

INDEX

Preface	2
Workshop report: Whigs in Space by Jacco Thijssen.....	4
NAKE teaching program 2000-2001.....	18
Program NAKE Day 2000.....	20
Registration form NAKE Day 2000 removable	middle page
Workshop report: Playing the Game of Integration by Mark Sanders.....	21

Preface

It is running towards the end of this rainy summer so we are getting ready for the new academic year. The successful 28th NAKE Workshop in a sunny Groningen finished off the year 1999-2000. It was the first workshop for Marty Roovers as NAKE-secretary and she did a great job. I would like to take the opportunity to thank her as well as the local organisers Erik Leertouwer, Philipp Maier, Remco van der Molen, Adriaan Soetevent and Linda Toolsema for the perfect organisation.

The following students received their NAKE diploma:

Hong Bo

Zolt Sándor

Dennis Botman

Edward Droste

Louise Grogan

Luc Moers

Congratulations with this achievement!

The reports on the interesting lectures during the Groningen workshop in June are not yet included in this issue of *NAKE Nieuws*. Instead, you will find the report on Ken Binmore's lectures on fairness at the December 1999 workshop that I promised you in the last issue of this periodical. In fact, as there were many very different, but equally good reports on these stimulating and provocative lectures, I decided to include two reports: one by **Jacco Thijssen** (Tilburg University) and another by **Mark Sanders** (UM).

This *NAKE Nieuws* also includes an overview of the NAKE Teaching Program of Utrecht courses for 2000-2001. As usual, many NAKE fellows were willing to teach a course and

I managed to compose a quite elaborate program, so there should be something to everybody's taste. You can enrol for the different courses via the NAKE homepage.

Elsewhere in this issue you find the program for the NAKE Day 2000, which will take place on October 13 at the Dutch Central Bank in Amsterdam. This year, the NAKE Day will be combined with the Tinbergen lecture organised by the Royal Netherlands Economic Association. This lecture will be delivered by Peter Diamond (MIT) and Willem Vermeend (Minister of Social Affairs). Note that the annual general meeting of members of the NAKE will take place over lunch during the NAKE Day. You can register for the NAKE Day either electronically (see the NAKE homepage) or with the registration form in this *NAKE Nieuws*.

I hope to see you in Amsterdam!

Best regards,

Lex Meijdam

Playing the Game of Integration

a Report on:

“Fairness” by Ken Binmore

NAKE Workshop Amsterdam

06-10 December 1999

by: Mark Sanders, Maastricht University

Professor Binmore’s lectures were beyond a doubt the most provocative and stimulating of this workshop. His game-theoretic approach to the evolution and function of fairness norms dominated discussions among students between lectures, during lunch and on occasion well into the night. This made me decide to write a report on his lectures. In an attempt to also illustrate some of the debate that went on outside the lecture hall, I have tried to tie the ongoing debate on the integration of minorities in. This debate was strongly revitalised during our stay in Amsterdam, due to the dramatic shoot-out in Veghel on Tuesday December 7 1999. This case will bring out the feelings of unease that many voiced during and outside the lectures. I stress that Prof. Binmore has not addressed any such issues in his lectures and cannot be held responsible for my mistakes in representing or applying his theory. But as he himself put it on several occasions, we should accept any conclusion provided it follows from sound reasoning. I leave it to the reader to assess whether my reasoning is sound. Prof. Binmore was the inspiration, the mistakes are mine.

Introduction

The integration of minorities in society is a problem with a long history. Pre-historic hunter gatherer groups only occasionally faced the problem of absorbing new members. In more

recent history there are many tales of violent clashes between groups of people that regarded each other as outsiders but later on managed to live alongside each other in peace. In modern day Western societies the increasing flow of economically and politically motivated migrants from developing countries is high on the political agenda. Since we do not regard a growing indigenous population or immigration from certain countries as a problem, apparently we do not feel that every new person poses a threat. It is people entering our society from another culture that we regard as a problem. The more deviant or unknown the culture these entrants come from the more they are regarded as a problem. Apparently these people have difficulty to function in our complicated societies. They come from another complicated society with a different set of cultural norms and values. This implies they sometimes behave in ways we cannot understand from our perspective and in much the same way it is hard for them to always understand how we behave from theirs. Thus the key to successful integration is finding a way to make these sets of norms and values more compatible. A first step in doing so is understanding where these norms and values come from and what function they play in the smooth running of society. A set of norms and values that are commonly shared among members of society can be referred to as a social contract. To understand its origin I will therefore first present the idea of the social contract and follow Prof. Binmore in putting it in a game theoretic perspective. This allows us to use game theory to work out a powerful theory of how a social contract evolves. Since pay-offs are a crucial element in game theory and some conceptual problems arise I devoted some paragraphs to a more detailed reproduction of Binmore's arguments for using cardinal utility functions in his theory. Then social evolution is presented as the driving force behind the norms and values that constitute a social contract and we may ask how the process of integration can be lubricated.

The Evolution of the Social Contract

In addressing the question: Where do social and cultural norms and values come from?, I will present Ken Binmore's theory on the game theoretic evolution of the social contract.

The social contract is defined as the set of all norms, values, regulations and the like that a society must and will adhere to in order to coordinate collective action and solve coordination problems that arise in the interaction between its members. In Binmore's terminology it is: *"... a set of commonly understood conventions about how behavior is to be coordinated, that is necessary to sustain an equilibrium in "the Game of Life".*" (Binmore 1994, p.6). Binmore's basic claim is that our capacity for moral judgement has evolved alongside the biological evolution of the human race to help us formulate just that. A set of commonly understood conventions that helps us cooperate and which are self-policing in the sense that no outside force, divine or otherwise, is needed to enforce these conventions. These conventions sustain a stable efficient and fair equilibrium in Binmore's Game of Life. Although all humans share this basic capacity the actual contents of the social contract can vary from one group of people to the next. Understanding the basic mechanisms that underlie the development of a stable, efficient and fair social contract is therefore helpful in understanding how differences in social contract may develop and how the integration of "aliens" in our society may be furthered.

In order to follow Binmore's argument we first need to refresh some basic game theory and see what constitutes the "Game of Life".

Life is a Game

A game in the game theoretical sense of the word is a well defined concept. A non-trivial game always has more than one players and these players can play more than one alternative strategies. Furthermore the reward or pay-off each player gets when a set of strategies is played usually depends on the strategy he has played himself and those played by others. Figure 1 represents a simple coordination problem as a game. The problem the drivers of two cars driving in opposite direction on the same road is: "On what side of the road should I drive?". In the Netherlands we have established a convention to answer this question and this presents no problem to any of us in every day life. One should, however, consider the problem before the convention is in place. Then all player A knows is the pay-

		Player A	
		Left	Right
Player B	Left	1	-10
	Right	-10	1

Driving Game*

		Player A	
		Confess	Deny
Player B	Confess	5	10
	Deny	10	2

Prisoners Dilemma**

* Pay-offs in utils i.e. high values are preferred

** Pay-offs in years in prison i.e. low values are preferred

Figure 1: Some examples of games

off he get when playing either a strategy to drive left or right as a function of the strategy played by player B. When A plays “drive left” and the B plays “drive left” they avoid an accident. The same holds if both play “drive right”. There is no reason to assume that either “social contract” is better than the other but it is definitely better than the contracts in which both choose a different strategy. If this is a one shot game and player A and B are not allowed to communicate the best they can do is flip a coin. Most situations, however, are repeated games. If player A and B come across the same road every day they can build a reputation for driving either right or left and knowing this the opponent can choose his best reply. If player A plays a best reply to player B and B plays a best reply to A we are in a Nash-equilibrium. In the driving game right-right and left-left, shaded in figure 1, are such “best replies” and happen to be optimal strategies in the sense that the pay-offs to both players are maximised. The second game, the famous prisoners dilemma, illustrates how Nash-equilibria can be sub-optimal. The game involves two criminals that are guilty of a joint crime but there is no evidence to convict them for it. They can be convicted for a lesser crime though. A confession of either would suffice to convict both. The optimal outcome

for the players would be that both deny and receive a light punishment for the lesser crime. They will, however, confess if the prosecutor offers the confessor to walk free. To see why, consider the pay-offs in this game. Assuming both dislike time in jail, are rational and the game is a one shot game, which implies there are no implications whatsoever from confessing other than walking free, the Nash-equilibrium is confess-confess. If we allow for someone from outside the game to police contracts, than our players can perhaps coordinate their actions. The godfather could be a very effective policeman to enforce the deny-deny equilibrium. Then, however, one should incorporate his actions into the pay-offs of both players and we no longer play a proper prisoners dilemma. In general the role of government and law enforcers is to do exactly that.

If the prisoners dilemma is played repeatedly, however, playing the one shot equilibrium every time is but one of the available equilibrium strategies and a policeman is not necessary to improve the outcome for both players. In the repeated prisoners dilemma sets of rules exist that can sustain the more efficient equilibrium deny-deny. A famous rule is the Tit-for-Tat rule, where player A and B both deny unless the opponent confessed in the preceding game. Such a rule may sustain the deny-deny equilibrium indefinitely. Several other rules exist and do the same. There are many more games we can think of and equilibrium strategies that players can play in these games. In that way we could construct the meta-game Binmore refers to as the “Game of Life”.

Binmore’s “Game of Life” is much too complex to be represented in the standard way as in figure 1, since it should incorporate all players that affect our lives at various points in our lives and deal with the infinite number of alternative strategies we and they can play. The “Game of Life” can be thought of as a game since the rules are set and are beyond the control of the players. Furthermore there is no outside agent that can alter the pay-offs to the players, prohibit or enforce any strategy or interfere in the game in any other way. The rules, strategies and pay-offs in the “Game of Life” are given to all the players and there is nothing we can do to change these settings of the game. One could think of our need for food or the laws of physics that drive and constrain our actions. Since society plays this game every day the Game of Life is truly a repeated game and it is therefore reasonable

to think of a social contract as the set of rules that sustains an equilibrium.

In the Game of Life the rewards are not in terms of years in prison or money but in terms of general evolutionary fitness. The feasible set of social contracts is limited to those that are stable or best replies as was explained above. Evolutionary stability is a stronger concept than game theoretic stability. An evolutionary stable equilibrium is also resistant to mutations. This means that non-equilibrium strategies played by a small subgroup of players cannot replicate faster than the equilibrium strategy players. This implies playing the non-equilibrium strategy must not yield a higher evolutionary fitness in a population dominated by equilibrium strategy players. For example in a population playing a repeated prisoners dilemma the Tit-for-Tat social contract is stable but mutant players that play confess all the time may come in and gradually become a sizeable share of this population. When their share in the population is big enough a small group of non-confessors can come in and take over the entire population. The so called Tat-for-Tit strategy might then come in and take over. This equilibrium has both game and evolutionary stability. Evolution will thus make sure the stable and efficient contracts eventually survive at the expense of the inefficient ones.

But how do we coordinate on one of the still potentially large set of evolutionary stable efficient social contracts in the meta-Game of Life? Binmore suggests that the social contract we actually operate is selected by an appeal to our notion of fairness. In Binmore's theory morality as such is only means to an end and no end in itself. Moral philosophers have frequently introduced some metaphysical perception of the "Good" in evaluating the fairness of outcomes and "Rights" to evaluate the fairness of the processes that yield outcomes. For Binmore, however, both have to be considered as social constructs that are specific to a social contract and can never be universal in nature. Binmore, placing himself in the tradition of David Hume, thus denounces the Kantian categorical imperative and claims no authority for metaphysical morality.¹

¹ It would take me too far astray to fully trace Binmore's theory in the history of moral philosophy. I refer the interested readers to volume I of his "Game Theory and the Social Contract", where he traces his own philosophical ancestry in great detail.

To operationalise fairness as a selection device for his purpose Binmore models the social contract that is actually implemented as the subject of a bargain between all players in Rawls' original position. That is a bargain where the players forget their roles in society and chance will allocate them a role after the bargain. In Rawls *A Theory of Justice* this theoretical construct is presented as the appropriate way to pass fair judgement on both outcomes and processes in social interaction. Rawls suggested we should not do unto others as we would have them do unto us by putting ourselves behind "a veil of ignorance" and completely forget our identity in society. In that situation every player will try to consider being in the other players shoes under the various possible social contracts. Rawls even went so far as to claim the only fair contract would then be the one that maximises the pay-off to the player worst off since everyone could turn out to be this player. Harsanyi (1977) showed that attaching equal probabilities to being any particular player leads to a social contract that maximises some weighted average of all players pay-offs as utilitarians often propose. Binmore presented us with evidence that the original position is very likely to be the basic mechanism by which all humans pass moral judgement over outcomes and processes. Nevertheless he differs of opinion with both Rawls and Harsanyi.²

Binmore frequently uses the bargain over a social contract between Adam and Eve in the original position to illustrate his point. I will follow him in that, however, to analyse the selection of a social contract in this simplified setting we first need to consider the representation of pay-offs in greater detail. Binmore uses von Neumann-Morgenstern utility and goes to great lengths to defend this representation of preferences in chapter 4 of vol. I, Binmore, (1994). I will shortly replicate his argument before proceeding with the analysis of social contract selection.

² For Rawls the veil is so thick that behind it we completely forget our identity. For Binmore we only pretend to forget but we actually bring our own extended preferences into the bargain. This implies Binmore differs with Rawls on his selection of a fair contract in the original position whereas he shares Rawls' egalitarian outcome. He agrees with Harsanyi on the thin veil type of bargain in the original position but claims Harsanyi's utilitarian conclusion is wrong because his theory lacks an explanation for the standards of inter-personal comparison. The following sections will clarify the importance of these differences.

Comparing Apples and Oranges

Suppose Adam and Eve have to decide on a “fair” social contract out of a potential set of stable and efficient contracts that specify what their role in society will be. In order to bargain over such a contract in the original position they must be able to put themselves in the other persons shoes to see their preferences over the potential contracts. To be able to do so Adam needs to know the order and intensity of his own and Eve’s preferences and vice versa. Furthermore, in a bargain behind the veil of ignorance, Eve’s preferences have to somehow be comparable to Adams. In every day life this is straightforward since we do this frequently but some problems arise when modelling this in a game. In economic terminology the above implies their preferences have to be measured cardinally and we have to introduce inter-personal utility comparison.

To rid ourselves of the problem that pay-offs in terms of years of prison, money or whatever other type of pay-off can have a different value to the same player under different circumstances we use the concept of utility. Having an umbrella when it rains might yield Adam say 1 util whereas having that same umbrella when the sun shines yields him disutility of -1 for having to carry it around. A utility function thus describes someone’s preferences by attaching a higher numerical value to more preferred outcomes of the game. These numbers are referred to as utils. One can compare this to measuring temperature in degrees. To use utils in game theory to represent pay-offs, however, we need to know not only which outcome is better than which but also by how much. This calls for a cardinal utility function. Intra-personal utility comparison is relatively unproblematic. We only make sure Adam values every util the same. To decide on a contract in the original position we furthermore need the same unit of utility for both (all) players. For such inter-personal comparison some way to allow Eve to convert Adam’s utils into her own and the other way around is necessary. This inter-personal standard of comparison is usually regarded as highly problematic by economists and moral philosophers alike since these standards seem rather arbitrary and yet are of crucial importance for the final outcome. Binmore claims these standards evolve from repeated bargaining over the social contract and can be

regarded as the core of what we consider as fairness in everyday life. This section proceeds first by deriving how a cardinal utility function over pay-offs can be constructed from an ordinal utility function over lotteries involving these pay-offs. Then we can derive how repeated bargaining behind the veil of ignorance leads to common standards of interpersonal comparison and how these help in selecting a fair social contract.

Gambling with Preferences

To derive a utility function that represents Adam's preferences in a cardinal way we start by asking Adam to express his preferences and indifferences over various lotteries where the prizes are in terms of the pay-offs we want him to value. In valuing the risk he implicitly values the assets at risk, even if that asset is his life for example, and assuming Adam is aware of the Laws of Probability makes his valuation cardinal in nature. The fact that probabilities are not always known presents a minor problem. We therefore only consider a simple example with known probabilities.

Suppose there are three possible outcomes, win (W), draw (D) and lose (L) and the probabilities of them occurring are known to be p , q and r in a lottery. Win is the most preferred outcome, draw is next in line and lose is least preferred. We will assume Adam has revealed preferences over lotteries $\mathbf{L}(p,q,r)$ involving these prizes at all values of p , q and r and there exists a function $V(\mathbf{L})$ that represents these preferences correctly by attaching a numerical value to each lottery. $V(\mathbf{L})$ is an ordinal utility function over lotteries. If preferences satisfy the Von Neumann and Morgenstern Axioms of choice we may develop this into a cardinal utility function over prizes. The Von Neumann Morgenstern representation of utility for $V(\mathbf{L})$ is then

$$V(\mathbf{L})=p*u(L)+q*u(D)+r*u(W) \quad (1)$$

where the function $u(X)$ is a cardinal utility function representing preferences over outcomes only and preferences over lotteries are a probability weighted average of the utility derived from prizes. In figure 1 this argument is illustrated. Lottery \mathbf{L} is presented in the first line. Von Neuman and Morgenstern now assume a rational being is indifferent between this simple lottery and one in which we play a lottery $\mathbf{M}(x)$ with probability x win vs $(1-x)$ lose instead of the outcomes lose, draw and win with probability $x=l$, probability $x=d$ and probability $x=w$ respectively for some l , d and w . Obviously $w=1$ and $l=0$ follow from $\mathbf{M}(0)=L$ and $\mathbf{M}(1)=W$. The value of d , however, captures Adam's attitude towards risk

and implicitly values the outcomes both ordinally and cardinally. If Adam believes in the laws of probability and lotteries **M** and **L** are independent, the complex lottery in line 2 can be simplified to a compound lottery **M**($pl+qd+rw$) and since Adam prefers winning he will

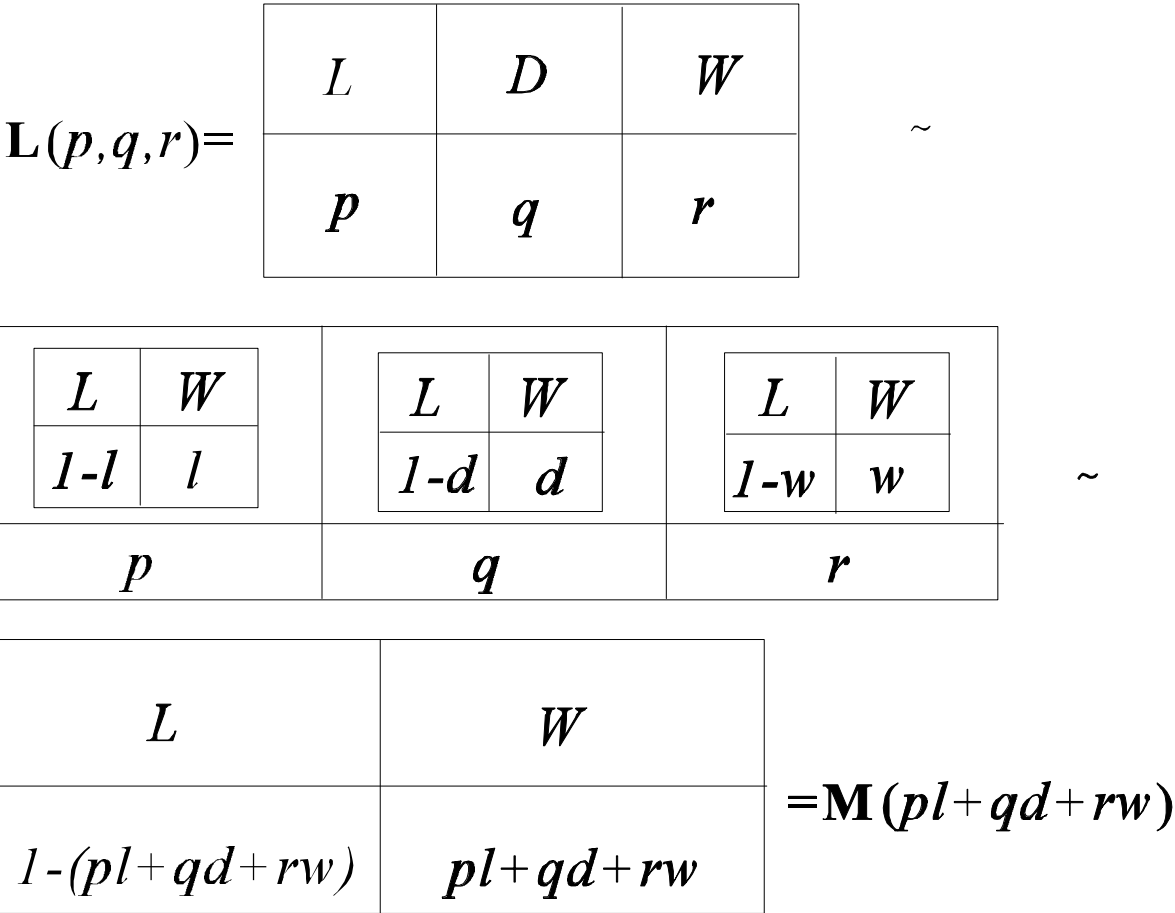


Figure 2: Von Neumann-Morgenstern’s Axioms

prefer lotteries with the highest value for $pl+qd+rw$.

Now a function $u(x)$ that satisfies $u(L)=l$, $u(D)=d$ and $u(W)=w$ is a cardinal utility function over outcomes and Adam’s preferences are adequately represented by the Von Neumann Morgenstern representation in (1). The unit of account, usually referred to as util, and the

zero of this function can be chosen arbitrarily by making linear transformations.³ What is crucial to note for our purpose is, however, that the difference in utils between two alternatives also contains information on the relative intensity of preferences.

To see this suppose Adam prefers A over B and C over D . If he feels stronger about the latter preference than Adam would always prefer lottery $\mathbf{A}(.5)$ over prices A and D over $\mathbf{B}(.5)$ over prices B and C . In terms of Adam's Von Neumann Morgenstern expected utility function this implies

$$.5*(u(A)+u(D))>.5*(u(B)+u(C)) \quad (2)$$

which implies the number of utils between A and B is smaller than between C and D thus reflecting adequately the relative intensities of the preferences. Von Neumann Morgenstern utility functions thus allow Adam to compare his pay-offs over the possible outcomes in a game.

The results above were derived under objectively known probabilities. The same argument can be made using subjective ones. One need only replace the objective probabilities with subjective ones and realise that people only have to act as if they have some conception of these probabilities.

Obviously the social contract Adam and Eve will select in the original position will exhaust all possibilities for increasing Adam's utility at no cost in terms of utils to Eve and vice versa. We should, however, also consider those contracts where a small sacrifice to Eve implies a large gain to Adam and the other way around. In constructing a criterium for evaluating the social contract teleological utilitarians construct a cardinal social welfare function that actually assumes an ideal observer with an implicit weighting scheme for aggregating individual utilities. Then it is possible to select the social contract that maximises the weighted sum of utils.

Binmore claims such inter personal aggregations are inadmissible. Fairness norms

³ The implicit assumption here is $u(L)=l=0$ and $u(W)=w=1$ and $u(D)=0<d<1$ but all transformations of the form $v(X)=a*u(X)+b$ yield a valid von Neumann Morgenstern cardinal utility function representing Adam's preferences

constitute such measures of inter-personal comparison as we see in next section.

If I were you...

Even though it may be impractical to construct Adam's Von Neumann Morgenstern utility function from his revealed preferences over lotteries involving all possible outcomes in the Game of Life, there is in principle no problem to do so. If we want Adam to compare his utils to those of Eve, as we ask of him in the original position, we need the above mentioned exchange rate or some standard of inter-personal comparison. Here empathetic preferences play a crucial role. They allow us to compare utils between individuals and provide the basis of rational (that is internally consistent) ethics. To discuss empathy, however, it is useful to discuss sympathy first.

Sympathy implies a particular type of identification with someone else. Their welfare effectively becomes your own. Eve's consumption of apples is for example in Adam's preference relations if he sympathises with her. Adam derives direct utility from Eve's eating an apple. In evolution we would expect sympathy to govern for example (extended) family relations, since the well-being of our kin increases the evolutionary fitness of our own genes since we share the same genes.

In considering non-family the best we can often do is empathise. We are aware of other peoples (revealed) preferences but do not derive utility ourselves directly from their well being. Combining man's ability to sympathise with his ability to envision himself in different situations as we did above, creates the possibility to extend our preferences to others by means of empathy. Empathy allows us to see things from someone else's point of view and predict behaviour and response to ones own actions. They allow an individual to say: I prefer being Adam with a fig leaf to being Eve with an apple.⁴ Empathetic preferences held by homo economicus can also be assumed to be consistent with Von Neumann Morgensterns Axioms of Choice. Homo economicus will make statements such

⁴ As opposed to: I prefer being Adam with a fig leaf to being Adam with an apple (personal preferences) or: I prefer being Adam while Eve has an apple to being Adam while Adam has a fig leaf (sympathetic preferences).

as the above one consistently and can represent his empathetic preferences in an extended von Neuman Morgenstern utility function.

This extended utility function attaches a number to each of the possible outcomes Adam faces in the original position. It attaches a value to being Adam having an apple, being Adam having a fig leave, being Eve having an apple and being Eve having a fig leave. Harsanyi (1977) tells us that Adam should attach equal probabilities to turning out to be Adam and turning out to be Eve in the original position. Thus Adam will maximise his expected utility⁵:

$$0.5*(v_A(C,A)+v_A(C,E)) \quad (3)$$

Assuming C has a worst and best outcome L and W respectively we can fix the scales of our personal utility functions but not our empathetic preference relation. For that we need to specify also the second argument. So arbitrarily we may set $v_A(L,A)=0$ and $v_A(W,E)=1$. As before the two remaining parameters $v_A(W,A)=U_A$ and $v_A(L,E)=1-V_A$ contain all the information we require to represent Adam's empathetic preferences in a Von Neumann Morgenstern cardinal utility function. To derive these parameters we set Adam's empathetic preferences when being Adam equal to a linear transformation of his own personal preferences,

$$v_A(C,A)=a* u_A(C) +b \quad (4)$$

and his empathetic preferences being Eve to a linear transformation of her (perceived) personal preference relation

$$v_A(C,E)=c* u_E(C)+d. \quad (5)$$

This implies

⁵ Binmore also presents the pay-offs under the Rawlsian assumption that probabilities are unknown and we therefore maximise the minimum.

$$v_A(C,A) = U_A^* u_A(C)$$

and

$$v_A(C,E) = 1 - V_A + V_A^* u_E(C). \quad (6)$$

Substituting into the expected utility Adam maximises in the original position and dropping constant terms we find Adam maximises:

$$U_A^* u_A(C) + V_A^* u_E(C), \quad (7)$$

a weighted average of both Adam's and Eve's personal utils, like a true utilitarian.

To be able to empathise with ones co-players in the game of life one has to thus know their personal preferences. In reality this seems to be a strong assumption but in many situations we only need to know our co-players preferences over a limited set of outcomes. The level of detail in the social contract determines the extent of the required information and we observe that for example one needs to know ones co-player a lot better to make a marriage work than to decide on which side of the road to drive.

The extended utility functions, defined to reflect empathetic preference relations for all players in all states of the world empathising with all adversaries, will include each players exchange rate for personal utilities from intra-personal empathetic preferences. The rates of exchange, however, are still idiosyncratic, that is individual. In the example above Adam and Eve may enter the original position with widely different U_i and V_i .

Still all moral judgement is captured by these "worthiness" coefficients. Whether Adam and Eve deem the contract fair or unfair depends crucially on their perception of their own and the others worthiness. If U happens to be zero for both and V happens to be one than the contract they select in the original position is the one that maximises Eve's utility at all costs to Adam. This does not imply Adam has no utility at all since he can have all the fig leaves if he is the only one that values them. The social contract will stipulate, however, that he must put all his efforts in collecting apples to please Eve. Both will accept Adam's servitude since both feel he is unworthy of consideration and Eve is worthy to be Adam's master. Both players will fail to come to an agreement if standards diverge and both

will bargain for the use of their standards with all the bargaining power they can muster. A social contract is only renegotiation proof when all share the same standards of inter-personal comparison.

Eventually society must and will converge upon such a common standard of inter-personal comparison through social evolution. Social evolution makes empathetic preferences converge as the U and V 's of winners are replicated and those held by losers are eradicated from the population.⁶ This implies all members of society will eventually go into the original position with the same standards of inter-personal comparison and will judge the proposed social contract fair or unfair according to the same shared standards. Such a social contract is an empathy equilibrium. It is not only proof to renegotiation but also to misrepresentations of members of society, who, in the original position, do not benefit from such misrepresentations. That is to say, since all have the same inter-personal standards of comparison there is no gain in leading others to believe yours are different.

The standards of comparison that evolve are then also used when we have to coordinate on new games. Facing a new situation, for example discovering the tree of wisdom, Adam and Eve still feel Eve is the only worthy person and Adam will offer her the apple. Only over time do empathetic preferences shift to support a new social contract that is more efficient in the new situation. If Adam and Eve could observe other couples finding the tree and those couples that share their standards of comparison being kicked out of Paradise, they would probably replicate the standards of those couples that were allowed to stay.

At this stage we can distinguish three equilibrating mechanisms running simultaneously in the "Game of Life". In the short run our empathetic and personal preferences are fixed. We decide what to do by choosing the strategy that maximises our expected utility given these preferences. This process takes place in economic time, typically on a daily basis. In social time our empathetic preferences may shift to adapt to new situations and reestablish a new stable and efficient empathetic equilibrium. This

⁶ Binmore (1998) is more explicit on this issue in chapter 2 of volume II.

social evolution is sometimes fast, sometimes slow but usually operates in a matter of months or years rather than generations or days. Finally our personal preferences may change over time as our biological evolution proceeds. These adjustments reestablishes evolutionary equilibrium after changes have occurred in our biological environment.

From empathising we construct social conventions (such as driving to the right) to co-ordinate on an (efficient) equilibrium quickly. In more intricate games we refer to fairness unaware of the fact that an empathy equilibrium also supports that convention. We are confronted with such hard realities only when our conventions fail to co-ordinate us on an equilibrium. This calls for social evolution to reach a new empathetic equilibrium.

Intruder Alert

Now we have a framework that explains the evolution of different sets of fairness norms in different societies we can proceed to analyse what happens if immigrants come to their new countries. As was argued above social evolution is required to change ones empathetic preferences and this is exactly what integration is about. Having similar standards of interpersonal comparison is the key to functioning properly in one's society and feeling happy about it too. To come back to the prehistoric hunter gatherers, in those times things were simple. You either adapt to the leader or the group makes you an outcast, reducing your chances of survival to zero. That is unless the intruder somehow has sufficient bargaining (muscle) power in the original position to force his standards on the group. The scarce evidence from isolated small hunter gatherer societies today suggests that these mechanisms are very powerful indeed.

In our modern day societies, however, many social contracts seem to operate alongside and on top of each other and friction is the obvious result. When new immigrants arrive they usually build up a social network among themselves and those who went before them, based on the social values and norms they had in their cultures of origin. This is logical and understandable since it is much easier to coordinate on equilibria in a sub-group that shares their standards of interpersonal comparison. The risk of costly mistakes

is a lot lower. The need to adapt and subscribe to a whole new social contract is thus limited. Italian, Irish, Hispanic and Chinese immigrants for example set up elaborate sub-cultures in the United States. These sub-cultures operate within the American society on their own sets of norms and values and only adapt what is essential for functioning in the new environment when it becomes essential.

Coming from the mafia dominated island of Sicily it is therefore no surprise that the Cosa Nostra set up shop in the US as well. The specifics of the arrangement were determined locally. Mafia organisations set up liquor smuggling, illegal gambling and prostitution in the American urban environment whereas they were involved in different activities in Sicily. The basic structure of the contract and the underlying standards of inter-personal comparison remained, however, unchanged. The authority of the head of the family was accepted and violent oppression of the weak by the strong a “legitimate” means of doing business. It is very difficult to judge the fairness of their way of life using Binmore’s theory. They would judge their own social contract unfair if they had my standards of inter-personal comparison but by the same token I would deem it fair if I had theirs.

As was argued above, this calls for a power struggle in the original position to resolve the differences and come to a mutual agreement. Understanding the source of the problem would not withhold many from strongly condemning their social contract as morally inferior and imposing their own standards through judicial and other means at their disposal and similarly the mafia resists these standards and tries to impose its on the rest of society by intimidation, corruption and the like.

This example also clearly illustrates that the indigenous people have all the reasons in the world to be worried when groups of immigrants come in. these sometimes pose a threat to the established way of life and these deviating standards of comparison may cause costly power struggles and result in inefficient strategies being played. Especially when sub-groups form it is unlikely that integration of these groups will leave their standards of comparison and way of life unaffected. In the previous example the power struggle is clear and most of us would have no problem choosing sides since most Dutch share the value that the weak are entitled to better treatment. Sometimes, however, the implications of Binmore’s positive theory go beyond the obvious and may even contradict our gut feelings about morality.

This brings me to the case at hand. On Tuesday, December 7th a 17 year old pupil at De Leijgraaf, a school in Veghel, Noord-Brabant, walked into the classroom and fired a gun

ten times, injuring 4 of his class mates and the teacher. The boy surrendered to the police and justified his actions by saying he was defending the family honour. The following days several headings in the paper illustrate the debate that his actions provoked. On Wednesday the 8th the headlines are objective in nature: "Five People wounded in Shoot-Out in School" and "Violence in Schools underestimated for a Long Time". Then first reactions came the next day: "Turks see Revenge as Justifiable Violence" and "Target of Shootout had been Threatened Before". After some reflection on Friday it read: "In the Netherlands Violence is a Crime" and "Revenge of Family Honour must be Punished under Dutch Law".

These headings show very clearly what went on in the minds of the people. First they took notice, then they tried to understand using additional information and finally they passed moral judgement after careful deliberation and came to the conclusion that such behaviour is not tolerable under the Dutch social contract. It is clear, however, that people recognise the importance of the Turkish social contract as an explanation for this boy's actions. It turned out that the target of the shoot-out had been involved in a love affair with the perpetrators sister in spite of her family's disapproval. Under the Turkish social contract this diminishes the worthiness of the (male) members of the family and even though Ali D. seemed to be well integrated in Dutch society he felt the urge or could be convinced to shoot the person responsible for this. The victim, also Turkish, and the Turkish community in the Netherlands claim to be shocked but insiders admit the violence does not surprise them.

Under the Turkish social contract, someone infringing upon one's honour is worthy of being violently straightened out by the person suffering the dishonour. According to a Turkish criminologist the majority of Turks living in the Netherlands would agree that violence to defend family-honour is not only justified but a duty to be carried out. According to him Turkish law allows judges to pass mild judgement on those defending their honour with violence and the custom is deeply rooted in Turkish tradition. The incident is distressing not only because of the victims and the associations with recent developments in the US. The fact that a Turkish boy, born and raised in the Netherlands, working at McDonald's, playing centerfield in Veghel's soccer team, maintaining good

relationships both in and beyond the local Turkish community, doing well at school, in short being the ideal integrated immigrant, still goes against the Dutch social contract when it comes to matters totally unimportant to the indigenous population.

We could also examine the position of the victim more closely. He was Turkish as well and hence can be expected to be at least aware of this aspect of the Turkish social contract. Living in the Netherlands, however, and being aware of the indigenous customs, may have led him to underestimate the threat of violent repercussions. The social contract works in Turkey because the social contract there justifies the use of violence and therefore makes the threat credible. Therefore the need to actually use violence is less likely to occur since it is an effective deterrent. In the Netherlands the credibility is reduced making violence more likely to occur. Binmore's theory thus allows us understand the event and this is an essential first step. However, what remains is: "How should we, the Dutch, respond?".

Playing with the New Kids on the Block

If our goal is to maintain our current Dutch social contract we have to eradicate the Turkish custom by punishing the perpetrator and hoping that will make him a loser not to be replicated both within and outside the Turkish community. From the headlines above, it seems this is exactly what will happen and there is little or no support for the position held by some that judges should take this boys cultural background into consideration.

Dutch judges, ironically, are instructed under our social contract to do exactly that. A judge will take all relevant issues into consideration and pressure from parents or social peer groups is usually taken into consideration in the Dutch legal system. If we would allow our judges to do so in this case, we allow this feature of the Turkish social contract to survive and possibly overtake our own conventions on the issue. If it does, one could say "so what?". Than family honour will be an issue worth killing and dying for for all of us in the long run. This would make the threat of violence credible and reduce the probability anyone will actually have to play the "shoot him" strategy. In Binmore's theory this is not bad or good, it is a fact.

But if we do not want our society to evolve in that direction, and we don't under our current standards of fairness, we should use every means available to convert the Turkish community to our standards of inter-personal comparison on this issue. A judge or for that

matter any Dutch citizens will realise the danger of allowing this to become a precedent by showing leniency. The key issue is, does this custom replicate and eventually overtake our current conventions. In this case we should convince them that family honour is worthless and worthiness does not depend on the extent to which the males in a family control the females.

Taken to the extreme this implies zero-tolerance for non-conformist behaviour and we would have to resist even the most innocent Turkish customs that reinforce the traditional Turkish role model for the family. Since we are the dominant cultural force in the Netherlands we have the power to impose our standards on minorities to quite some extent. Hence allowing Turkish women to wear traditional head covers at public schools or work would be out of the question. There are examples abound where use of our power is proposed by both left and right wing oriented people. Left wing feminists oppose tolerance for the dress code for Muslim women as they deem it oppressive and right wing politicians propose obligatory integration classes for new immigrants.

Fortunately in this case the custom is not very likely to replicate in our society and less drastic measures are required. The boy's parents apparently put pressure on him but it is unlikely that he will pass the idea of family honour on to his children to the same extent his father did. Mixed marriage, frequent interaction with Dutch mainstream culture and the gradual fading of parental influence as adolescents mature will probably be sufficient to eradicate this convention in the long run. The boy's punishment under Dutch law and the debate his actions provoked in the media will hopefully communicate to the Turkish community that this behaviour makes one a criminal, not an example.

I say fortunately because my (Dutch) standards of inter-personal comparison make me deem the proposed measures above unfair. Taking Binmore's theory to the extreme consequences of his reasoning and confusing his answers to the question "what is fairness?" with "what is fair?" caused most of the problems we discussed during lunch and over dinner. In Binmore's theory the social contract is void of moral content and no contract can be judged better than the next provided both are evolutionary stable and efficient. All problems we call moral dilemma's in modern society thus boil down to differences in standards of comparison and though Binmore's theory helps us understand

this, it is of no use in reconciling the dilemma's. Convergence of standards of comparison will eventually lead to the successful integration of minorities and in the long run we will all deem our social contract fair, whereas only bargaining power determines how it looks. The suggestion that all available power should therefore be used to force immigrants to accept our social contract is the only policy prescription one can formulate, if one accepts the preservation of the current social contract as the goal. Binmore would probably reply it is useless to formulate such a policy because the nature of power makes that it will even if another policy were formulated. It seems there is no Right or Wrong. If we have the power, we can impose our standards. If not, we cannot, and morality always justifies our actions, not because we do what is defined as Right or Good but because we define right and good by what we Do. Appeals to Fairness, Universal Rights and the Divine Good are no more and no less than rhetoric in our bargain behind the thin veil of ignorance.

References

Binmore, K. (1994), 'Playing Fair', in: *Game Theory and the Social Contract*, Vol. I, MIT Press, Cambridge MA

Binmore, K. (1998), 'Just Playing', in: *Game Theory and the Social Contract*, Vol. II, MIT Press, Cambridge MA

Van Dinther, M. (1999), 'Doelwit Schietpartij was al eerder Bedreigd', in: *de Volkskrant*, 9th December, p 3

De Graaf, P. (1999), 'Vijf Gewonden bij Schietpartij op School', in: *de Volkskrant*, 8th December, p 1

Luyten, M and M. Van Dinther (1999), 'Erewraak moet bestraft volgens Nederlands Recht', in: *de Volkskrant*, 10th December

Mesters, B. (1999), 'Turk ziet Wreken van Eer als Zinvol Geweld', in: *de Volkskrant*, 9th December, p 1

Various Authors (1999), 'Schietpartij', in: *de Volkskrant*, 9th December, p 13

Various Authors (1999), 'In Nederland is Geweld Misdadig', in: *de Volkskrant*, 10th December

NAKE teaching program 2000-2001

Block I: 8 September - 20 October 2000¹

10.00 - 12.00

00.78 Overlapping generations models Weddepohl

12.30 - 14.30

00.39 International environmental policies Jepma

00.73 Bayesian views on testing and model selection De Vos

15.00 - 17.00

00.49 Transfers and migration Meijdam & Verbon

00.74 Financial risk management De Vries & Lucas

Block II: 27 October - 1 December 2000

10.00 - 12.00

00.36 Nonlinear economic dynamics Hommes

00.11 Applied policy analysis Knot

12.30 - 14.30

00.43 Time series econometrics using state space methods Koopman

00.63 Competition and cooperation in the non-profit sector Ruys

15.00 - 17.00

00.56 Game theory Peters & Jansen

00.47 Competition and market coordination Maks

¹ There are no lectures on 13 October because this is the NAKE Day 2000.

Block III: 26 January - 2 March 2001**10.00 - 12.00**

- | | | |
|-------|--|------------|
| 00.07 | Applied General Equilibrium theory: Theoretical part | Keyzer |
| 00.80 | Behavioural modelling of government decision making | Van Winden |

12.30 - 14.30

- | | | |
|-------|--|---------|
| 00.30 | Recent developments in non-linear time series analysis | Franses |
| 00.38 | Adverse selection: Recent developments | Janssen |

15.00 - 17.00

- | | | |
|-------|--|--|
| 00.72 | Econometrics of panel data | Verbeek |
| 00.26 | Synthesis and comparison in economic research
(provisional) | Florax, Heijungs,
Nijkamp en Withagen |

Block IV: 16 March - 27 April 2001**10.00 - 12.00**

- | | | |
|-------|--|--------|
| 00.08 | Applied general equilibrium: Applied part | Keyzer |
| 00.66 | Experimental economics and the design of mechanism | Schram |

12.30 - 14.30

- | | | |
|-------|--|----------------|
| 00.17 | Agricultural policy analysis (provisional) | Burell & Oskam |
| 00.64 | Topics in oligopoly theory | Schoonbeek |

15.00 - 17.00

- | | | |
|-------|---|---------------------------------|
| 00.59 | New institutional economics | Potters |
| 00.28 | Capital market imperfections, investment
and monetary policy | Garretsen, Sterken &
Van Ees |

The course descriptions can be found on the NAKE homepage.
You can register electronically through the NAKE homepage.

NAKE Day 2000
October 13
Dutch Central Bank, Amsterdam

Program

- 9.00 Registration
- 9.15 Opening
- 9.30 Parallel sessions
- 11.00 Coffee
- 11.30 Parallel sessions
- 13.00 Lunch
- 13.30 General meeting NAKE members
- 14.00 Parallel sessions
- 15.30 Tea
- 16.00 Tinbergen Lecture: Peter Diamond
- 16.45 Tinbergen Lecture: Willem Vermeend
- 17.30 Drinks

Fee

The NAKE Day has a Dfl. 50,- fee, which includes your lunch and drinks.

Members of the Royal Netherlands Economic Association (Koninklijke Vereniging voor de Staathuishoudkunde; KVS) pay a reduced fee of Dfl. 20,-.

Please pay the exact amount cash at the registration desk on the NAKE Day.

Registration

You can register either electronically via the NAKE homepage:

<http://few.kub.nl/nake/nakedayinfo.htm> or with the registration form in this *NAKE Nieuws*.

Abstract

In case you want to present, please e-mail an abstract of the paper as soon as possible but no later than September 22, 2000 to the NAKE Secretariat:

NAKE@kub.nl

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 difference 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. Thijssen@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.

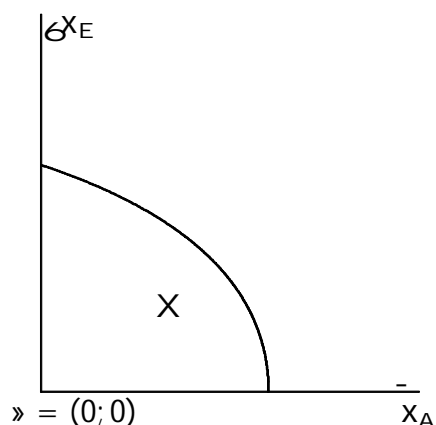


Figure 1: The Game of Life

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 utility 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,

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

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

Then, the set T is given by

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

The status quo \bar{z} is assumed to be given by $\bar{z} = \frac{1}{2}\bar{y} + \frac{1}{2}\bar{z}$, where \bar{y} and \bar{z} are the mappings of \bar{z} into X_{AE} and X_{EA} , respectively. This is depicted in Figure 2. The bargaining problem $(T; \bar{z})$ 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".

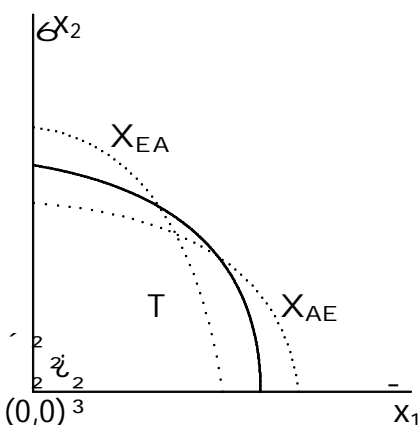


Figure 2: Harsanyi's original position

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 $z = (z_1, z_2)$. 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} .

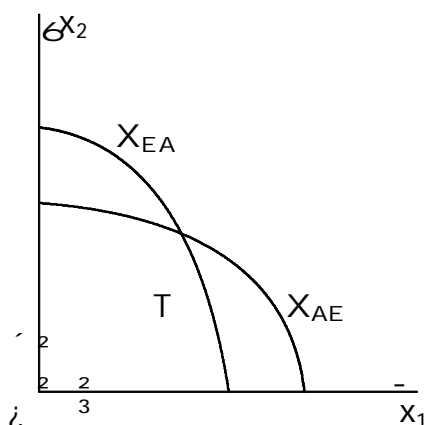


Figure 3: Rawls' original position

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

$$Ur_A = 1 - V(1 - r_E) \quad (11)$$

This is nothing else but the proportional bargaining solution with weights U and V for the bargaining problem $(X; \bar{c})$, where $\bar{c} = (0; 1 - V)$ (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).

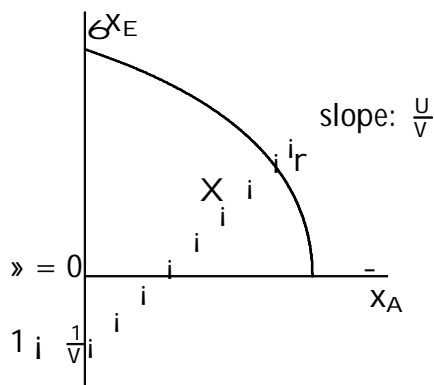


Figure 4: The social contract

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.