

Whigs in Space

A Naturalistic Approach to Fairness

Jacco Thijssen[¶]

Prelude

Starting with Plato, Western political philosophy has evolved addressing fairness issues, one of its most recent fruits being Rawls (1971). Professor Ken Binmore gave [¶]ve very interesting and thought-provoking lectures on how he thinks fairness issues should be handled.

Realizing to be far too crude in categorizing, two main lines of thought, "schools" maybe, are known since the late seventeenth century. The [¶]rst is the utilitarian school. Fairness is viewed solely in terms of gains and losses. A state is fair if the state provides "the greatest happiness for the greatest number" to quote Jeremy Bentham, one of the founding fathers. Also John Stuart Mill and Adam Smith can be seen as utilitarianists. Binmore describes utilitarianism to be a metaphysical teleological theory, the "cosmic order" or "Good", being determined by utility.

On the other hand Binmore distinguishes so-called metaphysical non-teleological theories. Exponents of this line are Locke (arguably), Rousseau, Kant and Rawls. The cosmic order in these theories is imposed by "the Good". Starting very prudent with Rousseau the concept of "duty" came into philosophy. With Kant it reached gigantic proportions culminating in a free will that tells you exactly what to do and what not. The transcendental "I" was born. In the nineteen seventies, John Rawls tried to vindicate Kant's ideas using a less vague concept, namely the original position.

John Harsanyi takes a rather unique position between these two schools. He tries to give a Rawlsian defence of utilitarianism (cf. Harsanyi, 1977). His main di®erence with Rawls concerns what happens within the original position. Basically, Rawls rejects

[¶]Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Ph: +31-(0)13-4662824; Fax: +31-(0)13-4663280; E-mail: J. J. J. Thi j ssen@kub. nl .

Bayesian decision theory and uses the maximin criterion instead. He is then led to an egalitarian outcome. Harsanyi on the other hand remains loyal to orthodox decision theory and arrives at a utilitarian outcome. In the remainder of this report both approaches will be described in more detail.

Apart from these metaphysical theories a third theory emerged, mainly in the side margins, namely naturalism. It was instigated by Thomas Hobbes and the torch was magnificently taken over by David Hume (cf. Hume, 1978), the unquestioned hero of Binmore's lectures. Naturalism is non-metaphysical, i.e. it does not presupposes notions of a "Good" or a "Right". To put it stronger, it rejects such ideas.

In this report we will first provide some intuitions for the naturalistic approach, based on results from psychology, anthropology, and game theory. Then, Harsanyi's and Rawls' stories will be retold using the rigor of game theory. It then becomes clear where both theories go astray. Finally, Binmore's own theory as described in Binmore (1994,1998) and in the lectures will be discussed. I realize that the aimed scope of this report is far broader than its contents. This fallacy is due to the many subtle though extremely important¹ arguments and thoughts that I cannot discuss because of limited space. The interested reader is referred to Binmore (1994) and Binmore (1998). In this report I assume that the reader is familiar with the theory of bargaining. Readers not familiar are referred to e.g. Osborne and Rubinstein (1994).

In Section 1 I will describe some ideas that form the basis of the naturalistic approach. In Sections 2 and 3 the naturalistic version of Harsanyi's and Rawls' theories will be discussed, respectively. In Section 4 some considerations are given on why we should reject utilitarianism. In Section 5 Binmore's own theory will be explained and finally, in Section 6 some conclusions are drawn.

1 Humean Fairness

One of the most important features of a naturalistic approach is the idea that, in David Hume's words, you can never derive an "ought" from an "is". A categorical imperative is like asking on the street: "Where should I go?". Naturalism hinges on hypothetical imperatives: "If you want to catch the 16.00 train, you ought to leave now." A hypothetical imperative inextricably links actions to goals, whereas a Kantian categorical imperative prescribes actions without making any reference to goals.

A naturalistic approach to morality then begins by first realizing that morality evolved

¹In political philosophy in particular, the seemingly trivial bears the greatest conceptual difficulty.

	C	D
C	(2,2)	(0,3)
D	(3,0)	(1,1)

Table 1: Prisoner's Dilemma

along with the human race as a system of self-policing conventions that promote cooperation. Especially the self-policing part is extremely important. If morals were not self-policing, there should be some external policeman to enforce the rules. The policeman then acts like a kind of philosopher-king in Platonian terms. An important consequence of viewing morality as a result of evolution is that there is no authority for preachers whatsoever.

It is by now well known that fairness norms evolved to get a society to an efficient social contract when a new source of surplus appears. The basic structure of these norms is Rawls' original position. It refers to a coordination problem behind a veil of ignorance. It is the only fairness norm that really works. One of the reasons why it works is that our whole life is filled with coordination problems. In the original position there is no incentive to create a disadvantage. This leads Rawls to the conclusion that the maximin criterion is used. To use the original position successfully however, empathetic preferences have to be used. Using empathetic preferences constitutes that every issue is viewed from within the social contract. Rawls strips away everything from people behind the veil of ignorance, even their empathetic preferences and puts back maximin in its place. Harsanyi also forgets about empathetic preferences without really replacing them with something else. He tries to build a wall without bricks so to say. We will return to empathetic preferences later to give a more rigorous account.

A social contract can be viewed as a consensus to coordinate on a particular equilibrium of the game of life. They are self-policing and hence require no external enforcement. There is an important misjudgment in political philosophy concerning the social contract that can be easily solved with some trivial game theory. Most political philosophers think that life is like a prisoner's dilemma (see Table 1). However, there is only one equilibrium in the prisoner's dilemma, namely for both players to defect. This leads philosophers to think that there needs to be some external force to make sure everybody cooperates, i.e. the rules of the game need to be changed. This is the wrong approach, which is also present in economics where governments are assumed to be able to enforce rules. It would be better to use mechanism design to make sure that everybody acts optimally, because it is optimal to do so.

	C	D
C	(5,5)	(0,4)
D	(4,0)	(2,2)

Table 2: Stag-Hunt Game

A more interesting game arises from Rousseau (1996) called the stag-hunt game which in a stylized form can be represented as in Table 2. This game has two pure Nash equilibria ((C,C) and (D,D)). The question is whether the game settles in the risk-dominant equilibrium (D,D) or in the Pareto-efficient (C,C) equilibrium. Harsanyi and Selten (1988) claim that the risk-dominant outcome will prevail. The question is then how to get society to the efficient equilibrium. Naturalists want to do this without external enforcement. Hence, it must be optimal to act optimally.

Another observation refuting the claim that life is like a Prisoner's Dilemma is that human cooperation is founded on reciprocity in repeated games. Hence, the dynamical aspect can not be ignored. Reciprocity boils down to: "I'll scratch your back if you scratch mine", or in the words of the ever gloomy Hobbes: "I won't scratch your back if you won't scratch mine". Deontological intuitions about rights and duties are derived from observing the rules for sustaining equilibrium strategies in repeated games, whereas political philosophy should be aimed at getting to a new equilibrium if situations change.

In this section we provided some arguments why deontologists go wrong and why a naturalistic approach seems to be the better one. In the next two sections Harsanyi's and Rawls' story will be retold using some of the ideas mentioned in this section, but in a more rigorous way.

2 Visitors in Eden I: Harsanyi

The scene is set as follows. The Game of Life is symbolized by two players, Adam (A) and Eve (E) being in the garden of Eden negotiating the terms for a marriage contract. For now we suppose there is a philosopher-king that can enforce the resulting contract being binding. The process of getting to the marriage agreement (the social contract) is free however.

Suppose that the status quo is given by $\gg = (0; 0)$. The bargaining process is a repeated game with a set of feasible solutions X like in Figure 1.² It can be shown that necessarily the set X is convex, comprehensive, closed, and bounded from above. The asymmetries

²Since there is a philosopher-king, all states in X are indeed feasible.

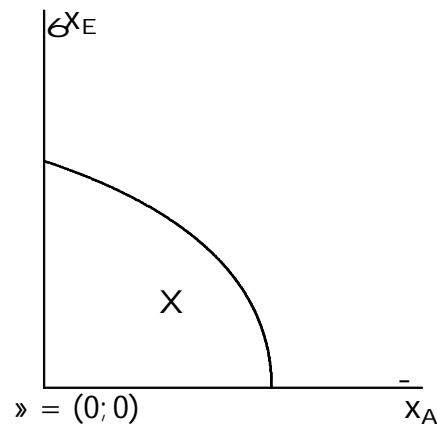


Figure 1: The Game of Life

of the set X register the ineradicable inequalities between Adam and Eve for which the original position is coming into play. In the original position Adam and Eve disappear behind a veil of ignorance, that is their names are "dropped". In the original position Adam will be player I and Eve will be player II. When the veil of ignorance is dropped there are two possible outcomes from the point of view of Adam and Eve: either player I is Adam and player II is Eve (denoted AE), or the other way around (denoted EA)³

It is here that empathy comes into play. A person has empathetic preferences if in her utility function, she includes both her own as the other player's personal utility function. Note that this is not equivalent to sympathy. A player has sympathetic preferences if the other player's utility function is a part of his own personal utility function. Empathetic preferences are implied for instance when saying: "I'd rather be Adam wearing a \bar{g} -leaf than Eve eating the apple". Let player I's preferences be given by his empathetic von Neumann-Morgenstern utility function v_1 . His beliefs are represented by a subjective probability distribution $(p_1; 1 - p_1)$, where p_1 is the probability that player I assigns to the event AE. Player I's expected utility for the contingent social contract (C; D) is then given by

$$w_1(C; D) = p_1 v_1(C; A) + (1 - p_1) v_1(D; E); \quad (1)$$

where $v_1(C; A)$ denotes the utility player I derives from social contract C if he turns out to be Adam. $v_1(D; E)$ can be interpreted similarly. Denoting the personal utilities of Adam and Eve for a social contract Y by $u_A(Y)$ and $u_E(Y)$, respectively, and by scaling correctly it can be shown that for some constants U_1 and V_1

$$v_1(Y; A) = U_1 u_A(Y) \quad (2)$$

³Of course we know that the true state is AE, but players I and II don't.

$$v_1(Y; E) = 1 - V_1(1 - u_E(Y)) \quad (3)$$

This means that Player I has an intrapersonal standard of utility comparison that equates V_1 of Adam's utils with U_1 of Eve's. Doing the same for Player II yields expected utilities in terms of the personal utility functions

$$w_1(C; D) = p_1 U_1 u_A(C) + (1 - p_1)[1 - V_1(1 - u_E(D))] \quad (4)$$

$$w_2(C; D) = p_2[1 - V_2(1 - u_E(C))] + (1 - p_2) U_2 u_A(D) \quad (5)$$

In the original position it holds that $p_1 = p_2 = \frac{1}{2}$. It should be noted that the introduction of empathetic preferences implies a fundamental difference with Harsanyi's original story. For Harsanyi and Rawls alike, people behind the veil of ignorance are stripped away with virtually everything. They only have a notion about people and some basic goods.⁴ They are somewhat like Kant's transcendental "I". Empathetic preferences are about treating unrelated as kin, with the difference that the implied degree of relationship is not determined genetically, but socially. Hence, not all sociability is stripped away from humans in the original position in Binmore's theory. By doing so, Binmore seems to take a communitarianistic approach (cf. Sandel, 1984).

Harsanyi assumes that behind the veil of ignorance the players are interested in the expected utilities only. Therefore they regard a contingent social contract that leads to the payoff pair y if AE occurs and to z if EA occurs, equivalent to the social contract $t = \frac{1}{2}y + \frac{1}{2}z$. Thus to get the correct bargaining set in the original position (denoted by T) the following transformations have to be made. First the set X must be transformed to a set X_{AE} and a set X_{EA} for the events AE and EA, respectively. That is,

$$X_{AE} = f(U_1 x_A; 1 - V_2(1 - x_E)) \cap (x_A; x_E) \subseteq X \quad (6)$$

$$X_{EA} = f(1 - V_1(1 - x_E); U_2 x_A) \cap (x_E; x_A) \subseteq X \quad (7)$$

Then, the set T is given by

$$T = ft = \frac{1}{2}y + \frac{1}{2}z \cap X_{AE}; z \cap X_{EA} \quad (8)$$

The status quo ζ is assumed to be given by $\zeta = \frac{1}{2}\zeta' + \frac{1}{2}\zeta^3$, where ζ' and ζ^3 are the mappings of ζ into X_{AE} and X_{EA} , respectively. This is depicted in Figure 2. The bargaining problem $(T; \zeta)$ can then be solved using Nash's theory of bargaining with commitment. It yields the Nash bargaining solution. Harsanyi argued that rational people in identical situations

⁴Rawls calls these goods "primal social goods".

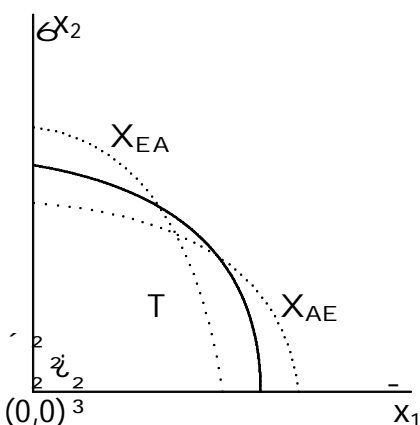


Figure 2: Harsanyi's original position

act identically. This Harsanyi doctrine boils in the original position down to assuming symmetry, i.e. $U_1 = U_2 = U$ and $V_1 = V_2 = V$. If the resulting equilibrium is translated back to the original set X it can be shown to maximize the weighted utilitarian social welfare function

$$W_h(x) = Ux_A + Vx_E: \quad (9)$$

In this way Harsanyi uses Rawls' theory to underpin utilitarianism. The advantage of the naturalistic approach is that we can directly put our finger at the weak spots, namely why should the Harsanyi doctrine hold? And how are U and V determined? We will come back to these issues in Section 5.

3 Visitors in Eden II: Rawls

Having set the scene in Section 2 makes it easy to come up with a naturalistic version of Rawls' theory. We use the empathetic preferences as described in eqs. (4) and (5). Again, we impose symmetry, to keep the analysis as close to Rawls' reasoning as possible. According to Rawls, people will apply the maximin rule in the original position. This implies that the value of a contingent social contract equals the value of the social contract in the worst case possible. That is, if player i receives y_i if the situation AE occurs and z_i if EA occurs then he values a social contract t_i the same if

$$t_i = \min\{y_i; z_i\}: \quad (10)$$

The bargaining set T is then simple to construct as the intersection of X_{AE} and X_{EA} , i.e. $T = X_{AE} \cap X_{EA}$ as depicted in Figure 3. The status quo equals $z = (0; 0)$. Since the problem is symmetric, using the symmetric Nash bargaining solution on $(T; z)$ yields as equilibrium point the intersection of the Pareto frontiers of X_{AE} and X_{EA} .

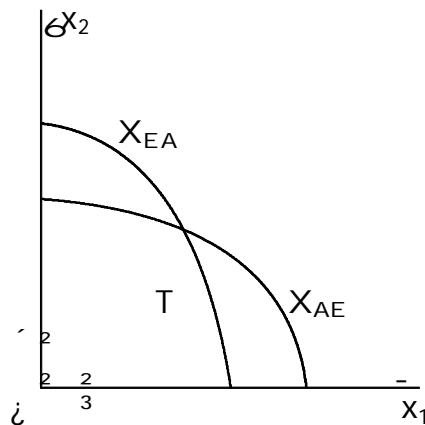


Figure 3: Rawls' original position

For both Harsanyi and Rawls, assuming symmetry ensures that the equilibria can be reached using the same social contract whatever the imaginary coin toss yields, AE or EA. However, it does not ensure that players are indifferent between the coin toss. In Harsanyi's case, they generally will not be indifferent.⁵ In Rawls' case it doesn't matter how the coin falls since in both cases the same utility is received. Rawls' theory is therefore highly egalitarian.

Let us return now to the original set of feasible social contracts X . Denote the transformed equilibrium by $(r_A; r_E)$. Then it can be shown that it must satisfy the condition

$$U r_A = 1 - V(1 - r_E) \quad (11)$$

This is nothing else but the proportional bargaining solution with weights U and V for the bargaining problem $(X; \bar{r})$, where $\bar{r} = (0; 1 - V)$ (see Figure 4). Thus, Rawls' theory too can be fashioned in a naturalistic way.

The equilibria found in the last two sections arise from empathetic preferences. They are therefore called empathetic equilibria. An empathetic equilibrium can best be described as answering no at the following question:

Suppose that you could deceive everybody into believing that your empathetic preferences are whatever you find it expedient to claim them to be. Would such an act of deceit seem worthwhile to you in the original position relative to the empathetic preferences that you actually hold?

Ken Binmore, *Game Theory and the Social Contract* vol. II, p. 224

⁵Rawls attacks the utilitarian approach exactly because of this, leading to his famous slaveholder's argument (see Rawls, 1971).

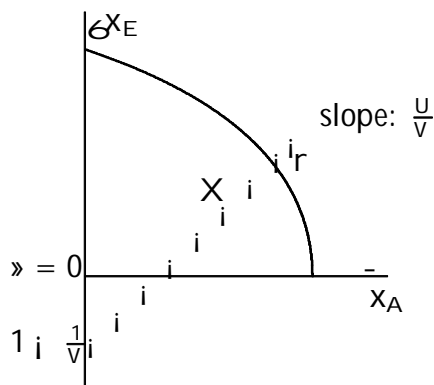


Figure 4: The social contract

It will be this equilibrium concept that is crucial to the approach of Binmore as described in Section 5.

4 A Moment Away from the Armchair

The arguments Rawls gives for using the maximin rule are vague and strong arguments are given against them in Binmore (1994). On the other hand, philosophers have spent a great deal of time devoted to criticizing the utilitarian approach. One of the famous arguments is that according to utilitarianism it is fair to kill someone if society as a whole benefits more from his death, than the deceased suffers. In this hypothetical case most people will agree that killing someone for the sake of society alone can not be considered fair.

However, this is theoretical morality.. Does it describe the everyday practice? At the risk of sounding Hobbesian, let us leave our armchair and walk the streets. Is it then fair to have elegant oriental rugs that are made by enslaved children? How often do we see beggars and other outcasts in the street not caring about their fate? The reason we do not care is that in our minds we have dehumanized them by attributing all kinds of features that might by no way describe the actuality: drug addiction, alcoholism, mental illness, etc.

Apparently, the exploitation of those powerless is accepted as an (unfortunate) consequence of the necessity that a productive society provides adequate incentives for its workers. This is utilitarian reasoning! So, much we like to picture ourselves as having strong ethical standards, we often do not live up to them.

Although the basic reasoning of utilitarianism might be correct, the outcomes are not. The reason lies in the implementation of utilitarian reasoning. For example, from a

utilitarian point of view it is first-best if all blind people get one eye from those who have eyesight. The idea however that one of our eyes might be surgically removed is so horrific that we seek a second-best solution in which we keep our eyes, no matter the consequences for the blind. The reason the first-best solution cannot be implemented and sustained is that there is no omnipotent philosopher-king. Therefore, the first-best solution can't be enforced. Thus, we reject utilitarianism not because it isn't first-best, but because it actually is.

5 The Social Contract in Eden

The theory that we will present here assumes that no external enforcement agency is possible. There will be no opposition between what is right and what is good, because these notions are the result of bargaining. Ideas of the right are then necessary to sustain equilibrium, while ideas of the good determine the selection of an equilibrium. In this context it becomes immediately clear, why life cannot be modelled as a Prisoner's Dilemma: this game only has one equilibrium. The selection component is therefore absent. In other words, the Good must already have been specified.

Rights are therefore prior to the Good, since it is always necessary to first find out what is feasible. Only then the question of optimality can be raised. The device used is the Game of Morals, which is an enriched form of the Game of Life. After each round of the repeated Game of Life, the players are allowed in the Game of Morals to appeal to the device of the original position. If an appeal is made, the players disappear behind the veil of ignorance with the empathetic preferences they have at that point in time, after which bargaining takes place. Nature then makes a chance move determining the actual social situation. A fair social contract is then defined as an equilibrium in the Game of Life, yielding strategies that if used in the original position never leaves a player with the incentive to appeal to the original position. Important is that people can cheat in the Game of Morals, but they don't have an incentive to. Hence, there is no need for external enforcement.

With respect to the set X we must interpret the points in it as payoff-ows of the Game of Life. As for the state of nature, we can no longer use the trivial status quo as before⁶. The correct status quo here would reflect Adam and Eve's payoff while bargaining, i.e. the state-of-nature point » is given by the pair of payoff-ows that Adam and Eve receive in the social contract currently being operated. In the remainder, it will be scaled to

⁶Recall that it was based on the possibility of external enforcement.

$\gg = (0; 0)$.

Behind the veil of ignorance players apply Bayesian decision making and are hence interested in maximizing expected payoff[®] (cf. Harsanyi). However, since there is no external enforcement the bargaining set in the original position, T , does not equal $\frac{1}{2}(X_{AE} + X_{EA})$ as in Figure 2. The reason is that this set includes points that after the coin has been tossed leads one of the parties to ask for a return to the original position. Actually, this is reciprocity working: "If you don't do something for me, I'll certainly won't do anything for you". The only reasonable feasible set is then $T = X_{AE} \setminus X_{EA}$, i.e. the same set that Rawls used. However, here the conclusion is reached by reciprocity arguments instead of bluntly applying the maximin rule. The status quo however is the same as in Harsanyi's approach: $\zeta = \frac{1}{2}(\zeta + \zeta^3)$. It can be shown that the original position is unworkable if $\zeta \notin \zeta^3$, hence we will assume $\zeta = \zeta^3$. This boils down to saying that both players regard the terms of the current social contract to be fair. We now must find a continuous path from ζ to the Nash solution.

The analysis in Sections 2 and 3 assumed identical empathetic preferences: $U_1 = U_2 = U$ and $V_1 = V_2 = V$. This implies a short-run analysis, where personal and empathetic preferences are fixed. In the medium-run we would think of personal preferences as being fixed, while empathetic preferences may change due to social evolution. In the long-run both personal and empathetic preferences are flexible. Let us consider the short-run for now. Recall that Rawls arrived at the proportional bargaining solution applied to the rather artificial game $(X; \textcircled{®})$. It can be shown that in our case $\textcircled{®} = 0$. In our analysis we apply the Nash bargaining solution in the original position, resulting in a continuous path to the equilibrium. Note that along this path no player has an incentive to cheat⁷. Translating this equilibrium back to $(X; 0)$ yields a social contract r in X at which the Rawlsian social welfare function

$$W(x) = \min\{U(x_A; \gg_A); V(x_E; \gg_E)\} \quad (12)$$

is maximized, i.e. the proportional bargaining solution. Thus, in this case the proportional bargaining solution equals the Nash bargaining solution. In this way we arrived at Rawls' conclusion using Harsanyi's tools.

Of course, an important question remains, namely how are U and V determined. It lies beyond the scope of this report to go into the details, but the question is equivalent to asking how social evolution takes place, since this determines U and V . The interested reader is referred to Binmore (1998, x4.6.6).

⁷The equilibrium path is important to avoid endless renegotiation. Along the equilibrium path even players that don't trust each other have no incentive to cheat.

Since social evolution changes empathetic preferences, the concept of fairness changes over time. This means that time will eventually erode all moral contents of a fairness norm. However, decisions are taken in the short-run, hence the evasion of morality does not imply that justice doesn't matter. As soon as some unexpected event changes the set of feasible social contracts X , Adam and Eve return in the original position behind the veil of ignorance. Thus, fairness matters at each point in time in the short-run, although it does not necessarily imply the same fairness norms over time in the medium-run.

6 Whigs in Space

Philosophers (and not only they) appear to have an intrinsic need for classification. Utilitarianists and libertarianist are on the extreme sides of the political spectrum. Binmore's approach is a little bit harder to classify. He takes an intermediate position using utilitarian principles to underpin egalitarianism. Binmore himself calls it Whiggery. Its meaning is explained by the poet Yeats:

What is Whiggery?

A levelling, rancorous, rational sort of mind
That never looked out of the eye of a saint
Or out of a drunkard's eye.

In Binmore's own words:

...whigs first recognize that utopian aspirations should not be allowed to conceal the fact that stability is the prime need of a society. When contemplating reform, the feasible set of new social contracts should therefore be restricted to equilibria in the Game of Life. Moreover, the new social contract should be reachable from the current social contract by a process that is not itself destabilizing.

Such a position implies that society should be reformed based on fairness norms that we encounter every day. These simple fairness norms evolved from the time we lived in anarchic hunter-gatherer communities, where all operating fairness norms necessarily were self-enforcing. Cultural aspects enter by determining the standards of interpersonal comparison (U and V). These standards reflect the underlying power structure of a society.

As can be concluded from the analysis, Whiggery takes its place between utilitarianism on the left side and libertarianism on the right side of the political spectrum. It then

takes a position that has a tolerant attitude (libertarianistic) as well as a need to share (utilitarianistic). So, instead of Robert Nozick's idea of a minimal protective state (cf. Nozick, 1974), Binmore joins Rawls in advocating a protective as well as a productive state.

Finally, I would like to remark that philosophically seen Binmore's approach is interesting, since it combines the three mainstreams in political philosophy: utilitarianism, deontologism and naturalism. Binmore completes the demythologization of Kant's transcendental "I". The reason he can do so where Rawls failed is that Rawls sticks to metaphysical premises. The only thing he can then do is tell a more fashionable story, namely the original position. By using a naturalistic approach, Binmore can translate the original position from space back to earth using game theory. It's fascinating to see how game theory can give political philosophy the rigour it needs. It makes you wonder where we go next...

7 References

- Binmore, K. (1994) *Game Theory and the Social Contract, vol I: Playing Fair*, MIT Press, Cambridge, MA.
- Binmore, K. (1998) *Game Theory and the Social Contract, vol II: Just Playing*, MIT Press, Cambridge, MA.
- Harsanyi, J. (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge.
- Harsanyi, J. and R. Selten (1988) *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge, MA.
- Hume, D. (1978) *A Treatise on Human Nature*, Clarendon Press, Oxford. First published: 1739.
- Nozick, R. (1974) *Anarchy, State, and Utopia*, Basic Books, New York.
- Osborne, M. and A. Rubinstein (1994) *A Course in Game Theory*, MIT Press, Cambridge, MA.
- Rawls, J. (1971) *A Theory of Justice*, Harvard University Press, Cambridge, MA.
- Rousseau, J.J. (1996) *Du contrat social, ou Principes du droit politique*, Bookking International, Paris. First published: 1762.
- Sandel, M. (1984). "The Procedural Republic and the Unencumbered Self," *Political Theory* 12, 81-96.