Cordes, Christian; Schubert, Christian

Working Paper

Toward a Naturalistic Foundation of the Social Contract

Papers on Economics and Evolution, No. 0501

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Cordes, Christian; Schubert, Christian (2005) : Toward a Naturalistic Foundation of the Social Contract, Papers on Economics and Evolution, No. 0501, Max Planck Institute for Research into Economic Systems, Jena

This Version is available at:
https://hdl.handle.net/10419/20020

PAPERS on Economics & Evolution

0501

Toward a Naturalistic Foundation of the Social Contract

by

Christian Cordes
Christian Schubert

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@mpiew-jena.mpg.de

ISSN 1430-4716

Max Planck Institute for
Research into Economic Systems
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

Toward a Naturalistic Foundation of the Social Contract

Christian Cordes and Christian Schubert[1]

Max Planck Institute for Research into Economic Systems

Evolutionary Economics Group

May 2005

Abstract

This paper delivers a step toward a naturalistic foundation of the social contract. While mainstream social contract theory is based on an original position model that is defined in an aprioristic way, we endogenize its key elements, i.e., develop them out of the individuals' moral common sense. Therefore, the biological and social basis of moral intuitions are explored. In this context, a key adaptation during evolution was the one that enabled humans to understand conspecifics as intentional agents. Since these behavioral aspects are considered to be an exaptation, they are not amenable to direct genetic explanations or to rationality-based approaches.

Keywords: Social contract theory – Fairness – Intentionality – Empathy – Human evolution

JEL Codes: A13, B52, D63, D71, P16

[1] Address: Kahlaische Str. 10, 07745 Jena, Germany E-mail: cordes@mpiew-jena.mpg.de and schubert@mpiew-jena.mpg.de

1. Introduction

Evolutionary Economics still lacks a "normative branch", i.e., a theoretical concept that allows to critically discuss given and develop new criteria for the evaluation of constitutional and political rule-making. When it comes to the examination of economic or technological novelty, for example, evolutionary scholars often, at least implicitly, take it to be "good" or "desirable" a priori, without further justification or discussion. As is well known, though, already Schumpeter pointed to the ambiguous welfare implications of novelty by describing it as entailing "creative destruction".

In the present paper, we wish to examine one possible way of using the general methodology of *social contract theory* as a basic tool to develop a normative branch within Evolutionary Economics or, if you prefer, an "evolutionary welfare theory" (Witt 2003a; Vanberg 2005). We argue that the contractarian method can be made compatible with an evolutionary world-view insofar as its key concepts can be successfully *naturalized*, i.e., developed in an endogenous way out of humans' basic behavioral dispositions, underlying moral values, and behaviorally relevant social norms. Traditional contractarianism starts from an aprioristic concept of the "original position", where what it means to argue from an impartial viewpoint is defined without an empirical, psychologically informed recourse to what real-world people understand by impartiality or "fairness". What is more, due to its radical subjectivist orientation, traditional contractarianism neither specifies the determinants nor the content of the individuals' constitutional preferences, thus running the risk of normative voidness, or, worse yet, arbitrariness.

On which insights does our naturalization project draw? Human values and conceptions of justice, such as notions of fairness, originate from cultural evolution, conditioned by products of biological evolution, yet of course not deducible from the latter (see, e.g., Dobzhansky 1962, p. 345; Sugden 2001). A naturalistic approach seeks to explore the biological and social facts on which man's moral intuitions and notions of fairness are based. So, to what extent does biology influence the forming of ethics? Can there be a naturalistic theory of fairness? In the following, it is shown that communication between philosophy, anthropology, evolutionary biology, and cognitive science can deliver new answers to these questions. Empirical and conceptual progress in research on the development of human social cognition yields new insights into the core architecture of moral psychology (see, e.g., Nichols 2001; Tomasello 1999a). In this way, a more detailed picture of basic moral capacities, such as the favoring of certain norms of fairness, can be gained.

Innate elements and dispositions have a lasting influence on actual behaviors. For this reason, a consideration of the evolved cognitive dispositions of humans is helpful for a sound understanding of economic behavior (see, e.g., Vromen 2001). It will be shown that these cognitive dispositions have implications for notions of fairness and moral judgments, which in turn have far-reaching influences on human behavior in a socio-economic context. The proposed naturalistic foundation of the social contract diverges from direct genetic explanations and rationality-based approaches. Moreover, this paper's approach is juxtaposed to Binmore's naturalizing strategy of the social contract that is, at the moment, the most prominent and elaborated one.

The paper is organized as follows. In Section 2 we briefly describe the methodology of contractarianism, as it is used in mainstream Constitutional Economics, as well as the major criticisms that have been aired against it from an arguably "evolutionary" perspective. Section 3 discusses the most prominent recent attempt to naturalize key concepts of the contractarian methodology, namely Ken Binmore's game-theoretical approach. Section 4 presents our strategy to naturalize the contractarian key concepts. This strategy draws on evidence from developmental psychology, evolutionary biology, anthropology, and cognitive science that shows how a unique human capability of social cognition – to understand people as intentional beings – emerges during ontogeny. It is shown that uniquely human ways of

behavior and thought, such as notions of fairness and empathy, are based on or considerably influenced by this cognitive disposition. Moreover, the evolutionary history of this capability is depicted. Section 5 sketches a possible application of our approach and concludes.

2. The Social Contract Methodology

Within normative economics, the contractarian methodology is not only widely accepted as a tool to develop well-founded evaluative statements about economic states and processes; it has, in particular, come to dominate Constitutional Economics (Vanberg 1999). What is more, it has also received renewed interest lately, as scholars with a background in Evolutionary Economics in general and Evolutionary Game Theory in particular have started to examine the positive assumptions and behavioral hypotheses on which the contractarian approach is (at least implicitly) based. This section will briefly introduce the mainstream approach of social contract theory as well as the major objections that have been aired against it from an evolutionary perspective.

In its most simple form, the contractarian argument proceeds as follows. A just society is characterized as being the product of a "fair" agreement among rational individuals on the "rules of the market game". For that purpose, first, an original position model is construed that reflects and operationalizes what is generally held to be an "impartial spectator" viewpoint. This model is characterized by, inter alia, restrictions on the information that individuals have at their disposal when gathering in the original position in order to decide on constitutional issues. This concerns chiefly information on which future social positions and interests will be individually held in the (sub-constitutional) market game. While information on future interests is, thus, lacking, theoretical knowledge on the working properties of alternative constitutional rules is generally assumed to be given.[2]

In other words, *fairness*, defined as a constitutional rule's ability to command general voluntary assent, is argued to be guaranteed mainly by the manipulation of the information that is available to the individuals in the original position. The underlying assumption is that "blindness" (to be understood as the exclusion of that subset of information which is considered to be "morally irrelevant") artificially created by a "veil of uncertainty" (Buchanan and Tullock 1965) or "veil of ignorance" (Rawls 1971) forces even rational and self-interested individuals to activate their "moral preferences", i.e., to adopt a "moral" viewpoint and to consider the equally weighted interests of all other individuals affected, thereby facilitating "fair" agreement on non-discriminatory constitutional rules. Note that the "veil" is generally held to be a normative assumption: in order to arrive at "fair" conclusions, the individuals *ought to* decide from behind a veil (of uncertainty or, for that matter, ignorance).

Thus, a constitutional rule is regarded as "fair" insofar as it passes this hypothetical test or "thought experiment", i.e., insofar as it can plausibly be reconstructed as being part of a set of mutual behavioral constraints that benefit all individuals concerned. It is an adequate decision *procedure* that guarantees the normative quality of the social contract. The rules agreed upon constitute the market order, i.e., they define both the scope of exchange opportunities that are available to the market participants on the sub-constitutional stage as well as the meaning of "efficiency". Allocations are efficient relative to the given framework of constitutional rules. The latter are efficient insofar as they are able to command general assent.

This standard model has been criticized from two main angles. On the one hand, the objections concern the *material hypotheses* put forward by constitutional theorists about, first, the characteristics of the individuals' constitutional preferences and, second, the assumptions about the situational (i.e., above all, informational) restrictions under which they decide in the original position. To illustrate the first point, if no information whatsoever is available to the

[2] On this unsatisfying assumption see Buchanan and Vanberg (1989), Buchanan and Vanberg (1991).

constitutional theorist about the content, structure, and relative weight[3] of the individuals' preferences, then it is hard to see how any well-founded conclusions can be drawn regarding these individuals' decisions in the original position. The question is if we can formulate any material hypotheses at all, given the widespread normative postulate that the structure of the individuals' constitutional preferences should *not* be materially specified ex ante (Buchanan 1975; van Aaken and Hegmann 2002). To illustrate the second point, there is the danger of implicitly biasing the contractarian argument by arbitrarily assuming that agents in the original position lack any information whatsoever on their future (sub-constitutional) interests, or that in the original position, there are only relatively "general" constitutional rules on the agenda, whose impact on individual interests is hard to predict in detail.

On the other hand, the *formal methodology* of traditional constitutional economics has been criticized for being based upon an aprioristic "armchair" way of specifying the characteristics of the original position, i.e., in particular the situational restrictions under which the individuals operate there (Binmore 1994; Sugden 2001). Thus, the formal objections concern the way the material assumptions are introduced, i.e., the *origin* of the information required by the material critics. In this context, the contractarians' neglect of informal social norms and their order-constituting function has been criticized by various authors, partly inspired by David Hume's (1748/1992) classic rejection of the social contract metaphor (Binmore 2001; Voigt 1999a). Let us discuss these two main objections in turn.

To illustrate the *material* criticism, a classic argument on what rational individuals would agree upon from behind a "veil of uncertainty" has been developed by Harsanyi (1953; 1955). According to him, perfect uncertainty about one's own position in the sub-constitutional market game (combined, to be sure, with theoretical knowledge about the working properties of alternative constitutional rules) leads rational individuals to choose that set of rules which maximizes the average utility level.[4] In particular, individuals would not choose the *maximin* criterion, as had been famously proposed by the political philosopher John Rawls (1971).[5] Rule-utilitarianism is thereby legitimized by contractarian means. Put differently, the contractarian methodology itself adds nothing of substance to the classic normative approach of utilitarianism.

Harsanyi's argument obviously hinges on the assumption that all individuals decide on rational grounds, using the Laplace rule[6], and that they care exclusively about their own self-interest, i.e., they are indifferent both with respect to distributional patterns as well as to the way specific allocative outcomes are brought about. These are strong substantial assumptions about the content of individual preference orderings. Moreover, the rule-utilitarian conclusion, implying, as it does, the global aggregation of individual utility levels, is clearly at odds with liberal and contractarian intuitions about the non-instrumental value of individual welfare and the meaning of "Normative Individualism" (Buchanan 1991).

Another problem concerning the material assumptions refers to a key assumption of Constitutional Economics, viz., that "as the veil's "thickness" increases so will the prospect of achieving agreement" (Vanberg and Buchanan 1989: 54). Müller (1998) shows that if the constitutional rules serve to avoid sub-constitutional Prisoners' Dilemma conflicts, then this assumption depends on the relative sub-constitutional *preference intensities* of the individuals involved. Thus, some cardinal information on the preferences would seem to be needed, before any material conclusions could be drawn about the constitutional agreement.

[3] Note that the latter may differ inter-individually, due to, for example, the heterogeneous distribution of power in the original position, as in, for instance, the approach by Buchanan (1975).
[4] See Vickrey (1945: 328ff) for a related argument
[5] On this, see Harsanyi (1975: 595): "If anybody really acted this way he would soon end up in a mental institution".
[6] According to the Laplace rule, when no information whatsoever is available on how the probabilities for alternative consequences are distributed, a rational decision-maker should act on the assumption that all possible consequences have the same probability.

Finally, if the veil's "thickness" is not directly justified on normative grounds (as, e.g., in Harsanyi's contributions), but is rather taken to be a *positive* assumption (as, e.g., in Buchanan and Tullock 1965), i.e., if it is assumed that in empirical "constitutional moments", only highly general rules tend to be de facto decided upon whose specific implications are unknown to the individuals concerned, then this is, first, counter-factual. For empirically, even the individuals negatively affected by it sometimes *do* agree upon non-general, i.e., discriminatory rules. Related to this, real-world constitutional rule-making is necessarily not only about highly abstract rules; on the contrary, most of it concerns the gradual modification and practical application of abstract rules to concrete, often novel problems of social interaction. This is a key function of the judiciary (Voigt 1999b). Second, to positively assume that only general rules will be decided upon from behind the veil of uncertainty or ignorance also implies the danger of circularity (Müller 2002): if, in the context of rule design, "fair" means "non-discriminatory", then what has to be guaranteed by procedural means (the rules' non-discriminatory character) is already within the procedural assumptions (only general rules are on the agenda anyway). Thus, as a *positive* assumption, a "thick veil" does not seem to make much sense. Rather, it is reasonable to explicitly treat the "veil" as a *normative* model.

Hence, it is hard to derive anything of substance about what it is that the individuals will agree upon without introducing some positive assumptions about their constitutional preferences in the original position. If assumptions are made, then this should of course be done in an explicit and refutable way, for the normative conclusions reached at the end of the contractarian argument critically hinge on these positive assumptions. As van Aaken and Hegmann (2002) show, this postulation (that can, of course, be directly derived from Max Weber's postulate to strictly separate positive from normative science) has repeatedly been neglected even by Buchanan himself.

Thus, we are left with a dilemma: if we are not allowed to make *any* positive assumptions about the preferences of the individuals in the original position, then the contractarian conclusions are indeterminate. The social contract model is only of limited use for policy advice purposes. If the contractarian theorist's task is, in Vanberg's (2004: 155) words, to "identify constitutional reforms that are in the relevant constituency's common constitutional interests"[7], then only reforms can be proposed that are generally acceptable independently of the content of individual preferences. This, however, only applies to a quite narrow set of conceivable constitutional rules. The principle of "Normative Individualism", as it is commonly understood, bans any restrictions of the content of conceivable preferences. Put differently, the principle restricts the amount of economic information that can be used to design the original position model in such a strict way that the contractarian argument cannot be applied any longer.[8] After reviewing some of the key objections against the material assumptions of contractarianism it has become clear that in order to finally arrive at some substantial normative conclusions, the nature of the original position has to be specified somehow. Restrictions have to be introduced regarding the individuals' preferences and their informational endowment (the "veil"). To be sure, this implies the danger of ideological abuse, since almost any desired normative result can be reached by applying a suitably specified original position model.

To conclude, some, ideally well-justified, material assumptions on the structure of the original position model are clearly needed. We then face the question of how to provide for the underlying information without running the risk of biasing the whole argument by

[7] Italics omitted. Cf. also (Ibid., italics omitted): "Contractarian constitutionalism...is about telling people what, in light of our theoretical constitutional knowledge, is prudent for them to do, in terms of their own interests and purposes".

[8] Or maybe we should understand the principle differently – as just banning a certain *kind* of restrictions, namely "external" or "arbitrary" or "artificial" ones? This question, though, is beyond the scope of the present paper.

"arbitrarily" smuggling in concealed value judgments. Next, we discuss the "formal" problem of where to look for the information needed.

Turning, then, to the *formal* criticisms of contractarianism, there is one strand of thought that may point toward a way out of the deadlock just described. David Hume's classic criticism of the contractarian methodology – outlined in Hume (1748/1992) – can be used constructively to provide a systematic place for informal *social norms* and individual *value judgments* within the contractarian model.[9] This implies a refocusing of the contractarian thought experiment. It is then no longer asked what rational *homines oeconomici* would choose under "fair" conditions, with "fair" being defined in an aprioristic way. Rather, it is asked which rules would plausibly be accepted as legitimate by those individuals whose individual perceptions of their constitutional interests and consequently whose constitutional preferences are influenced by informal social norms and value judgments that are prevalent in the society under review. In a nutshell, one can say that constitutions are reconceptualized as conventions.[10] This change of perspective has far-reaching consequences for the scope, force, and scientific status of contractarian conclusions.

According to Hume, the key assumption of the first "modern" contractarian, Hobbes, viz, that a centralized, omnipotent Leviathan is necessary to establish social order is fundamentally flawed. For anarchy can be (and has been historically) overcome by decentralized means, namely informal *conventions*. First of all, after Hume, these spontaneously evolving and self-enforcing institutions are historically prior to any formally designed rules. This is, however, no strong objection against the social contract metaphor, for modern contractarians do not use Hobbes' story as an empirical hypothesis.[11] According to Hume, though, informal conventions are also prior in a normative sense: For there is no reason to assume that a *fictitious* contract – or a real one, agreed upon centuries ago, for that matter – has any binding force per se on contemporarily living *real* individuals.[12] Even my own agreement to a contract, given yesterday, does not *in itself* bind me today. Assuming otherwise means to commit a logical fallacy, viz. a logically invalid inductive conclusion: if it is rational to make a certain decision in situation A (say, the original position), it does not logically follow that it is also rational to make the same decision in situation B (say, a real-world setting).[13] Put differently, if it is rational for agent i to agree upon the terms of a social contract, it does not logically follow that it would also have been rational for agent j to do so. Rather, in order to get the necessary binding force, some prior social norm or convention has to be introduced, as for instance the social norm to abide by contractual agreements ("pacta sunt servanda"). Hume rhetorically asks the reader: "[W]hy are we bound to keep our word? Nor can you give an answer, but what would, immediately, without any circuit, have accounted for our obligation to allegiance" (Hume 1748/1992: 456).[14] This normatively expected social norm is logically prior to any formal, "artificially" set up social contract.

---

[9] Schlicht (1998) is a good introduction on key aspects of Hume's work. See also Hayek (1991). Binmore (1998) explicitly relies on Hume's anti-contractarianism when developing his own social philosophy. See also Binmore (2001) and section three below.

[10] Hardin (1990) is an early proponent of this theoretical strategy; see also Voigt (1999a).

[11] Hume's arguments against an empirical interpretation of the social contract metaphor does, then, miss its target (with the exception of Locke (1689/1988: §§ 100f) and, at times, Buchanan): in Hume (1748/1992) he observes that "[a]lmost all the governments, which exist at present, or of which these remains any record in story, have been founded originally, either on usurpation or conquest, or both, without any pretence of a fair consent, or voluntary subjection of the people."

[12] Cf. Hume (1748/1992).

[13] Cf. Müller (2002); see Sugden (1998) for an inquiry into Hume's theory of inductive inferences.

[14] In the words of Binmore (1994: 37, italics in the original): "The [traditional contractarian, C.C./C.S.] argument (based on a quasi-legal interpretation of the social contract) takes for granted that, because one would have *wished* to have made a commitment and perhaps therefore have uttered appropriate words or signed a piece of paper, *therefore* a commitment would have been made. But without a mechanism for making commitments stick, such gestures would be empty. For a person to have *claimed*, whether hypothetically or actually, that he is

Thus, for Hume, the contract metaphor itself becomes obsolete: either you abide by the contractual terms because of some internalized social norm; or it is in your pure rational self-interest to do it anyway. Hume's social philosophy disposes of the social contract altogether – it gets eliminated by Occam's razor. Concluding from this exercise in applied logic, we can say that in the long-run, constitutional rules are only viable if they are enforced by either (i) some underlying corresponding and effectively binding social norm or (ii) by the pure rational self-interest of the individuals concerned. Anyhow, some *motivating* force must be specified.

While, as we have seen, the reference to unspecified "pure rational self-interest" does not lead us very far in developing well-founded implications for constitutional reform, the focus on social norms is arguably more promising. Hume's point of view that, first, social norms are (not only historically, but also) normatively prior to consciously designed rules and that, second, individual compliance with a constitutional rule positively depends on the latter's compatibility with social norms, has interesting implications for contractarian theory. First of all, however, the concept itself has to be defined.[15] Social norms (a synonym, in this paper, to informal institutions) are behavioral regularities that are (i) common knowledge in a certain society and that are (ii) at least normatively, but often also positively expected to be followed. To be sure, normative expectation implies that there is an informal sanctioning mechanism (instead of a specialized enforcement organization, as in the case of consciously designed legal rules), i.e., at least a sub-group of the members of society is willing to incur positive private costs in order to punish individuals who defect.[16] Well-researched examples of social norms include rules of outcome-oriented (distributive) or procedural justice (Konow 2003; Benz et al. 2004).

In the context of the emerging theoretical debate about how to refocus the contractarian methodology in order to cope with Hume's critique stress has been laid on informal self-enforcing *conventions*, a subset of social norms. Conventions solve coordination problems, i.e., they are to be seen as those equilibria that get actually selected in coordination games (with or without distributional conflict) which exhibit multiple equilibria. As Schelling (1960: 56-58, 68f) has argued, the ensuing equilibrium selection problem cannot be solved by purely rational reasoning alone.[17] Homines oeconomici, operating in an institution-free environment, may often fail to coordinate on one pair of strategies and instead find themselves trapped in an infinite reasoning regress. Real-world individuals, by contrast, often are quite successful in spontaneously solving this problem, because they are intuitively able to predict their opponent's strategy choice. This ability relates to the existence of "salient" or "prominent" equilibria, i.e., on behavioral regularities that are somehow collectively expected to be followed as a kind of common practice in a given social context. Some behavioral patterns seem to be "anchored" psychologically, while others are not, although those others could just as well be rationalized as solutions to the underlying coordination game when played by rational homines oeconomici. Thus, according to Schelling and Sugden (among others), aprioristic approaches to model human behavior are not sufficient to explain the emergence of

committed to a course of action is not the same as that person *being* committed to the course of action." See also Harsanyi (1987: 343f, our italics): "[P]eople cannot rationally feel committed to keep any contract unless they have *already accepted* a moral code requiring them to keep contracts. Therefore, morality cannot depend on a social contract because...contracts obtain all their binding force from a *prior* commitment to morality". See also Müller (2002: 479). Relatedly, Voigt (1999a: 287) argues that "[t]he existence of conventions is prerequisite for the ability to establish constitutions".

[15] Cf. Witt (1989).

[16] The reason for this can be put as follows: "[A] proper coordination equilibrium [in the sense of a normatively expected behavior, C.C./C.S.] is a combination of strategies, one for each player, such that for every player i, if all the other players choose their equilibrium strategies, it is strictly best (i.e. payoff-maximizing) *for every player* that i chooses his equilibrium strategy too" (Sugden 1998: 3, italics in the original), cf. also Lewis (1969: 8-24).

[17] See also Sugden (1995) and the classic contribution by Lewis (1969: 24-36)

conventions. What is rather needed is an *empirical*, psychologically informed theory about how real-world individuals learn about collectively shared conventions, how they form subjective beliefs, and how they project their past experiences of coordination games into the future (Sugden 1998).

Individual *value judgments* are another key concept in the literature that is relevant here. Value judgments can be regarded as individually held convictions about the "fairness" of alternative behavioral strategies of other players as well as about alternative macro results (such as, e.g., distributional patterns) of the interplay of these strategies. Moreover, value judgments influence the perceived legitimacy of constitutional decision processes and constitutional rules (Tyler 1990).[18] These value judgments are themselves to be seen as conventions, hence as products of cultural learning processes. As such, they both influence the individual decision to adopt (other) social norms and are in turn influenced by the frequency distribution of (other) social norms in a social group. This interdependence is however not crucial for the purpose of the present paper. What is essential is that value judgments and social norms more general are a key determinant of individual (market and voting) behavior in that they influence what the individual perceives to be in her interest. What she perceives to be in her interest, in turn, shapes her individual *preferences*. Preferences express a clear better-worse-relationship between two alternatives, such as, for example, two alternative constitutional rules in the case of the subset of preferences that is important here, viz. *constitutional* preferences. Finally, as Buchanan and Vanberg (1989) have elaborated, constitutional preferences consist of a normative or *interest* part and a positive-instrumental or *theory* part. While the latter reflects individually held beliefs about the technical working properties of constitutional rules, the former reflects individually held value judgments and individually adopted social norms (i.e., "*given our theoretical beliefs*, in this strategic situation, this is the kind of behavior we accept in our society" or "this kind of constitutional rule we find illegitimate").

After having briefly defined the key concepts, what has all this to do with Hume's criticism of traditional ("Hobbesian") contractarian thought? Hume just provides a bridge between economic (mostly evolutionary) thinking about the emergence and maintenance of social norms on the one hand, and the contractarian methodology on the other hand. In other words, his work calls for a systematic place for social norms and value judgments within the theoretical architecture of contractarianism. This concerns the attempt to specify, at least in a basic way, (i) the individuals' (constitutional) preferences, as well as (ii) the original position (as reflecting a fundamental fairness norm), i.e., to develop a model of individuals' fairness norms that is plausible in the light of what we know about human nature. In the contractarian framework, as we have seen, the combined assumptions on the content of individual (constitutional) preferences and on the characteristics of "fair" situational restrictions serve as antecedence conditions that allow to deductively derive material hypothetical statements about what constitutional rules the individuals will finally agree upon.

3. On Binmore's naturalization approach

As we have seen in the preceding section, the traditional contractarian methodology suffers from two shortcomings: first, in order to being able to develop meaningful conclusions, it needs to specify, at least to a certain degree, the content of individual preferences and of the situational restrictions they face in the original position. In this respect,

[18] In the words of Harsanyi (1955: 315), moral values may be thought to influence an individual's "ethical preferences" (as opposed to his "subjective" preferences) that express "what he prefers only in those possibly rare moments when he forces a special impartial and impersonal attitude upon himself". See also Harsanyi (1982: 44-48) on "moral value judgments". Vanberg's (2004: 166, EN 5) concept of "constitutional interests", as opposed to "action interests" is related to this.

it has to find a middle way between, on the one hand, a radical subjectivist stance that leaves everything open, and, on the other hand, the "smuggling in" of unfounded (i.e., "artificial") or, even worse, only implicit assumptions about, for example, the individual preferences.

Second, traditional contractarianism suffers from a logical problem (of fallacious inductive inferences), insofar as it is based on the idea that today's individual-citizen i can be thought of as being normatively *bound* by a behavioral restriction that an (real or hypothesized) individual-citizen j agreed upon yesterday. This Humean insight has two implications: first, the scientific status of contractarian conclusions has to be downgraded somewhat; they cannot be regarded as rock-solid products of deductive reasoning, but rather as more or less well-founded hypothetical statements about the ability of alternative constitutional rules to command (ideally) general assent. In this sense, social contract theory can be seen, somewhat condescendingly, as a mere "means of psychological suasion" (Müller 2002: 466) or as a useful analytical device that is badly needed in order to ascertain how existing constitutional arrangements "might potentially be improved to serve [the involved individuals', C.C./C.S.] common interests better" (Vanberg 2004: 155) and, thus, to guide mutually beneficial constitutional reform.

The second implication concerns the way social contract theory performs this (latter) function. If individual citizens, when casting their vote on constitutional reform issues, are effectively motivated (predominantly) by informal social norms and value judgments, then "[t]he fact that constitutions have to be based on spontaneously arisen internal institutions...contains a search-instruction if one wants to modify or extent a constitution" (Voigt 1999a: 295). Methodologically, this means that the contractarian theorist should "define the constitutional decision making situation according to the value judgments which belong to his addressees' set of deeply rooted moral convictions" (Müller 2002: 474). This opens the way toward a material specification of a given society's social contract. If the contractarian argument and its conclusions are systematically aligned with the given informal institutional setting of a society, then both the compliance with the constitutional rules and the general stability of the social contract over time can be expected to be increased. Note, though, that this aspect points to a second "downgrading" of the status of at least some of the contractarian conclusions: while a first subset of the naturalistically founded conclusions can claim universal validity because of it being based on empirical insights into universal mechanisms of human cognition, a second subset, being based on insights into the culturally and historically contingent content of a society's network of informal social norms, obviously cannot claim any universal validity. Rather, these latter conclusions are to be seen as hypothetical statements that refer to a specific socio-cultural context only. In this sense, then, they reflect a moral relativist stance. At first sight, this may be bad news for contractarian theorists. Note, though, that if it is indeed possible to provide a systematic theoretical place for social norms within the contractarian methodology, many long-standing theoretical and logical objections against the traditional approach can be effectively countered. Hence, by reorienting the contractarian methodology towards the informal institutional background of a given society, social contract theory can both gain material specificity (when it comes to identifying the individuals "common interests") and increase the probability of voluntary compliance with its conclusions. But how exactly can this "reorientation" be envisaged?

Starting from the task of opening the contractarian "black box" of individual preferences, eliminating unfounded and developing well-founded positive assumptions about them, Binmore (1994; 1998) proposes to elaborate upon a "naturalistic" version of social contract theory. That means that insights from the sciences into the origin, determinants, and change of individual preferences shall be explicitly integrated into the contractarian theory framework. Binmore attempts to achieve this with the help of game theory's toolbox. He conceptualizes constitutional reform as the conscious choice of one "desirable" among many possible equilibria of real-world coordination games.

Binmore's approach is ultimately based on the most simple model of a *coordination game*, where two or more agents try to adjust their individual behavior (strategy choice) in such a way as to realize a joint surplus. They do so within an endlessly repeated "game of life" which denotes that sphere of human interactions that is subject to the invariant laws of physics and psychology. The key problem that the agents face, however, is the multiplicity of possible equilibria (institutional arrangements) with different distributional characteristics. This amounts to the well-known problem of *equilibrium selection*. Hence, the agents have an incentive to reach a self-enforcing equilibrium, where no one will defect ex post, because everyone gets a higher payoff than in alternative equilibria.

What is essential now is that Binmore *defines* the set of all realizable self-enforcing institutional equilibria as a society's "social contract". For him, a social contract is an "implicit self-policing agreement between members of society to coordinate on a particular equilibrium in the game of life" (Binmore 1994: 35). Thus, the social contract metaphor loses its traditional quasi-legal connotations. Consequently, Hume's criticism does not concern Binmore: his agents are solely motivated by their rational self-interest. Note, however, that this does not make the following conclusions (concerning the content of the social contract) indeterminate, since in the original position of the "game of life", Binmore's agents interact with each other and, thus, face the equilibrium-selection problem. They solve this problem spontaneously by applying the assumedly inborn, universal human capacity of *empathy*: by putting themselves imaginatively in their opponents' shoes, the agents are able to figure out which strategy their opponents will presumably choose and which convention will eventually result.[19] Thereby, mutually harmful distributional conflicts are avoided. Hence, at this point, institutional arrangements, social norms, and morality emerge spontaneously.

A first essential aspect of this approach deserves to be emphasized: explicitly, Binmore does not follow Kant and the Kantians (such as, in his view, Rawls) in postulating some moral norms and ethical convictions (that the individuals have when placed in the original position) in an aprioristic "armchair" way. Rather, he aims at *endogenizing* morality as a spontaneously emerging institution.

Besides the "game of life", a parallel "game of morals" is the second pillar of Binmore's theoretical edifice, and it concerns the normative value judgments of the individuals. In the course of the "game of morals", a "just" equilibrium is chosen out of the many technically realizable equilibria. Thus, delicate normative issues arise at this point. As already alluded to above, according to Binmore, we all, as human beings, are genetically endowed with a special capacity, viz. empathetic preferences, that helps us to put ourselves in the shoes of our fellow men. This capacity helps the agents to figure out how a quick solution to coordination and bargaining problems could look like that would avoid wasteful conflicts. Binmore regards empathetical preferences as "Nature's answer" to the universal human problem of equilibrium selection. As it turns out, Binmore hypothesizes not only that "Nature's answer" happens to satisfy the von Neumann-Morgenstern axioms, but that in the end, the classic Nash bargaining solution will prevail. This bargaining solution, then, is seen as a plausible model of universal human normative conceptions of fairness. This is so for two reasons: first, Binmore

[19] As regards the relationship between empathy and sympathy, this paper's notion differs from the one of modern rational choice theory: as will be shown in detail in the next section, the unique human capability to take another's perspective and to understand his intentions forms the basis for empathy. According to rational choice theory, empathy enables an agent to identify with fellow human beings and to potentially figure out the content of their preferences. This does not necessarily imply the consideration of another's well-being (see Sugden 2002). However, what is more, this capability essentially also provides the cognitive foundations of sympathy: the ability to see the self in others allows for an understanding of others as sentient beings like the self, i.e., similar affective states are aroused on the part of the observer. Getting in touch with the emotions of other persons contributes to the motivating of human behavior. Therefore, in the next section, a more encompassing concept of empathy will be applied that also includes sympathetic aspects and that does not belong to the ontology of rational choice theory.

conceptualizes the mutual "figuring out" as, indeed, a *bargaining* process (albeit one where no one speaks a word; the bargaining is purely imaginative). Second, he assumes the Nash bargaining solution to be the product of long-term processes of cultural evolution, since it arguably turned out to maximize the genetic fitness of the social groups that employed it in those ancestral environments where the basic behavioral dispositions of man were shaped.[20]

What is interesting for the purposes of the present paper is Binmore's concept of the original position. He observes that what individuals regard as "just" relates closely to the use of their empathetic preferences: "[T]he interest in the original position lies in the fact that it represents a stylized version of do-as-you-would-be-done-by principles that are *already* firmly entrenched [sic] as joint decision-making criteria within the system of commonly understood conventions that bind society together" (Binmore 1998: 112, italics in the original).[21] Thus, Binmore at least embarks on the task of explaining and, potentially, further specifying the key role that the original position with its essential characteristic, the veil of uncertainty or ignorance, plays as a model for Kant's widely accepted "golden rule" (i.e., an ethical rule of universality) instead of simply defining it in an aprioristic way. Thus, he tries to anchor the contractarian methodology within the informal institutional framework of society.

According to Binmore, existing constitutional rules are "just" if they can be reconstructed as emanating from the kind of "imaginative bargaining" sketched above. The scientific observer can test this by applying a general consensus rule. During the game of life, every agent unsatisfied with his actual expected life-time utility has, at any point in time, the right to lower the veil again and to re-open the game of morals. Note that this normative rule, too, is part of the model that, according to Binmore, reconstructs universal normative human concepts of fairness. Hence, a constitutional rule is "just", insofar as it can be reconstructed as being a *simultaneous* equilibrium both in the game of life and in the game of morals.[22] What is eventually qualified as "just" must necessarily be qualified as institutionally realizable (within the specific cultural-institutional context of a given society) beforehand. If a coordination game has to be "artificially" established by a hierarchical body, as in the case of PD games, institutional solutions are "just" that do not depart "too much" from those fairness norms that would have guided individual behavior in the case of a lucky decentralized solution (Binmore 2001).

To briefly summarize the general structure of a naturalization approach, it comprises three key elements: first, some statements about the individuals' constitutional preferences; second, a model (however basic) of a plausible impartiality norm (this model will include some informational restrictions); third and most importantly, a set of statements that explain or justify both, i.e., the statements about individual preferences and those about the impartiality norm. From an evolutionary viewpoint, this explanation or justification will include an account of how the basic behavioral elements that underlie both preferences and the impartiality norm (or norms) have come about in the course of cultural evolution which, in turn, is arguably based on some basic genetically anchored capacities. The last point will now be examined in more depth in the following.

4. The evolutionary origins of man's ideas of fairness

As has been expounded above, this paper's inquiry takes a naturalistic starting-point; human conceptions of justice and fairness are understood as the product of biological and social processes of evolution and learning (see Sugden 2001). In this section, insights from

[20] Cf. Skyrms (1996: ch. 2) on some qualifications to this heroic assumption.
[21] See also Binmore (1994: 336); Binmore (1998: 178, 209).
[22] "An equilibrium of the natural game G [the "game of life", C.C./C.S.] is said to be fair if its play would never give a player reason to appeal to the device of the original position under the rules of the morality game M" (Binmore 1998: 11).

evolutionary biology, anthropology, and cognitive science provide an explanatory basis for man's moral intuitions, notions of fairness, and empathetic potential. What is more, these findings can serve as a starting-point for the development of a naturalistic foundation of the social contract: first, some material hypotheses about the individuals' constitutional preferences can be formulated. As has been argued before, this is a prerequisite to say something about the social contract's content. Second, humans' empathetic potential is the basis on which the notion of an "impartial spectator" rests, who is asked to take other people's perspective in an unbiased way. In this manner, characteristics of "fair" situational restrictions in the original position can be derived. Moreover, crucial differences to the naturalizing strategy brought forward by Binmore are outlined.

During human evolution, man adapted for culture in ways other primates did not.[23] Biologically evolved novel forms of social cognition and cultural learning formed the basis for cultural evolution and exclusively human ways of behavior. In this context, the key adaptation is the one that enables humans to understand other individuals as intentional agents like the self – a capability necessary for reproducing another's behavioral strategies and for taking other persons' perspectives (Carpenter et al. 1998; Tomasello 1999b; Tomasello 1999a).[24] This unique cognitive skill of man underlies behavioral patterns such as joint attentional activities, discourse skills, the learning to use tools, the creation and use of conventional symbols, the participation in and creation of complex social organizations and explains, as will be shown in this section, some aspects of the motivational underpinnings of human inclinations, such as a tendency toward cooperation and fairness.

In the case of human children, the ontogeny of these species-specific cognitive skills begins at the end of the first year of life. Carpenter et al. (1998) have shown that, at around one year of age, infants display qualitatively new behaviors that indicate a newly emerging understanding of other persons as intentional beings. Children start to recognize that the attention and behavior of others to outside objects may be shared and directed in various ways. While six-month-old infants interact dyadically with objects and other people, children at approximately 9-12 months of age begin to engage in interactions that are triadic, i.e., they involve the referential triangle of child, adult, and an outside entity to which they share attention. In the course of human ontogeny, these triadic social skills emerge together as a group. Baron-Cohen (1995) has suggested that there exists – in the case of humans – a neurocognitive mechanism that is concerned with establishing a shared focus of attention with another organism, i.e., with the construction of triadic representations (see also Premack 1990). Such phenomena of triadic social skills and interactions are characterized by the term "joint attention" (Dunham and Moore 1995).[25] A prerequisite for this engagement in joint attentional interactions is the understanding of other persons as intentional agents (Premack 1990; Tomasello 1995).[26] Meltzoff (1995) has shown that eighteen months old children

[23] The following discussion draws on Cordes (2004).

[24] In this context, intentional agents are defined as animate beings capable of controlling their spontaneous behavior, having goals, making active choices among behavioral means of attaining those goals, and choosing what they pay attention to in pursuing their goals.

[25] It is known from research into autism that without a fully developed ability for joint attention, human beings fall into a grievous state of pathology. Moreover, autism is considered to be a disorder of biological origin (Baron-Cohen 1990). Without joint attention, humans cannot construct and coordinate the shared social realities that constitute everyday life. The hypothesis is that in autism a specific impairment in the development of a "theory of mind" prevents sufferers from understanding and predicting much of behavior due to the fact that they can not refer to mental states, such as intentions, emotions, believes, etc. (Bruner 1995; Kasari and Sigman 1995).

[26] These cognitive achievements are followed by the acquisition of linguistic communication skills (Carpenter et al. 1998: 116). By learning and using linguistic symbols in a productive manner, children are demonstrating their understanding that other persons have points of view about a situation that may differ from their own. The capacity of language to represent objects and states of affairs in the world is an extension of the more

observing an adult who tried, but failed, to perform certain target acts can infer the adult's intended act by watching these failed attempts. They go beyond just duplicating what was actually done and instead enact what the adult intended to do by taking his perspective. The infants situate people within a psychological framework that distinguishes between the surface behavior of agents and a deeper behavioral level consisting of goals, perspectives, and intentions. What is more, only persons are understood within this psychological framework but not inanimate objects. Human behavior is seen as purposive. Furthermore, it is assumed that there is an innate tendency to attribute intentions to humans.[27]

The cognitive capability of infants to understand other persons as intentional agents like the self is the ontogenetic precursor to various social skills, such as the understanding of the thoughts and beliefs of others that emerges later in life and set the stage for more sophisticated skills of perspective taking, cultural learning, communication, and a "theory of mind" (Carpenter et al. 1998: 118ff; Goldman 1992; Premack and Woodruff 1978). The functional properties of joint attentional experiences are implicated in a broader array of developmental phenomena comprising the infant's understanding of the mental life of others: first, the most basic level in social cognition involves an understanding that others attend to and have intentions toward outside entities; infants begin to take part in "joint visual attention". On a second level, the specific content of the psychological stances of others is understood, i.e., the specific behavioral intentions to change a certain environmental stage and the specific perceptual intention to attend to some things are identified. After having attained this level of understanding, children begin to follow the specific focus of attention and behavioral intentions of others and start to attempt to manipulate these specific psychological states. Infants begin to act on another's attitude by, for example, re-directing the other person's gaze. Three year olds are able to anticipate the reactions of people whose current mental stance differs from their own. Four year olds understand that someone might hold a false belief and realize that the content of the belief may depict reality in a misleading or outdated fashion.

Consequently, the sophisticated human skills of social cognition do not just mimic the surface structure of an observed behavior, they also mean a reproduction of an instrumental act understood intentionally. Humans do not just reproduce the behavioral means but also the intended end to which the behavioral means was applied. By participating in social and communicative interactions with other persons understood intentionally, human children come to mentally represent the world in some unique ways and to appropriate the accumulated wisdom of their social group as embodied in material artifacts, symbolic artifacts, collective cognition, and conventional social practice. This unique capability of intersubjective and perspectival cognitive representation provides the basis for cultural evolution. Humans can only understand a cultural symbolic convention if they experience their communicative partner as an intentional agent with whom one can share attention and whose perspective they can internalize.

The intentional behaviors of human children in combination with an understanding of the intentional behaviors of others constitute a recursive process. Infants' ability to identify with fellow human beings – to perceive others as "like me" – contributes to infants' experience of their own intentionality. This process provides the basis for the infant's earliest self-concept and for simulating the perspective of others toward the self. Soon after infants begin to behave in a clearly intentional way, they begin seeing the behavior of others as intentional as well and that others experience the world in ways similar to the ways in which they themselves do (Carpenter et al. 1998: 124ff). One hypothesis is that children do this most readily with

biologically fundamental capacities of the cognitive apparatus to relate the organism to the world by way of mental states such as beliefs, desires, and intentions (Bates 1990; Searle 1983: vii).

[27] Nonhuman primates do not show these kinds of social-cognitive skills. Their forms of social learning and cognition do not require a comprehension of others as intentional agents with whom they can align themselves.

respect to shared emotions, i.e., that they are able to give subjective meaning to the emotional expressions they see in others (see Meltzoff 1990). They infuse the expressions seen in others with meaning from personal experience, i.e., infants perceive regularities between their own expressions and emotional states. Finally, these capabilities enable children to develop a "theory of mind" that includes the other as a sentient being. A growing understanding of others is applied to oneself and, conversely, the knowledge of oneself is applied to others. Thus, knowledge of self and others develops simultaneously. To come to understand conspecifics in terms of their intentions, beside some cognitive predispositions, the learner must him- or herself be treated as an intentional agent encouraged by other organisms to attend to or behave toward some object of mutual interest. The species-unique aspects of human cognition are socially constituted, i.e., human social organization is an integral part of the process that results in the special characteristics of human cognition. Social mirroring provides an important foundation for the development of self, for elaborating the similarity between self and others, and for understanding that fellow human beings, like the self, are sentient beings with thoughts, intentions, and emotions. A fundamental facet of human nature is the ability to include others in the definition of self and to see the self in others (Brown et al. 2002). Moreover, man's capacity to attribute self-awareness to persons is central to an understanding of the psycho-dynamics of an individual and the human social order.

Approaches in cognitive sciences to the social cognition theme reveal differences in opinion about the timing of an infant's development of an understanding of others. However, all accounts assume the social understanding of attentional and intentional states of others to be the ontogenetic foundation on which a more complex "theory of mind" can be built, including an understanding of goals, emotions, desires, references, and beliefs (see, e.g., Astington and Gopnik, 1991; Baron-Cohen 1995; Wellman 1991).[28] What is more, distinguishing the surface behavior of people from another deeper level involving intentions, emotions, and goals lies at the core of humans' commonsense psychology, communication, notions of fairness, and moral judgments (see, e.g., Meltzoff 1995).[29]

A fundamental anthropological question, whose answer is important for any attempt of a naturalistic foundation of the social contract, is where the complex behavioral practices and their cognitive basis came from. Recent research on human phylogeny has provided evidence showing that the species-specific aspects of human cognition mentioned above must have arisen rapidly when considered in an evolutionary context (for references see Tomasello 1999a: 2ff). Given these insights, there has not been enough time for natural selection forces to have – one after the other – created each of the cognitive skills necessary for modern humans' cultural and social achievements. Hence, there must be a small difference that made a big difference, i.e., an adaptation that changed the process of primate cognitive evolution in fundamental ways (see Carpenter et al. 1998; Tomasello 1999b). Culture and its cognitive foundations are considered to make this big difference. Human beings do have some species-unique modes of cultural transmission and social cognition. The evolved unique forms of

[28] Given the general developmental synchrony with which these cognitive capabilities emerge in ontogeny, it seems highly implausible that these behaviors are conditioned one at a time, each under an own set of reinforcement contingencies.

[29] Kahneman and Tversky (1982: 203) have run an experiment in which the subjects were given the following example: "Mr. Crane and Mr. Tees were scheduled to leave the airport on different flights, at the same time. They traveled from town in the same limousine, were caught in a traffic jam, and arrived at the airport 30 minutes after the scheduled departure time of their flights. Mr. Crane is told that his flight left on time. Mr. Tees is told that his flight was delayed, and just left five minutes ago. Who is more upset?" In this experiment, 96 percent of the subjects said that Mr. Tees would be more upset. An explanation of these experimental results can be found in the human capability to imaginatively project themselves into the shoes of the two protagonists. The subjects decided that they themselves would be more upset in Tee's situation than in Crane's. To do so, they "simulated" Tee's and Crane's mental states and generated an affective state that is ascribed to the corresponding agent. Similarly, predictions of behavior would be derived (see also Goldman 1992).

social cognition – with the key adaptation that enabled individuals to understand other persons as intentional agents like the self – are biological adaptations for culture. In the further course of human evolution, such a capability for culture greatly enhanced the reproductive fitness of an organism.[30] However, as will be shown below, these features of human cognition, which were previously shaped by natural selection for a particular function, can be coopted for new, possibly non-adaptive, uses.

In contrast to these views, sociobiologists, who seek to explain the social behavior of organisms, typically assume that all facets of behavior are the equilibrium result of natural selection. According to their approach, individuals behave so as to maximize some (utility) function by their procurement of various resources. This utility function should have been closely related to the number and quality of an organism's offspring. Moreover, social behavior among unrelated individuals results from the interaction of selfish rational individuals. Hence, no tendency toward cooperation or fairness can be observed unless these behaviors are in accordance with the interests of these actors.[31] Furthermore, sociobiologists argue that since the capacity for culture arose by the process of natural selection, the resulting ways of behaving must be adaptive.[32]

For example, Binmore (1998: ch. 4) and other authors (see, e.g., Rubin 1982) claim that intuitive ethical notions are already entrenched among the innate instincts that regulate human life. According to their arguments, humans have evolved certain perceptions of ethics that are explicable in terms of the environments in which human phylogeny occurred. Binmore believes that the appeal of Rawls' original position lies in the fact that humans recognize it as a stylized version of a principle that is unconsciously applied in interaction with others. He argues that the cognitive apparatus that enables an adoption of a Rawlsian original position evolved during human phylogeny as a means to come to agreements in prehistoric food-sharing between members of the same family.[33] Hence, according to Binmore (2001), fairness is an evolved device to select an equilibrium in real-life games. The argument put forward in this paper differs from Binmore's in that an evolved cognitive disposition is proposed that is a general adaptation for culture. As will be shown later, this disposition – as a side effect – gave rise to certain notions of fairness. The postulated evolutionary history for the device of the original position differs from Binmore's.

Moreover, Binmore has, in order to apply the formal structure of game theory, to naturalize rational choice and expected utility theory by showing that biological or social evolution selects decision-making behavior that satisfies the corresponding axioms (see Sugden 2001). In addition, since in his original position the contracting parties rely on empathetic preferences, Binmore has to naturalize such preferences that at the same time satisfy the axioms of expected utility theory and the von Neumann-Morgenstern axioms respectively. Binmore's empathetic preferences exist only due to the fact that they can be applied in solving equilibrium selection problems. However, as is shown in this paper, for naturalizing these preferences, there is no need to invoke equilibrium selection or the original

[30] Considering the origins of the human capacity for culture, a period of co-evolution of cultural and natural evolution can be discerned. This phase of a mutually interactive relationship ultimately allowed forms of human behavior to emerge that had a strong relative reproductive success and resulted in an ending of natural selection as a shaping force. Behavioral variety of man increased notwithstanding adaptive value in terms of genetic fitness (Boyd and Richerson 1980; Witt 2003).

[31] In game theory, cooperative attitudes are regularly interpreted as a matter of tastes or preferences shaped during human phylogeny (see, e.g., Güth and Yaari 1992; Binmore 1998). Cooperation is then taken as an expression of human genes.

[32] Sociobiology, applied to such organisms as, for example, social insects, can explain the evolution of behaviors. However, sociobiological arguments applied to explain all facets of social behavior in higher mammals face severe problems (see, also for references, Maryanski 1994).

[33] Moreover, Binmore (1998: 413ff) assumes that, with respect to transactions with strangers, the hardwired fairness algorithm has been modified by cultural evolution to be also applied to these cases. He maintains that humans learned to adopt strangers into the family clan by treating them as relatives.

position. What is more, there is no need that empathetic preferences satisfy the von Neumann-Morgenstern axioms induced by a process of selection. This paper's naturalistic approach is informed by evidence from other disciplines, such as cognitive science, that helps to understand the origins of man's empathetic capacities and his notions of fairness.

An important question in this context is whether or not the whole variety of social behaviors is adaptive, i.e., the product of selection for the trait in question. This point is central to the argument of Gould and Vrba (1982), who emphasize the crucial distinction between historical genesis and current utility of an organism's characteristics, for example, the faculty to understand other persons as intentional agents like the self. A once adaptive trait can be converted for other functions than its original utility. As a result, the direct historical relation between behavior and adaptation may be skewed. To tackle this problem, Gould and Vrba introduced a new term to the taxonomy of evolutionary morphology. They contributed the concept of "exaptation" that accounts for the evolution of biological features that actually enhance an organism's fitness but were not built by natural selection for their current role.[34] In contrast, the term "adaptation", as suggested by Darwin (1859), refers to features built by selection for their current role, i.e., the origin and perfection of such a design can be attributed to a long period of selection for effectiveness in a particular role.[35]

Gould and Vrba strictly discriminate between the historical genesis of biological dispositions and their current utility, in order to avoid viewing natural selection as so dominant among evolutionary mechanisms that historical process and current product necessarily become one. Characteristics previously shaped by natural selection forces for a particular function (an adaptation), can be coopted for a new use (an exaptation).[36] Furthermore, many evolved features of organisms are non-adapted but available for advantageous cooptation in descendants. Consequently, there are two sources of exaptation: features may have been adaptations for another function, or they may have been non-adaptive structures, for example, structures correlated with features contributing to fitness.[37] Many cases in biological evolution can be considered to be exaptations: selection for the initial development of feathers in an ancestor could have been for the function of thermoregulation and not for flight – a fundamental innovation that had far-reaching, incidental consequences. Bones could have evolved initially as an adaptation for storing phosphates and for mineralization, both needed for metabolic activity. Only later in evolution did bone replace the cartilaginous endoskeleton and adopt the function of support. What is more, some contributions to evolutionary biology invoke, at least implicitly, the phenomenon of exaptation to assess the capabilities of the human mind (see, e.g., Williams 1996: 14, Wilkins and Dumford 1990).

Gould and Vrba (1982) themselves state that the human brain is undoubtedly built by natural selection for some complex set of functions but can, as a result of its intricate structure, work in an unlimited number of ways that are quite unrelated to the selective forces that constructed it. Although many of these capabilities, such as the taking of another's perspective, are important for survival in later social contexts, a current application or utility does not automatically carry implications about historical origins. Most aspects of cultural evolution, whose natural foundation is man's evolved cognitive apparatus, that enhanced mankind's survival can be ascribed to the dominance of exaptation (see also Gould 1991; Lewontin 1990). Hence, behavior can be quite different from that predicted by sociobiology

[34] For further references for this strand of thinking see Chipman (2001).
[35] Though, also the concept of "exaptation" has its origins in the writings of Darwin.
[36] Other authors use the term "pre-adaptation" for the same phenomenon (see, e.g., Corning 2003).
[37] Adaptations that had been converted to an exaptation of different effect set the basis for a subsequent adaptation. Any coopted structure does not necessarily arise perfected for its new effect. Complex biological features can evolve by a mixture of exaptations and adaptations. As a result, this process leads to evolutionary transformations that probably could not have arisen by adaptation alone. Exaptations originate randomly with respect to their subsequent effects.

or many economists (see Boyd and Richerson 1980; Durham 1976).[38] These thoughts about man's evolutionary origins have far-reaching implications for evolutionary thinking about human behavior, for example, in the context of evolutionary social theory.

Although, as has been shown above, a significant genetic event happened in human cognitive evolution that opened the way for some new and powerful social and cultural processes, this event does not specify the detailed outcomes of behavior we see today. It just provided the basis for cultural evolution that, in the following and with no further genetic events, entailed many of the most distinctive characteristics of human cognition (see Tomasello 1999b). It is the phenomenon of evolved traits, which emerged because they solved a particular evolutionary problem, but that then gave rise to completely different capabilities of an organism. A genetic event changed the nature of primate social cognition, which revolutionized the social-cultural transmission process and led to a series of cascading sociological and psychological changes. The process of cultural evolution is based on a new form of social cognition that involves the understanding of other persons as intentional agents like the self. This perspectively based cognitive representation enabled several forms of cultural learning, sophisticated ways of socio-cultural transmission, and empathy-based behaviors.[39] The following hypothesis is proposed:

> The evolved unique human cognitive faculty to understand other persons as intentional agents like the self, i.e., the capability to take another's perspective, which is man's basic biological adaptation for culture, has far-reaching implications with respect to moral judgments, notions of fairness, empathy, and human behavior in a socio-economic context in general. This interrelationship can explain the motivational underpinnings of a variety of human inclinations and behaviors, such as a tendency toward cooperation and fairness. Since these aspects of human behavior can be considered to be an exaptation, they are not amenable to a direct genetic explanation.

The understanding of other persons as intentional beings is the ontogenetic foundation on which "theories of mind" are based, comprising an understanding of thoughts, goals, and emotions. Due to this evolved unique cognitive faculty, humans are able to give subjective meaning to the emotional expressions and experiences of others, who are understood as sentient beings like the self. Mental states of others can be simulated and affective states, which are ascribed to the corresponding agent, can be generated, thereby contributing to the motivating of human behavior.

This becomes obvious when considering further evidence from developmental psychology: human infants respond to signs of emotions of fear, disgust, and distress in others early in life. They begin to show prosocial responses to the distress of others by 12 months of age and show a full range of prosocial behaviors by 20 to 24 months (for references see Kasari and Sigman 1995).[40] Prosocial behaviors and empathic concern develop over the second year of life. Buchsbaum and Emde (1990) assume that self-awareness and early morality occur in an integrated manner and are mediated by social referencing. A prerequisite for both dispositions is a consistent, emotionally available sense of self, other, and we. The authors expect empathy to have major roots in biology and to enhance the understanding of self by virtue of emotional monitoring of another's experience, which may differ from one's own. Mental perspectival representations of others are connected to one's own emotional

[38] The breaking up of direct genetic mechanisms prevents *a priori* mathematical models from being applied as stylized descriptions of social evolution (see, e.g., Sugden 2001).

[39] Psychologists use the label "perspective taking" for the phenomenon of empathy.

[40] Matsumoto et al. (1986) have shown that young children are capable of a wide range of morally sensitive behaviors in Prisoner's Dilemma situations. In addition, evidence from experiments with children (for details see Buchsbaum and Emde 1990) hints at a basic reciprocity built into fittedness for human social interaction. This operates from earliest infancy and may be the starting point of a child's sense of fairness about reciprocity between humans. Reciprocal behavior is a component of fairness-driven behavior (see Falk et al. 2003).

centers in a manner similar to self-representations (see, e.g., Gierer 1998). Due to this universal "affective core", humans are able to get in touch with the feelings of others. Thus, empathy arises from imaginatively adopting the perspective of another. Such "pretend" states are then operated upon by psychological processes that generate feelings, attitudes, or affects that facilitate an empathizing with the target individual's states.[41] This affective dimension explains why social relations have subjective value for human beings.

Within modern rational choice theory, there is a categorical distinction between sympathy and empathy (see Sugden 2002): sympathizing means an integration of another's welfare as an argument in one's own utility function. As a consequence, the agent's feelings are not only affected by the perception of another's feelings, but she is also always motivated to perform actions that benefit this person. Empathy, on the other hand, denotes the ability to identify with another person to discover her preferences and beliefs without caring for her welfare. In contrast to these views, this paper's concept of empathy is more encompassing for it includes aspects of sympathy as understood by rational choice theory. In the context of our argument, empathy is based on the evolved human faculty to understand other persons as intentional agents like the self and on how – based on this capability – one person's affective state can influence another's thus motivating behavior without involving the axioms of rational choice theory (for that notion of the role of empathy see also Damasio, 2003: 270). Therefore, empathy understood this way also comprises – as a sympathetic aspect – the capability to get in touch with the feelings of other persons arousing similar affective states in the observer, i.e., pleasurable feelings if the other person's state is pleasurable or painful if it is painful (see Hume, 1740/1978, who has a similar argument). As Adam Smith (1759/1982) has argued, this capacity of empathy is tightly linked to the approval or disapproval of other people's sentiments by a process of aligning them with one's own sentiments and the thereby induced norms of "propriety of sentiment" – or notions of fairness – within groups of interacting persons. David Hume (1740/1978: 468) has said that moral facts are facts of psychology giving rise to certain kinds of sentiments of approval and disapproval. The human capacity of empathy is the basis for these sentiments and establishes an important demand on an "impartial spectator", who is explicitly asked to employ her empathetic potential to come to agreeable rules.

Given these insights into the human psyche, social scientists can be enabled to make sense of the calculus underlying the coordinative effort on an outcome, for example, in the context of a social contract, that people afterwards describe as "fair".[42] It can be explained why such a consensus is firmly in favor of some empathy-based type of, as Binmore (1998: 8) puts it, "do-as-you-would-be-done-by" principle. The capability to experience empathetically another's affective states, negative or positive ones, motivates people to follow this principle as a basis of their notions of fairness. Thus, a substantial naturalistic perspective on notions of fairness and moral judgments delivers sound arguments to show why it is morally imperative for humans to follow such a "golden rule" without assuming an inevitable correspondence

[41] The motivational underpinnings of, for example, human helping behavior, namely a merged identity with the victim, general negative affect, and true altruism, i.e., empathetic concern, are all based on the human cognitive capacity to take the perspective of others (see, e.g., Batson 1991; Brown et al. 2002; Davis et al. 1996). Empathizing with one's conspecifics promotes mutual aid and inhibits injurious behavior (see Goldman 1992). Though, empathy tends to be biased. Subjects are more empathic to persons who are familiar and similar to themselves than to persons who are different.

[42] The subjects of a study conducted by Ames and Marwell (1981), for instance, have shown a surprising unanimity of thought regarding what was considered a fair contribution to a public good. Experiments on fairness in social psychology have led to an empirically based behavioral rule that resolves problems of social exchange by equalizing the ratio of each agent's allotment to his worth. People who are considered to be worthy get more than others. Normally, this theory is called "modern equity theory" (see, e.g., Mellers 1982). A prerequisite for such behavior is the capability to imagine oneself in the shoes of others, i.e., to see things from their point of view. Then it may be acceptable that those who invested more effort in a common affair are rewarded by receiving a correspondingly higher share of the benefits.

between decision-theoretic concepts and actual mental processes that are based on evolved cognitive dispositions. The unique human capability to understand other people as intentional agents like the self and the accompanying faculty to take the perspective of conspecifics is presented as the cognitive foundation of the above-mentioned "golden rule". The latter is also specifying a characteristic of the original position.

There is a lot of evidence showing that many basic results of economic theory conflict with the intuitive ethical or fairness notions of people (see, e.g., Selten, 1998, Fehr and Schmidt 1999, Bolton and Ockenfels 2000, Falk et al. 2003; Rabin 1993). Agents often show behaviors that are inconsistent with some of the most fundamental principles of economics. In many circumstances, humans display levels of cooperative behavior that contradict the predictions of economic models of rational self-interested individuals. Laboratory experiments have delivered robust empirical results showing that subjects offer significant amounts of money to other players in dictator games, contribute to the funding of public goods even if this action is not in their immediate self-interest, do less free riding than economic theory predicts, and achieve non-Nash levels of cooperation in repeated dilemma games (see, e.g., Ames and Marwell 1981; Davis and Holt 1993, Kagel and Roth 1995, Ferraro et al. 2003; Güth and van Damme 1998). Many authors offer "other-regarding" preferences as an explanation of these behaviors, however, without regarding the exact nature or the origins of these preferences: for example, according to Selten (1978), one motivation of people to cooperate in games is the endeavor of the agents not to be "mean" in the sense that they disappoint the other player's trust.[43] This argument only holds true if the disappointment of the other player is somehow experienced or comprehended in place of this agent. The vicarious experiencing of others disadvantage translates into an active concern for their welfare. Obviously, individuals derive satisfaction from contributing to the welfare of others in order to achieve a "fair" outcome (see also Bolton and Ockenfels 2000). Thus, the common assumption in economic theory that more complex patterns of behavior result from the interaction of selfish, unrelated individuals may be untenable. Adding fairness and empathetic potential to economic models substantially alters conclusions. Moreover, it can be argued that "fairness" is a component of people's constitutional preferences. Therefore, the approach put forward in this paper can provide an explanation for some of these unclear motivational underpinnings of moral judgment or fairness and provides an empirical basis to refine theories of economic behavior in general and social contract theory in particular.

5. Conclusions

In order to prepare the ground for using the contractarian methodology as a basic toolbox for a normative branch within Evolutionary Economics, we have proposed a way in which its key concepts might be naturalized in the sense of being developed out of insights into the psychological foundations of (i) value judgments concerning fairness, and (ii) the positive and normative role of empathetic preferences.

The unease with the traditional aprioristic approach to model the original position starts from Hume's logical objections against contractarian thinking. In order to avoid a logical fallacy by deriving strong normative commitments from the merely hypothesized agreement of individuals on a social contract, it has been proposed to try to reorient the social contract model toward value judgments and social norms as forces that effectively motivate human beings. The integration can concern the material and/or the formal aspect of the original position model. While the former concerns material hypotheses on the origin and content of individual constitutional preferences, the latter concerns the question how people's normative use of empathy can be (i) explained and (ii) modeled by means of an original position model.

[43] Adam Smith (1759/1982) has argued for a "natural desire to please" as a fundamental characteristic of man.

Relative to existing attempts to naturalize the social contract model (such as the one outlined in Binmore's "Game Theory and the Social Contract"), the argument that has been brought forward in this paper is more substantial and less speculative with respect to the evolutionary mechanisms that are underlying man's notions of fairness. On the other hand, positive *predictions* about the substance of constitutional preferences as well as about what "impartiality" means to real-world individuals turn out to be much more difficult than was supposed by Binmore.

Phenomena such as the feeling of compassion or empathy toward other persons, the existence of certain notions of fairness and moral judgments as well as many other aspects of human behavior relevant in the context of constitutional reform are not adequately explained by the neo-Darwinian or the rationalist paradigm. This paper has therefore suggested a perspective from which some moral judgments or notions of fairness can be interpreted in naturalistic terms: based on insights from cognitive science, developmental psychology, anthropology, and evolutionary biology, it has been shown that some "objective moral values", such as the consideration of another's welfare due to empathy phenomena that are based on evolved cognitive dispositions, put pressure on the human will, emotions, and interests. The empathic human potential for certain feeling states rests on the unique human cognitive faculty to understand other persons as intentional agents like the self and may act as a natural, non-relativistic platform upon which various universal, or principled, moral standards can be built. Since these aspects of human behavior can be considered to be an exaptation, they are not amenable to a direct genetic explanation.

As regards the policy implications of our approach, two aspects are worth stressing. First, and more generally, constitutional reform is particular likely to be accepted and judged legitimate by the individuals affected if it is grounded on and aligned with humans' conceptions of fairness. Second, besides this probable gain in legitimacy, there is, in many policy areas, a genuine pragmatic need for the application of non-arbitrary and, better yet, well-founded fairness norms. To illustrate, consider the standard legal policy case of "Law & Economics", as first outlined in Coase (1960): having produced a negative external effect due to their mutually incompatible land uses, two parties bargain with each other over how to re-allocate the relevant property rights (such as, e.g., the "right to breathe clean air" or the "right to emit smoke"). Coase assumes that the prospect of realizing a positive joint surplus suffices to induce a cooperative solution in that the disputed property right ends up with the parties that values it the most. However, this solution presupposes a solution to an underlying distributional (hence, zero-sum) game: in order to reach a bargaining solution, the parties have to agree upon how to distribute the positive joint surplus. As soon as asymmetric information is introduced into the model, there is, though, no guarantee that such an agreement will be reached. Gains from exchange may therefore be left unexploited.

The problem can be solved, however, when the parties dispose of a *distributional norm* that specifies ex ante how the joint surplus should be partitioned. This norm may be introduced "externally", for instance by the state, or it may be have been emerged "internally", i.e., in the course of multiple former attempts to solve the bargaining problem. In the latter case, as has been empirically demonstrated by Ellickson (1991), the norm is part of the cultural context of the parties involved. Hence, real-world individuals may spontaneously solve bargaining problems that would not have been easily solved by hyper-rational *homines oeconomici*. While Ellickson (1991) studied rural communities in Northern California, where the existence of firmly rooted distributional norms may plausibly be assumed, in densely populated urban areas (where, of course, most tricky land use conflicts surface) this no longer holds. Hence, there is the need to introduce "external" norms – which arguably should be acceptable to the individuals involved. The case of urban land use conflicts may, then, be one important area of application of our naturalized social contract approach. To be sure, further research is required in this regard.

References

van Aaken, A. and Hegmann, H. (2002). Konsens als Grundnorm? *Archiv für Rechts- und Sozialphilosophie* 88: 28-50.

Ames, R.E. and Marwell, G. (1981). Economists free ride, does anyone else? *Journal of Public Economics* 15: 295-310.

Astington, J.W. and Gopnik, A. (1991). Developing understanding of desire and intention. In A. Whiten (Ed.), *Natural theories of mind*, 39-50. Oxford: Basil Blackwell.

Baron-Cohen, Simon (1990): "Autism: A specific cognitive disorder of 'mind-blindness'", *International Review of Psychiatry*, Vol. 2, pp. 81-90.

Baron-Cohen, S. (1995). The eye direction detector (EED) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In P.J. Dunham and C. Moore (Eds.), *Joint attention: Its origins and role in development*, 41-59. Hillsdale: Lawrence Erlbaum.

Bates, E. (1990). Language about me and you: Pronominal reference and the emerging concept of self. In M. Beeghly and D. Cicchetti (Eds.), *The self in transition*, 165-182. Chicago: University of Chicago Press.

Batson, C.D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale: Lawrence Erlbaum.

Benz, M., Frey, B.S., Stutzer, A. (2004). Introducing procedural utility: Not only what but also how matters. *Journal of Institutional and Theoretical Economics* 160: 377-401.

Binmore, K. (1994). *Game theory and the social contract I: Playing fair*. Cambridge MA: MIT Press.

Binmore, K. (1998). *Just playing: Game theory and the social contract II.* Cambridge MA: MIT Press.

Binmore, K. (2001). Natural justice and political stability. *Journal of Institutional and Theoretical Economics* 157: 133-151.

Bolton, G.E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review* 90: 166-193.

Boyd, R. and Richerson, P.J. (1980). Sociobiology, culture and economic theory. *Journal of Economic Behavior & Organization* 1: 97-121.

Brown, S., Cialdini, R.B., Luce, C.L., Maner, J.K., Neuberg, S.L. and Sagarin, B.J. (2002). The effects of perspective taking on motivations for helping: Still no evidence for altruism. *Personality and Social Psychology Bulletin* 28: 1601-1610.

Bruner, J. (1995). From joint attention to the meeting of minds: An introduction. In P.J. Dunham and C. Moore (Eds.), *Joint attention: Its origins and role in development*, 1-14. Hillsdale: Lawrence Erlbaum.

Buchanan, J.M. (1954). Social choice, democracy and free markets. *Journal of Political Economy* 62: 114-123.

Buchanan, J.M. (1975). *The limits of liberty.* Chicago: University of Chicago Press.

Buchanan, J.M. (1991). The foundations for normative individualism. In J.M. Buchanan (Ed.), *The economics and the ethics of constitutional order*, 221-229. Ann Arbor: University of Michigan Press.

Buchanan, J.M. and Tullock, G. (1965). *Calculus of consent*. Ann Arbor: University of Michigan Press.

Buchanan, J.M. and Vanberg, V.J. (1989). Interests and theories in constitutional choice. *Journal of Theoretical Politics* 1: 49-62.

Buchanan, J.M. and Vanberg, V.J. (1991). Constitutional choice, rational ignorance and the limits of reason. *Jahrbuch für Neue Politische Ökonomie* 10: 61-78.

Buchsbaum, H.K. and Emde, R.N. (1990). 'Didn't you hear my mommy?' Autonomy with connectedness in moral self emergence. In M. Beeghly and D. Cicchetti (Eds.), *The self in transition*, 35-60. Chicago: University of Chicago Press.

Carpenter, M., Nagell, K. and Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development* 63.

Chipman, A.D. (2001). Developmental exaptation and evolutionary change. *Evolution & Development* 3: 299-301.

Coase, R.H. (1960). The problem of social cost. *Journal of Law and Economics* 3: 1-44.

Cordes, C. (2004). The human adaptation for culture and its behavioral implications. *Journal of Bioeconomics* 6: 143-163.

Corning, P. (2003). *Nature's magic: Synergy in evolution and the fate of humankind.* Cambridge: Cambridge University Press.

Damasio, A.R. (2003). *Looking for Spinoza.* London: William Heinemann.

Darwin, C. (1859). *On the origin of the species by means of natural selection.* London: J. Murray.

Davis, D.D. and Holt, C.A. (1993). *Experimental economics.* Princeton: Princeton University Press.

Davis, M.H., Conklin, L., Smith, A. and Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology* 70: 713-726.

Dobzhansky, T. (1962). *Mankind evolving.* New Haven: Yale University Press.

Dunham, P.J. and Moore, C. (1995). Current themes in research on joint attention. In P.J. Dunham and C. Moore (Eds.), *Joint attention: Its origins and role in development*, 15-28. Hillsdale: Lawrence Erlbaum.

Durham, W.H. (1976). The adaptive significance of cultural behavior. *Human Ecology* 4: 89-121.

Ellickson, R.C. (1991). *Order without law: How neighbors settle disputes.* Cambridge MA: Harvard University Press.

Falk, A., Fehr, E. and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry* 41: 20-26.

Fehr, E. and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114: 817-868.

Ferraro, P.J., Rondeau, D. and Poe, G.L. (2003). Detecting other-regarding behavior with virtual players. *Journal of Economic Behavior & Organization* 51: 99-109.

Gierer, A. (1998). Networks of gene regulation, neural development and the evolution of general capabilities, such as human empathy. *Zeitschrift für Naturforschung* 53(c): 716-722.

Goldman, A.I. (1992). Empathy, mind, and morals. *Proceedings and Addresses of the American Philosophical Association* 66: 17-40.

Gould, S.J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues* 47: 43-65.

Gould, S.J. and Vrba, E.S. (1982). Exaptation – a missing term in the science of form. *Paleobiology* 8: 4-15.

Güth, W. and van Damme, E. (1998). Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study. *Journal of Mathematical Psychology* 42: 227-247.

Güth, W. and Yaari, M.E. (1992). Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In U. Witt (Ed.), *Explaining process and change*, 23-34. Ann Arbor: University of Michigan Press.

Hardin, R. (1990). Contractarianism: Wistful thinking. *Constitutional Political Economy* 1: 35-52.

Harsanyi, J.C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434-435.

Harsanyi, J.C. (1955). Cardinal welfare, individual ethics and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309-321.

Harsanyi, J.C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls' theory. *American Political Science Review* 69: 594-606.

Harsanyi, J.C. (1982). Morality and the theory of rational behavior. In A.K. Sen and B. Williams (Eds.), *Utilitarianism and beyond*, 39-62. Cambridge: Cambridge University Press.

Harsanyi, J.C. (1987). Morals by agreement. Review of: David Gauthier, Morals by agreement. *Economics & Philosophy* 3: 339-351.

Hayek, F.A. (1991). The legal and political philosophy of David Hume (1711-1776). In W.W. Bartley and S. Kresge (Eds.), *The collected works of F.A. Hayek*, Vol. III., 101-118. Chicago: University of Chicago Press.

Hume, D. (1740/1978). *A treatise of human nature.* Oxford: Clarendon Press.

Hume, D. (1748/1992). Of the original contract. In T. Hill Green and T.H. Grose (Eds.), *David Hume. The philosophical works*, Vol. 3, 443-460. Aalen: Scientia.

Kagel, J.H. and Roth, A.E. (1995). *Handbook of experimental economics.* Princeton: Princeton University Press.

Kahneman, D. and Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, 201-208. Cambridge: Cambridge University Press.

Kant, I. (1793/1977). Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht in der Praxis (On the Old Saw: This May be Right in Theory, but it won't work in Practice). In W. Weischedel (Ed.), *Immanuel Kant Werkausgabe*, Vol. XI, 127-172. Frankfurt/M.: Suhrkamp.

Kasari, C. and Sigman, M. (1995). Joint attention across contexts in normal and autistic children. In P.J. Dunham and C. Moore (Eds.), *Joint attention: Its origins and role in development*, 189-203. Hillsdale: Lawrence Erlbaum.

Konow, J. (2003). A positive analysis of justice theories. *Journal of Economic Literature* 41: 1188-1239.

Lewis, D. (1969). Convention. A philosophical study. Cambridge MA: Harvard University Press.

Lewontin, R.C. (1990). The evolution of cognition. In D.N. Osherson and H. Lasnik (Eds.), *An invitation to cognitive science*, 229-246. Cambridge MA: MIT Press.

Locke, J. (1689/1988). *Two treatises on government*. 2nd ed. Cambridge: Cambridge University Press.

Maryanski, A. (1994). The pursuit of human nature in sociobiology and evolutionary sociology. *Sociological Perspectives* 37: 375-389.

Matsumoto, D., Haan, N., Yabrove, G., Theodorou, P. and Cooke Carney, C. (1986). Preschooler's moral actions and emotions in prisoner's dilemma. *Developmental Psychology* 22: 663-670.

Mellers, B.A. (1982). Equity judgment: A revision of Aristotelian views. *Journal of Experimental Psychology* 111: 242-270.
Meltzoff, A.N. (1990). Foundations for developing a concept of self: The role of imitation in relating self to other and the value of social mirroring, social modeling, and self practice in infancy. In M. Beeghly and D. Cicchetti (Eds.), *The self in transition*, 139-164. Chicago: University of Chicago Press.
Meltzoff, A.N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31: 838-850.
Müller, C. (1998). The veil of uncertainty unveiled. *Constitutional Political Economy* 9: 5-17.
Müller, C. (2002). The methodology of contractarianism in economics. *Public Choice* 113: 465-483.
Nichols, S. (2001). Mindreading and the cognitive architecture underlying altruistic motivation. *Mind & Language* 16: 425-455.
Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition* 36: 1-16.
Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a 'theory of mind'? *Behaviour and Brain Sciences* 4: 515-526.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281-1302.
Rawls, J. (1971). *A theory of justice*. Cambridge MA: Belknap Press.
Rubin, P.H. (1982). Evolved ethics and efficient ethics. *Journal of Economic Behavior & Organization* 3: 161-174.
Rubin, P.H. (2001). The state of nature and the evolution of political preferences. *American Law and Economics Review* 3: 50-81.
Schelling, T.C. (1960). *The strategy of conflict.* Cambridge MA: Harvard University Press.
Schlicht, E. (1998). *On custom in the economy.* Oxford: Clarendon Press.
Searle, J.R. (1983). *Intentionality.* Cambridge: Cambridge University Press.
Selten, R. (1978). The chain store paradox. *Theory and Decision* 9: 127-159.
Selten, R. (1998). Features of experimentally observed bounded rationality. *European Economic Review* 42: 413-436.
Skyrms, B. (1996). *Evolution of the social contract.* Cambridge: Cambridge University Press.
Smith, A. (1759/1982). The theory of moral sentiments. In A.L. Macfie and D.D. Raphael (Eds.), *The theory of moral sentiments*. Indianapolis: Liberty Fund.
Sugden, R. (1995). A theory of focal points. *The Economic Journal* 105: 533-550.
Sugden, R. (2001). Ken Binmore's evolutionary social theory. *The Economic Journal* 111: F213-F243.
Sugden, R. (2002). Beyond sympathy and empathy: Adam Smith's concept of fellow-feeling. *Economics and Philosophy* 18: 63-87.
Tomasello, M. (1995). Joint attention as social cognition. In P.J. Dunham and C. Moore (Eds.), *Joint attention: Its origins and role in development*, 103-130. Hillsdale: Lawrence Erlbaum.
Tomasello, M. (1999a). *The cultural origins of human cognition*. London: Harvard University Press.
Tomasello, M. (1999b). The human adaptation for culture. *Annual Review of Anthropology* 28: 509-529.
Tyler, T. (1990). *Why people obey the law.* New Haven: Yale University Press.
Vanberg, V.J. (1999). Markets and regulation: On the contrast between free-market liberalism and constitutional liberalism. *Constitutional Political Economy* 10: 219-243.
Vanberg, V.J. (2004). The status quo in contractarian-constitutionalist perspective. *Constitutional Political Economy* 15: 153-170.
Vanberg, V.J. (2005). Human intentionality and design in cultural evolution. In C. Schubert and G. v. Wangenheim (Eds.), *The evolution of designed institutions.* London: Routledge, forthcoming.
Vanderschraaf, P. (2000). Game theory, evolution, and justice. *Philosophy & Public Affairs* 28: 325-358.
Vickrey, W. (1945). Measuring marginal utility by reactions to risk. *Econometrica* 13: 319-333.
Voigt, S. (1999a). Breaking with the notion of social contract: Constitutions as based on spontaneously arisen institutions. *Constitutional Political Economy* 10: 283-300.
Voigt, S. (1999b). Implicit constitutional change: changing the meaning of the constitution without changing the text of the document. *European Journal of Law and Economics* 7: 197-224.
Vromen, J. (2001). The human agent in evolutionary economics. In J. Laurent and J. Nightingale (Eds.), *Darwinism and evolutionary economics*, 184-208. Cheltenham: Edward Elgar.
Wellman, H.M. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind*, 19-38. Oxford: Basil Blackwell.
Wilkins, W. and Dumford, J. (1990). In defense of exaptation. *Behavioral and Brain Sciences* 13: 763-764.
Williams, G.C. (1996). *Adaptation and natural selection*. Princeton: Princeton University Press.
Witt, U. (1989). The evolution of economic institutions as a propagation process. *Public Choice* 62: 155-172.
Witt, U. (2003a). Economic policy making in evolutionary perspective. *Journal of Evolutionary Economics* 13: 77-94.
Witt, U. (2003b). *The evolving economy – essays on the evolutionary approach to economics*. Cheltenham: Edward Elgar.