that the payoffs at the end of the game tree – which show up in the sequence corresponding to (i, j) – are meant to represent two (!) value functions each at the same time. There is, to put it paradoxically, an "objective objective function" in substantive payoffs, and there is a "subjective objective function" in preference representing payoffs for each of the individuals, i, j. For instance, for i the first value function measures what is good for the actor i in the substantive sense directly related to evolutionary success. The second function for i measures the "apparent good" or the evaluation as made by the individual i in her evaluative judgments. The same applies for j. This individual also has subjective preferences – captured by the one interpretation of the numerical values – and is objectively "evaluated" in terms of (relative) objective success – as captured by the other interpretation.

This mirrors the dual approach of the Alchian model. What is "good for" the individual in the judgmental sense is relevant for her intentional behavior and its consequences, while what is "good for" the individual in terms of objective success is directly bringing about the consequences in the relevant environment.

[179] See for a somewhat eclectic approach Fukuyama (1995) and for the "full truth on trust", Lahno (2002).

[180] On individualistic foundations of evolutionary economics, see also Witt (1987).

[181] With respect to the objective payoffs, one may think of money earned or offspring etc.

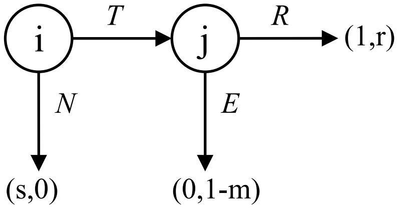

Figure 6.1: The simple game of trust with subjective parameter m

So, assume that all the measures of "good for" in the objective sense – directly relevant for evolution – also represent good for in the subjective sense – only indirectly related to evolutionary success via intentional behavior. When playing E, an additional purely subjective factor $m \geq 0$, which represents the trustee's internal rule-following motivation is relevant. There is *no* objective side to that. The payoff in objective terms is "1" in subjective terms it is, however, "1 – m." It is assumed that m is either strong enough to restrain "exploitation," i.e. $m \geq 1 - r$, or too weak, i.e. $m < 1 - r$ (in which case we can assume $m = 0$ since behaviorally the specific value of m does not matter). [182]

Rule-following is intentional. The effects the intentions have on preferences show up in the purely subjective parameter "m." The parameter m *represents* whatever it is that brings individuals to behave differently from what the value function as formulated in substantive payoffs (or what is objectively good for the actor) would dictate. An additional dimension of evaluation – representing generalized constraints – enters into decision making once the chance of exploiting a trusting first mover emerges. Only if the opportunity of exploitation is reliably foregone in such situations can the social good of inter-individual trust-based co-operation be realized in society.

In an evolutionary setting in which evolutionary pressure towards opportunistic behavior is present, the adherent to a theory that is based on "moral behavior" must show how individual rule-following behavior can be to the advantage of the individuals who show that behavior. Commitments to rule-following behavior must pay relatively better for individuals than dispositions to take opportunities whenever they come up. To get a more precise handle on the problem, assume:

> *There is an infinite population of individuals* (modeling that there are many)

[182] In view of the fact that subjective, preference-representing utility functions are unique only up to positive affine transformations, there are degrees of freedom allowing for re-scaling the representing function such that objective and subjective payoff measures would look more divergent without any alteration in the substance of the argument.

> *There is an infinite number of rounds of play* (modeling the long perspective)
>
> *The individuals are randomly matched to play a simple trust game on each round of the evolutionary process* (modeling settings like large anonymous markets in which the shadow of the future does not do the trick and, therefore genuine trustworthiness is necessary)[183]
>
> *The individuals know the type composition of the population or the general share p of trustworthy individuals* (modeling that they are strategic actors rather than automata).

These formally rather demanding conditions are introduced because, otherwise, a closed model could not be formulated. Each of the conditions can be weakened in theory and approximated in the laboratory. Yet, making the model "more realistic" is not what I am interested in here. My focus is rather on the general patterns of evolutionary processes that emerge and whether in such processes generalized trustworthiness will not be driven out by particular or case-by-case maximization.

All depends on how well an individual in the role of the trustor can discriminate between committed and uncommitted types. Distinguishing three qualitatively different conditions the following observations can be made:

Condition 1: Extreme case of perfect type information

If individuals have perfect information about their randomly assigned co-player's type, then all individuals in first mover roles will always and *exclusively* co-operate with trustworthy individuals. Thus, on every round of play, those committed to playing by the rules will gain more in the second mover role than those who are not so committed. Over the long haul, the trustworthy will out-compete those who are not committed to rule-following behavior of that kind.

To put it slightly otherwise, if virtue can be recognized perfectly and without charge by potential partners, then it pays to be virtuous.[184] This seems obvious enough, but one might want to note that it is not sufficient to behave "as if" the virtuous character trait were present. It is necessary to actually *be* virtuous because

[183] Of course, there could be individuals who freely choose to interact with each other for extended periods of time under conditions of free entry and exit. This would create a shadow of the future for these interactions by the free choice of individuals. I focus here on the extreme case of one-shot interactions since being able to trust even in such interactions is the hallmark of a great, free contract society which exploits the advantages of the division of labor to their full extent, and it is also a way to keep the group selection effect completely out and to stick to the individual; see on Axelrod type models with free exit and entry originally Schüssler (1990), and, in the same spirit, Vanberg and Congleton (1992)

[184] This is akin to the extreme conditions of Gauthier (1986) but puts them into evolutionary perspective.

the demand is for true virtue rather than for the mere appearance thereof.[185] If such commitments as a matter of fact are "technologically" viable for individuals of the human kind and their presence can be known, then trustworthiness can be sustained in evolutionarily stable equilibrium. Under conditions of perfect type information, the selection of partners will drive out the untrustworthy.

In sum, if there is perfect type discrimination, then the evolutionary dynamics are such that only the monomophic population of exclusively trustworthy types will be evolutionarily stable. If p is the share of trustworthy types, then we get

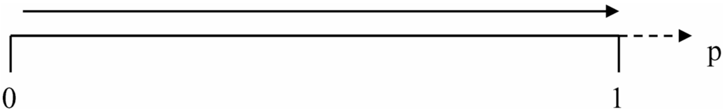

Figure 6.2: Population dynamics under perfect type information

To state the preceding in even more simple terms, whenever there is a trustworthy individual, then this individual will be singled out by others. The trustworthy will find better terms in interaction with others than the non-trustworthy. The proportion p of trustworthy individuals will go straight towards 1 if the presence of the quality can be detected perfectly.

Condition 2: Extreme case of completely private type information

If the personal virtues are purely private information, then individuals will trust when assigned first mover roles as long as sufficiently many virtuous individuals are around, i.e. as long as the expected value of showing trust in first mover roles, which is $p1 + (1 – p)0$, is greater than s, the payoff accruing from choosing N. Being informed about the type composition, they will not trust if only insufficiently many trustworthy individuals are around. It is not worthwhile then to show trusting behavior, i.e. if $s > p$, no trust will be shown since it would be a bad bet. As long as all show trust in the first mover role, i.e. as long as $p > s$, those who are non-trustworthy will always fare better than the trustworthy. That advantage will be slight if the type-composition of the population becomes such that nobody will rationally bet on trustworthiness anymore, i.e. once $s > p$. In that case, the process must be mistake-driven (a kind of trembling as in Selten (1983)). Once in a while some would show trusting behavior by mistake and since the trustworthy fare worse than the untrustworthy in these rare cases, their population share will slowly decline.

[185] "Gang of four" type arguments, according to which a suitable uncertainty about the presence of true commitments is sufficient to induce individuals to behave as if committed would not work here, see Kreps et al. (1982), Kreps and Wilson (1982). According to such arguments it would be necessary that some truly committed individuals exist and therefore the crucial assumptions that such commitments are in fact possible must be made in any event.

In sum, the presence of the virtue of trustworthiness in society cannot be sustained at all in evolutionarily stable equilibrium if the presence of that quality cannot be detected in an individual. The only evolutionarily stable population composition is characterized by a population parameter p* = 0.

In the figure below, the dotted arrow indicates the threshold from which point on evolution is merely mistake driven.

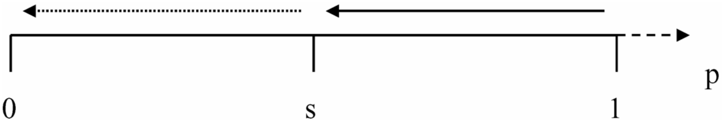

Figure 6.3: The population dynamics under private type information

Condition 3: Intermediate case of some type information at some cost
The most interesting case is, of course, the intermediate – or non-extreme – one in which some specific, imperfectly reliable type information about the randomly assigned co-player is available at some cost. The crucial parameters then are the costs of using the technology of type recognition C and its reliability (which, for the sake of simplicity, I take as given here). The triangle in the next figure depends in its shape and height on these parameters. As long as p < s, trust is a bad bet and individuals would not trust unless they received a signal that the specific partner they are facing is trustworthy.

Up to p = s, the technology would be used to find trustworthy individuals. If p > s, then it would be used to discriminate against the untrustworthy.

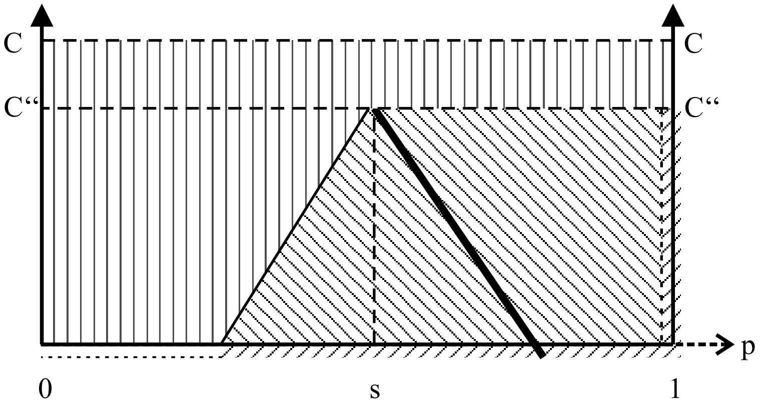

Figure 6.4: The population dynamics with imperfect type information

To the left and to the right of the triangle, the technology is too costly to be used. To the left, the gain in expected value is insufficient because finding a trustworthy individual by means of the technology is too unlikely. To the right, finding an untrustworthy individual is too unlikely to make it worthwhile to expose the untrustworthy at cost C. To the left, the decline of the trustworthy will be slow mistake-driven, but to the right, it will be fast since all the untrustworthy will be trusted all the time and will gain the extra payoff of exploitation without being ever discriminated against.

Now, consider – for a given reliability – the population and cost pairs (p, C). If (p, C) lies within the triangle, then the technology will be utilized. Drawing a horizontal line from an initial p_0 at the height of the cost parameter C towards the fat right border of the triangle will give the evolutionarily stable type composition parameter p* for that C and initial p_0 at (p*, C) on the fat line. If the initial parameter p_0 lies to the right of this point, the population share of the trustworthy will shrink to p* and then stabilize at (p*, C) since individuals will start to utilize the technology in the trustor role when p = p* at cost C. If the initial parameter p_0 is too low to make it worthwhile to use the technology, then the only evolutionarily stable population share of trustworthy individuals is p* = 0.

For costs higher than C´´, the technology will never be used. Moreover, the higher C´´ and the steeper the sides of the triangle, the less likely is it that the population share can be stabilized at some p* > 0.

In sum, the process of the evolution of a share of committed individuals in the population is *path-dependent.* The process will lead to an evolutionarily stable positive share of trustworthiness in society only if the initial conditions p_0 are such that under the given technology it pays to invest in control.

The preceding model suffices to illustrate that a general disposition to abide by the rules is not only desirable, it is also viable within the social evolutionary world even of a Great Society.[186] It is true enough that there will always be a niche either for individuals who are not rule abiding – in the case of a bi-morphism with some trustworthy and some non-trustworthy (0 < p < 1) – or for some amount of non-rule abiding opportunism – in the case of a monomorphic population playing a mixed behavioral strategy. However, within the limits of the possible, we may expect some of the crucial social good of individual rule abiding behavior to be maintained in evolutionarily stable equilibrium.

I am content to let it rest with this defense of the merits and viability of strict rule-following. It seems that an indirect evolutionary model can be used to show that the *generalized* disposition of strict rule obedience can itself meet the test of evolutionary survival. Since it meets this test, it can be regarded as a virtue in objective evolutionary terms. In addition, as long as and to the extent that this virtue is present in society, the spontaneous co-operation under rules without

[186] A specific real world example is eBay with its costly reputation mechanism; see for a more detailed account Güth and Kliemt (2004) Interestingly enough, eBay obviously started with a very high p_0. This was the good luck or path-dependence part.

external enforcement can exist. Small group moral dispositions need not necessarily be eroded by living in a Great Society.

In sum, a generalized disposition to abide by rules can survive as an evolutionarily stable outcome in evolutionary competition in a Great Society. Though there is a niche for opportunism (or opportunists), there is one, too, for trustworthy individuals and mutually advantageous co-operation under a general intrinsic motivation to abide by the evolved specific rules.

The preceding shows how in principle the subjective and the objective dimension of choice making can be related to each other. The indirect evolutionary approach is a way to bridge the gap between the two ways of world making, the subjective and the objective perspective. This is important in itself. Yet it also shows that there is in all likelihood a niche for genuine rule following behavior and how such behavior can be incorporated into RCM. Before I turn to morals on the basis of what I have said about methods and models let me draw some conclusions concerning the first volume.

7 Conclusions

7.1 Rational community and the participant's attitude

In distributed computing several computers must coordinate their tasks and, therefore, in some way or other "know" what the others are doing and what they "know." However, it would seem somewhat far-fetched if we claimed that in understanding the coordination of the several machines we put ourselves in their several shoes. Of course, we all tend to talk to computers and to shout at them if programmed by our common friend and enemy Bill ("barbarians at the gates"); however, we do not in earnest form some kind of community with them. Nor do we, when we observe computers communicating with each other, try to understand what they do by emulating their behavior, at least we do not do so in the full sense of that term. We may, however, go through the steps of the same algorithms that we used to program them.

Returning to reasoning about knowledge amongst humans, we can observe that we, like the computers, try as well to reach some common ground of reasoning. As far as that is concerned, it has been argued that some ascription of rationality plays a crucial role in particular in game theoretic modeling from a participant's point of view. However, ascribing some kind of ideal reasoning process symmetrically to all players in the game, it becomes very unclear whether we as analysts can truly adopt a participant's attitude to such an idealized interaction. After all, we are as a matter of fact only boundedly rational and not perfectly rational beings ourselves. How could we participate then in the full sense or at least emulate participation in the reasoning processes of such an idealized knowledge community?

It seems that we are almost as far detached from such an interaction as from that between a set of computers engaged in performing some task of distributed but coordinated "reasoning." Moreover, according to the way we normally use the common knowledge assumption along with that of symmetrically rational, and, for that matter, perfectly rational individuals, each and every individual is assumed to reason the same way about the game. We in effect have reduced the problem of reasoning in an interactive situation to the reasoning of a representative ideal individual who knows the game in full and shares this knowledge by virtue of the common knowledge assumption with each and every other participant. The game theorist and the participants in the game are in the same situation. Everybody comes exactly to the same conclusions as everybody else when thinking about the game before the specific play of the game starts.

In sum, as far as the reasoning itself is concerned we are not talking about some interactive reasoning practice. It is rather an ideal type of reasoning to which all ideal type reasoners are assumed to "converge." It is the reasoning of a representative ideally rational individual.

The higher forms of reasoning based on models of the action situation seem to be uniquely human. As opposed to behavior in an actual interaction – a play of the game – analyzing the game beforehand requires higher faculties of understanding, which only humans seem to command.

Pushing to its limits in theory the idealization of the specifically human faculties, in particular, those characteristically human abilities to which the principle of intervention refers, may well be seen as expressing an ideal type of analysis performed with a participant's attitude. After all, we are talking about reasoning and, thus, "behavior" that is following its own logic (i.e. reasons) rather than about behavior occurring according to behavioral laws. We are dealing with a community that might be and not with one that is. Yet, that we are doing this is in itself a fact. Theoretical reflection does exist and thus what might be can exert an influence on the real world when envisioned as such.

7.2 Moral community and the participant's attitude

It does not seem to be personhood per se but rather membership in a *community* of persons allowing for specific interpersonal relationships that is bringing about the difference in our attitudes towards other participants of interaction. There must be a quality in the relationship between Crusoe and Friday that is absent in that between Crusoe and a chimp. This difference must allow for approaching the other person with a participant's attitude.

As is already clear from the fact that at least in some classification systems chimpanzees would qualify as persons, not all sets of persons can form a community. This raises the question of how to characterize the relevant communities and membership in them. According to one possibility, membership in the community would require merely being human. However, in view of the fact that some humans – as opposed to some chimpanzees – quite obviously are not endowed with the characteristics of personhood, being human can only be a necessary condition for membership in the relevant community. A somewhat more convincing minimum condition seems to be that members of the relevant community must be both: human and a person. Yet, this raises two follow-up questions. On the one hand, a Kantian might argue that the relevant community is that of all rational beings and, thus, goes beyond human kind – though not including chimpanzees and the like; on the other hand, the relevant community might still be some subgroup of that of all human persons.

Though it may well be that in the future we can make contact with rational beings from some other galaxy or perhaps create computers who are rational persons[187], those possibilities are sufficiently utopian or remote to be dismissed without further ado in the present context. We should focus on the issue of how to determine the relevant community of humans.

In the spirit of modern times and our human rights' declarations, it seems as if only the community of all rational human beings might be chosen. However, it is perhaps not by accident that many so-called primitive tribes seem to classify those who are not members of their own tribe as non-human rather than as simply humans of another group. Human nature, therefore, clearly does not rule out the possibility of forming subgroups of rational human beings, which are seen as the exclusive recipients of that interpersonal respect, which we express when approaching another individual with the participant's attitude.

Speaking the same language clearly facilitates adopting a participant's attitude to an interaction, but that condition is not a sufficient one, even if it is for the simple reason that we can always also approach another individual with an objective attitude. It seems rather implausible also that sharing a common language is necessary for forming a community with another person such that adopting a participant's attitude becomes viable[188].

Yet, what else could be the basis of such a common understanding? Even though there may be some biologically fixed ways of all facial and other corporal expressions that signal specific emotions in a way that all humans naturally understand, this cannot be what sets an interaction among persons who approach each other with a participant's attitude apart from other forms of interaction. In fact, we all naturally understand the natural signals of anger that a dog or a chimpanzee may send. However, nobody would say that this naturally brings us to the point where we adopt the participant's attitude towards interaction with them. We adopt that attitude only when there is some common ground for signaling and understanding rather complex intentions and attitudes of humans that form some kind of community with us.

[187] By passing, say, the Turing test according to which human actors communicating with the program from a distance could not find out that they are not interacting with a human being.

[188] Going back to the original story of Crusoe and Friday, the community between the two cannot be based on speaking the same language. So, if there is some common understanding allowing for adopting a participant's attitude towards the interaction with each other, that common understanding cannot have been created by verbal means in the ordinary sense.

The assumption that among humans there is a natural way of signaling more complicated intentions seems to be far-fetched. Any common understanding between Crusoe and Friday, if there is any, that would allow for adopting a true participant's attitude to each other, must be based on something other than and beyond a common language or common natural signs. It is quite telling that in his original story Daniel Defoe implicitly makes unwarranted assumptions about conventional meanings of acts taken as signals. For instance, he seems to assume that by rescuing Friday, Crusoe signals his good intentions to Friday. According to Defoe's own cultural background, it seems indeed natural that Crusoe by rescuing Friday from being eaten by the cannibals makes his resentment against cannibalism pretty clear. However, for Friday in his own cultural background the most plausible story must have been that he was on Crusoe's menu. Why in the world should somebody risk his own life to rescue another completely alien individual if not to secure some essential benefit to himself?

This shows that shared interest or, for that matter, sympathy are not crucial. Friday, ascribing to Crusoe the aim of using him as a source of protein as the most plausible explanation for Crusoe's act is expressing a participant's attitude towards another. For, the actions of the other are understood in terms of a teleological framework. Moreover, Friday, putting Crusoe's foot on his neck seems to express his submission from the point of view of persons like Crusoe. However, as we know, even such gestures have conventional meaning only. In some societies, shaking your head means the same as nodding it in others. Likewise, putting the foot of another on your neck, rather than being a gesture of submission, might be a claim to superiority in some culture or other.

It may be that the mere ability to speak some language at all – perhaps being in command of the rules of the universal deep grammar of language if there is any – is constitutive for some kind of moral community membership and, in that sense, sufficient for delineating the relevant group of persons. Possibly trade and mutual advantage are enough for forming a kind of moral community.

Instead of going on and on with reflections let us simply note that we are dealing with an issue that is possibly decided in quite arbitrary ways by the fact that humans do perceive themselves in a moral community with some individuals and possibly not with others. The person who perceives herself in certain ways as being in a "moral" community with another individual approaches the individual differently, at least in part according to her own whim. It may well be that some perceive themselves as being in a community with their pets. At the same time, they may be of the opinion that a human being coming from another culture has nothing in common with them or at least not sufficiently so that they could approach that human being in ways other than strategic manipulation. They may feel that they do not share enough with the other even to deal with her or him strategically in the full sense of the term. Symmetry may just not make sense.

Crusoe may not make distinctions between Friday and the chimpanzee and vice versa.

In sum, the limits of the relevant community to which we adopt a participant's and possibly a moral attitude seem to be self-selected. Who is in and who is out of the range of receiving personal respect is decided by those who are *in fact* adopting the participant's attitude towards others.

Going back to the initial example again, it seems that Crusoe could do one of three things: either always show an objective attitude towards Friday, or always adopt a participant's point of view towards Friday, or sometimes the one and sometimes the other. The last of the three possibilities seems to be the one most in line with regular attitudes of human beings towards each other. It is not the case that they exclusively approach each other with the participant's attitude. They may well try to manipulate each other with an objective attitude once in a while. However, as long as they command the ability to look at each other in ways other than manipulative ones and sometimes exercise that faculty, human interpersonal interaction gains a quality absent in other relations either among humans or among humans and other beings. This other quality will become visible if we look at the world through the participant's window. There is nothing that forces us to look at the world from this point of view. Yet, if we do so, we will see different things and hear different voices whispering different suggestions into our ears than otherwise (see for instance Gehlen (1978)). And, it would be a gross mistake to ignore this fact in dealing with idealized rational choice modeling.

7.3 Against nature or other minds?

Whether there would in fact be a role for anything but playing games against nature from an objective point of view among ideally rational individuals is hard to say. Yet, classical political economy, classical political philosophy as well as classical game theory are based on another conceptualization of action. They can play a legitimate role among boundedly rational humans because for them they are a way to look beyond their own limitations (notwithstanding the fact that they remain within these limits when doing so). Taking this look of what there might be in principle (but not in fact) is not scientific if science is restricted to fact finding. Yet, whether it be science or not, there is a legitimate role for a non-science within RCT.

Pursuing the issues of this non-science, we may be well aware that we are as a matter of fact merely boundedly rational individuals. At the same time, we may as a matter of fact be interested in spelling out the requirements of rationality in a setting in which all individuals symmetrically command the same form of ideal rationality and behave accordingly. What we are looking for under this "contrary to fact" assumption are theories that are fully absorbable under ideal conditions and in wide reflective equilibrium (see on this concept also volume 2). These are theories that can be known to all individuals without providing any of them a reason to alter them or an incentive to deviate from what the theory predicts and suggests.

Note that besides the philosophical interest that the fiction of a world of ideally rational beings per se may command, it can be of interest for beings with limited capacities as we are in their ordinary lives as well. For, even though we may not be able to live up to the theories and even though they may be rather far away from our actual behavior, the theories of ideal behavior may play a crucial role in shaping our attitudes towards other rational beings in social interaction. Ascribing to them what they may not actually have as a property, we will deal with them in specific ways that differ from other ways. Living in a different perceptual world may influence behavior and, in this stronger sense, the world itself. An illusion about the facts may distort the facts but that it as such prevails can be a fact nevertheless. An idealization may only be in our heads and not "out there" (not even approximately!), but it is as a matter of fact in our heads. This subjective aspect is part of the objective world. As we shall see in the next volume, this holds good as well for moral evaluations, perceptions, and ideals.

8 References to volume 1

Ainslee, G. (1992): Picoeconomics. Cambridge.

Ainslee, G. (2002): Break Down of the Will. Princeton.

Albert, H. (1967): Marktsoziologie und Entscheidungslogik. Neuwied/Berlin.

Albert, H. (1985): Treatise on Critical Reason. Princeton.

Alchian, A. A. (1950): Uncertainty, Evolution, and Economic Theory. Journal of Political Economy, Vol. 58, 211–221.

Alchian, A. A. (1984): Specificity, Specialization, and Coalitions. Zeitschrift für die gesamte Staatswissenschaft, Vol. 140, 34 ff.

Alchian, A. A. and Woodward, S. (1988): The Firm Is Dead; Long Live the Firm. A Review of Oliver E. Williamson's "The Economic Institutions of Capitalism". Journal of Economic Literature, Vol. XXVI(March), 65–79.

Arendt, H. (1951): The Origins of Totalitarianism. New York.

Aumann, R. J., 1981, Survey of Repeated Games. In: Robert et al. Aumann (Ed.), Essays in Game Theory and Mathematical Economics. Bibliographisches Institut BI, Mannheim, pp. 11–42.

Aumann, R. J. (1987): Correlated Equilibrium as an Expression of Bayesian Rationality. Econometrica, 55(1), 1–18.

Axelrod, R. (1984): The Evolution of Cooperation. New York.

Baurmann, M. (2002): The Market of Virtue, vol. 60. Dordrecht.

Berninghaus, S., Güth, W. and Kliemt, H. (2003): From teleology to evolution. Bridging the gap between rationality and adaptation in social explanation. Journal of Evolutionary Economics, 13(4), 385–410.

Beth, E. W. (1965): The Foundations of Mathematics. Amsterdam.

Binmore, K. (1987/88): Modeling rational players I&II. Economics and Philosophy, 1987/88 (3 & 4), 179–214 & 179–155.

Binmore, K. (1992): Fun and Games – A Text on Game Theory. Lexington.

Binmore, K. (1994): Game Theory and Social Contract Volume I – Playing Fair. Cambridge, London.

Binmore, K. (1998): Game Theory and Social Contract Volume II – Just Playing. Cambridge, London.

Binmore, K. (2005): Natural Justice. New York.

Brennan, G. and Hamlin, A. (2000): Democratic Devices and Desires. Cambridge.

Brennan, G. and Pettit, P. (2006): The Economy of Esteem. Oxford.

Brennan, H. G. and Kliemt, H. (1994): Finite Lives and Social Institutions. Kyklos, 47(4), 551–571.

Brennan, H. G. and Lomasky, L. (1984): Inefficient Unanimity. Journal of Applied Philosophy, 1(1), 151–163.

Brennan, H. G. and Lomasky, L. E. (1993): Democracy and Decision. Cambridge.

Buchanan, J. M. (1965): Ethics, Expected Values, and Large Numbers. Ethics, LXXVI, 1–13.

Buchanan, J. M. (1975/1996): An Ambiguity in Sen's Alleged Proof of the Impossibility of a Pareto Liberal. Analyse & Kritik, 18(1), 118–125.

Buchanan, J. M. (1985): What Should Economists Do. Indianapolis.

Buchanan, J. M. (1999): The Logical Foundations of Constitutional Liberty, vol. 1. Indianapolis.

Buchanan, J. M. (1999 ff.): The Collected Works of James M. Buchanan. Indianapolis.

Buchanan, J. M. (2001): Game Theory, Mathematics, and Economics. Journal of Economic Methodology, 8(1), 27–32.

Buchanan, J. M., Güth, W., Kliemt, H., Schwödiauer, G. and Selten, R. (2001): John von Neumanns und Oskar Morgensterns 'Theory of Games and Economic Behavior'. Düsseldorf.

Coleman, J. S. (1988): Free Riders and Zealots: The Role of Social Networks. Sociological Theory, 6(Spring), 52–57.

Dacey, R. (1976): Theory Absorption and the Testability of Economic Theory. Zeitschrift für Nationalökonomie, 36(3-4), 247–267.

Dacey, R., 1981, Some Implications of 'Theory Absorption' for Economic Theory and the Economics of Information. In: Joseph C. Pitt (Ed.), Philosophy in Economics. D. Reidel, Dordrecht, pp. 111–136.

Danielson, P. A., 1998, Introduction to Modeling Rationality, Morality and Evolution. In: Peter A. Danielson (Ed.), Modeling Rationality, Morality and Evolution. Oxford University Press, New York and Oxford, pp. 3–9.

Davis, D. D. and Holt, C. A. (1993): Experimental Economics. Princeton.

de Jasay, A. (1995): Social Contract – Free Ride. Oxford.

de Jasay, A. (1997): Against Politics: On Government Anarchy and Order, vol. 7. London and New York.

Diekmann, A. (1985): Volunteer`s Dilemma. Journal of Conflict Resolution, Vol. 29/4, December 1985, 605–610.

Dixit, A. K. and Nalebuff, B. (1991): Thinking Strategically. New York.

Eigen, M. and Winkler, R. (1975): Das Spiel. Naturgesetze steuern den Zufall. München.

Fagin, R., Halpern, J. Y., Moses, Y. and Vardi, M. Y. (1995): Reasoning about Knowledge. Cambridge, MA / London.

Fehr, E. and Gächter, S. (2002): Altruistic Punishment in Humans. Nature, 415(January), 137–140.

Flache, A. and Hegselmann, R. (1998): Understanding Complex Social Dynamics – A Plea For Cellular Automata Based Modelling. Journal of Artificial Societies and Social Simulation, 3.

Frank, R. (1987): If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? The American Economic Review, 77/4, 593–604.

Frank, R. (1988): The Passions within Reason: Prisoner's Dilemmas and the Strategic Role of the Emotions. New York.

Fukuyama, F. (1995): Trust. The Social Virtues and the Creation of Prosperity. New York.

Gaus, G. (2008): On philosophy, politics, and economics. Belmont, Ca.

Gauthier, D. P. (1969): The Logic of Leviathan.

Gauthier, D. P. (1986): Morals by Agreement. Oxford.

Gehlen, A. (1978): Der Mensch. Wiesbaden.

Gigerenzer, G. and al., e. (1989): The Empire of Chance. How probability changed science and everyday life. Cambridge.

Gode, D. K. and Sunder, S. (1993): Allocative Efficiency of Markets With Zero Intelligence Traders: Markets as a Partial Substitute for Individual Rationality. Journal of Political Economy, 101, 119–137.

Granovetter, M. (1985): Economic action and social structure: The problem of embeddedness. American Journal of Sociology, 91(3), 481–510.

Güth, W. (2000): Boundedly Rational Decision Emergence – A General Perspective and some Selective Illustrations. Journal of Economic Psychology, 21, 433 – 458.

Güth, W. and Kliemt, H. (1994): Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes. Metroeconomica, 45(2), 155–187.

Güth, W. and Kliemt, H. (1998): Towards a Fully Indirect Evolutionary Approach. Rationality and Society, 10(3), 377–399.

Güth, W. and Kliemt, H. (2000a): Evolutionarily Stable Co-operative Commitments. Theory and Decision, 49, 197–221.

Güth, W. and Kliemt, H., 2000b. From full to bounded rationality. The limits of unlimited rationality, Center for Interdisciplinary Research (ZiF), Bielefeld.

Güth, W. and Kliemt, H. (2004): The Evolution of Trust(worthness) in the Net. Analyse & Kritik, 26, 203 – 219.

Güth, W. and Kliemt, H., 2007, The Rationality of Rational Fools. In: Fabienne Peter and Hans Bernhard Schmid (Eds.). Oxford University Press, Oxford, pp. 124–149.

Güth, W., Kliemt, H. and Ockenfels, A. (2001): Retributive Responses. Journal of Conflict Resolution, 45(4), 453–469.

Güth, W., Kliemt, H. and Peleg, B. (1999): Co-evolution of Preferences and Information in Simple Game of Trust. German Economic Review, 1(1), 83–110.

Güth, W. and Yaari, M., 1992, An Evolutionary Approach to Explaining Reciprocal Behavior in a Simple Strategic Game. In: U. Witt (Ed.), Explaining Process and Change – Approaches to Evolutionary Economics. The University of Michigan Press, Ann Arbor, pp. 23 ff.

Hahn, S., 1998, Reflective Equilibrium-Method or Metaphor of Justification? Schriftenreihe der Wittgensteingesellschaft. Hölder-Pichler-Tempsky, Wien, pp. 237–243.

Handel, M. (2000): Masters of war. London.

Hardin, R. (1982): Exchange Theory on Strategic Basis. Social Science Information, 2, 251 ff.

Harsanyi, J. C. (1967–8): Games with Incomplete Information Played by Bayesian Players. Management Science, 14, 159–182, 320–134, 486–502.

Harsanyi, J. C. and Selten, R. (1988): A general theory of equilibrium selection in games. Cambridge, Mass.

Hart, H. L. A. (1961): The Concept of Law. Oxford.

Hayek, F. A. v. (1972): Die Theorie komplexer Phänomene. Tübingen.

Heinimann, F. (1987/1945): Nomos und Physis. Darmstadt.

Hempel, G. and Oppenheim, P. (1948): Studies in the Logic of Explanation. Philosophy of Science, 15(2), 135–175.

Heyd, D. (1982): Supererogation. Its Status in Ethical Theory. Cambridge et al.

Hobbes, T. (1651/1968): Leviathan. Harmondsworth.

Hofstadter, D. R. (1979): Gödel, Escher, Bach: An eternal golden braid. New York.

Hofstadter, R. (1969): Social Darwinism and American Thought. New York.

Holland, J. (1975): Adaptation in Natural and Artificial Systems. Ann Arbor.

Hume, D. (1739/1978): A Treatise of Human Nature. Oxford.

Jacobsen, H. J. (1996): On the Foundations of Nash Equilibrium. Economics and Philosophy, 12(1), 67–88.

Kagel, J. H. and Roth, A. E. (Eds.), 1995. The Handbook of Experimental Economics. Princeton University Press, Princeton.

Kahneman, D. and Tversky, A. (1984): Choices, Values and Frames. American Psychologist, 39(April), 341–350.

Klein, D. B. (Ed.), 1997. Reputation. The University of Michigan Press, Ann Arbor.

Kliemt, H. (1985): Moralische Institutionen. Empiristische Theorien ihrer Evolution. Freiburg.

Kliemt, H. (1986a): Antagonistische Kooperation. Freiburg und München.

Kliemt, H. (1986b): Las institutiones morales. Barcelona/Caracas.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, R. (1982): Rational cooperation in the Finitely-Repeated Prisoners' Dilemma. Journal of Economic Theory, 27, 245–252.

Kreps, D. M. and Wilson, R. (1982): Reputation and Imperfect Information. Journal of Economic Theory, 27, 253–279.

Lahno, B. (2002): Der Begriff des Vertrauens. Paderborn.

Lewis, D. (1969): Convention. Cambridge, Mass.

Mackie, J. L. (1980): Hume's Moral Theory. London.

Mackie, J. L. (1982): Morality and the Retributive Emotions. Criminal Justice Ethics, 1982, 3–10.

McClennen, E. F. (1990): Rationality and Dynamic Choice – Foundational Explorations. New York / Port Chester / Melbourne / Sydney.

McClennen, E. F., 1998, Rationality and Rules. In: Peter A. Danielson (Ed.), Modeling Rationality, Morality and Evolution. Oxford University Press, New York and Oxford, pp. 13–40.

Mises, L. v. (1949/1966): Human Action. Chicago.

Morgenstern, O. and Schwödiauer, G. (1976): Competition and Collusion in Bilateral Markets. Zeitschrift für Nationalökonomie, 36(3-4), 217–245.

Muzzio, S. D. (1982): Watergate Games. Strategies, Choices, Outcomes. New York and London.

Nash, J. (1951): Non-Cooperative Games. Annals of Mathematics, 52(2), 286–295.

Nelson, R. R. and Winter, S. G. (1982): An Evolutionary Theory of Economic Change. Cambridge, MA.

Ockenfels, A. (2003): Reputationsmechanismen auf Internet-Marktplattformen: Theorie und Empirie. Zeitschrift für Betriebswirtschaft, 73(3), 295–315.

Ofek, H. (2001): Second Nature. Cambridge.

Olson, M. (1965): The Logic of Collective Action. Cambridge, Mass.

Ostrom, E. (1990): Governing the Commons. The Evolution of Institutions for Collective Action. New York.

Pearl, J. (2000): Causality. Models, Reasoning, and Inference. Cambridge.

Radner, R., 1998, Economic Survival. In: Donald P. Jacobs, Ehud Kalai and Morton I. Kamien (Eds.), Frontiers of Research in Economic Theory. The Nancy Schwartz Memorial Lectures, 1983–1997. Cambridge University Press, Cambridge, pp. 183–209.

Raiffa, H. (1973): Einführung in die Entscheidungstheorie. München.

Raphael, D. D. (1977): Hobbes. Morals and Politics. London.

Raphael, D.-D. (Ed.), 1969. British Moralists. Oxford University Press, Oxford.

Robertson, D. H. (1956): Economic Commentaries. London.

Rubinstein, A. (1989): The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge". American Economic Revue, 79(3), 385–391.

Schelling, T. C. (1978): Micromotives and Macrobehavior. New York and London.

Schelling, T. C. (1984): Choice and Consequence. Cambridge, MA.

Schneider, L. (Ed.), 1967. The Scottish Moralists on Human Nature and Society., Chicago und London.

Schumpeter, J. A. (1959): The Theory of Economic Development. Cambridge, MA.

Schüssler, R. (1990): Kooperation unter Egoisten. München.

Selten, R. (1978): The Chain Store Paradox. Theory and Decision, 9, 127–159.

Selten, R. (1983): Evolutionary Stability in Extensive Two-Person games. Mathematical Social Sciences, 5, 269–363.

Selten, R. (1990): Some Remarks on Bounded Rationality, vol. 172. Bonn.

Selten, R. and Stöcker, R. (1983): End Behavior in Finite Prisoner's Dilemma Supergames. Journal of Economic Behavior and Organization, 7, 47–70.

Sen, A., 1982/1976, Rational fools. In: Amartya Sen (Ed.), Choice, Welfare and Measurement. Blackwell, Oxford, pp. 84–106.

Sen, A. K., 1973/1982, Behaviour and the Concept of Preference, Choice, Welfare and Measurement. Basil Blackwell, Oxford, pp. 54–73.

Skyrms, B. (1990): The Dynamics of Rational Deliberation. Cambridge.

Skyrms, B. (1996): Evolution of the Social Contract. Cambridge.

Smith, V. L. (Ed.), 2000. Bargaining and Market Behavior. Cambridge University Press, Cambridge.

Smith, V. L. (2008): Rationality in Economics . Constructivist and Ecological Forms. New York.

Spinoza, B. d. (1670/1951): A Theologico-Political Treatise. A Political Treatise. New York.

Stöckler, M., 1991, A short history of Emergence and Reductionism. In: E. Agazzi (Ed.), The Problem of Reductionism in Science. Kluwer, Dordrecht, pp. 71–90.

Strawson, P. F. (1962): Freedom and Resentment. Proceedings of the British Academy, 187–211.

Sugden, R. (1986): The Economics of Rights, Co-operation and Welfare. Oxford, New York.

Sugden, R., 2002, Credible worlds: the status of theoretical models in economics. In: Uskali Mäki (Ed.), Fact and Fiction in Economics. Cambridge University Press, Cambridge, UK.

Sumner, W. G. (1914): The Challenge of Facts and Other Essays. New Haven et al.

Sumner, W. G. and Keller, A. G. (1927): The science of society. New Haven.

Taylor, M. (1976): Anarchy and Cooperation. London u. a.

Taylor, M. (1987): The Possibility of Cooperation. Cambridge.

Taylor, M. and Ward, H. (1982): Chickens, Whales, and Lumpy Goods: Alternative Models of Public-Goods Provisions. Political Studies, 30, 350–370.

Urmson, J. O., 1958, Saints and Heroes. In: I Melden (Ed.), Essays in Moral Philosophy. University of Washington Press, Seattle/London, pp. 198 ff.

Vanberg, V. and Buchanan, J. M. (1988): Rational Choice and Moral Order. Analyse und Kritik, 10/2, 138 ff.

Vanberg, V. J. and Congleton, R. (1992): Rationality, Morality and Exit. American Political Science Review, 86(2), 418 ff.

Vining, R. (1956): Economics in the United States of America. A Review and Interpretation of Research. Paris.

Walsh, V. and Gram, H. (1980): Classical and Neoclassical Theories of General Equilibrium. Oxford.

Witt, U. (1987): Individualistische Grundlagen der evolutorischen Ökonomik. Tübingen.

Young, H. P. (1998): Individual Strategy and Social Structure. An Evolutionary Theory of Institutions. Princeton.

Zahavi, A. (1975): Mate Selection – A Selection for Handicap. Journal of Theoretical Biology, 53, 205–214.

Zahavi, A. and Zahavi, A. (1997): The Handicap Principle. A Missing Piece of Darwin’s Puzzle. New York, Oxford.