JUSTICE AS A NATURAL PHENOMENON

Ken Binmore

Inventing Right and Wrong

This article is my latest attempt to come up with a minimal version of my evolutionary theory of fairness, previously summarized in my book *Natural Justice.* The naturalism that I espouse is currently unpopular, but Figure 1 shows that the scientific tradition in moral philosophy nevertheless has a long and distinguished history. John Mackie's *Inventing Right and Wrong* is the most eloquent expression of the case for naturalism in modern times. Mackie's demolition of the claims made for *a priori* reasoning in moral philosophy seem unanswerable to me.

In Mackie's view, human morality is an artefact of our evolutionary history. To study it, he tells us to look at the anthropological facts presented in such pioneering studies as Westermarck's *Origin and Development of the Moral Ideas.* And for a framework within which to make sense of such anthropological data, he directs our attention to Von Neumann's theory of games.

My theory takes up where Mackie left off, but it doesn't treat the efforts of the metaphysical school as worthless. Their theories may rest on unsound foundations, but they are no less adept at observing ordinary people coping with ordinary life than Mackie. In particular, a naturalized version of the original position, independently formulated by the philosophers Harsanyi and Rawls, plays a leading role in my theory of fairness norms. But where Harsanyi and Rawls see a procedure for operationalizing Immanuel Kant's categorical imperative, I see a stylized version of the deep structure of the fairness norms that have evolved in our species to solve the equilibrium selection problems

doi:10.1017/S1477175609000025
Think 22, Vol. 8 (Summer 2009)

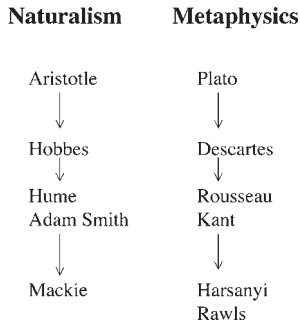

Figure 1: Naturalism and metaphysics in the history of moral philosophy.

that arise in the multitude of picayune coordination games of which everyday social life largely consists.

Game Theory

A game is any situation in which people or animals interact. The plans of action of the players are called strategies. Figure 2 shows two examples. The game on the left is the Prisoners' Dilemma. The game on the right is the Stag Hunt, which game theorists use to illustrate a story of Jean-Jacques Rousseau.

Each game has two players, whom I call Adam and Eve. Adam has two strategies, *dove* and *hawk* that are represented by the rows of the payoff table. Eve also has two strategies, *dove* and *hawk*, represented by the columns of

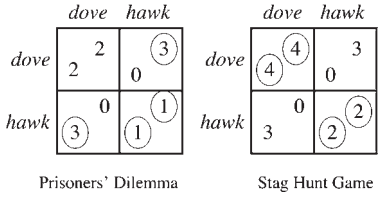

Figure 2: Two toy games.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

the payoff table. The four cells of the payoff table correspond to the possible outcomes of the game. Each cell contains two numbers, one for Adam and one for Eve. The number in the southwest corner is Adam's payoff for the corresponding outcome of the game. The number in the northeast corner is Eve's payoff.

The payoffs may correspond to money or to biological fitness, but they don't need to. With mild assumptions, economists have shown that any consistent behaviour whatever can be modeled by assuming that the players are behaving as though seeking to maximize the average value of *something.* This abstract something – which obviously varies with the context – is called utility. When assuming that a player is maximizing his or her expected payoff in a game, we aren't therefore taking for granted that people are selfish, or victims of their genes. We make no assumptions about their motivations, except that they pursue their goals – whatever they may be – in a consistent manner.

It would be easy for the players to maximize their expected payoffs if they knew what strategy their opponent was going to choose. For example, if Adam knew that Eve were going to choose *dove* in the Prisoners' Dilemma, he would maximize his payoff by choosing *hawk*. That is to say, *hawk* is Adam's best reply to Eve's choice of *dove*, a fact indicated in Figure 2 by circling Adam's payoff in the cell that results if the players choose the strategy profile (*hawk*, *dove*). However, the problem in game theory is that a player doesn't normally know in advance what strategy the other player will choose.

A Nash equilibrium is any profile of strategies – one for each player – in which each player's strategy is a best reply to the strategies of the other players. In the examples of Figure 2, a cell in which both payoffs are circled therefore corresponds to a Nash equilibrium.

Nash equilibria are of interest for two reasons. If it is possible to single out the rational solution of a game, it must be a Nash equilibrium. For example, if Adam knows that Eve is rational, he would be stupid not to make the

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

best reply to what he knows is her rational choice. The second reason is even more important. An evolutionary process that adjusts the players' strategy choices in the direction of increasing payoffs can only stop when it reaches a Nash equilibrium.

Because evolution stops working at an equilibrium, biologists say that Nash equilibria are evolutionarily stable. Each relevant locus on a chromosome is then occupied by the gene with maximal fitness. Since a gene is just a molecule, it can't *choose* to maximize its fitness, but evolution makes it seem as though it had. This is a valuable insight, because it allows biologists to use the rational interpretation of an equilibrium to predict the outcome of an evolutionary process, without following each complicated twist and turn that the process might take.

Why, for example, do songbirds sing in the early spring? The proximate cause is long and difficult. This molecule knocked against that molecule. This chemical reaction is catalyzed by that enzyme. But the ultimate cause is that the birds are signaling territorial claims to each other in order to avoid unnecessary conflict. They neither know nor care that this behaviour is rational. They just do what they do. But the net effect of an immensely complicated evolutionary process is that songbirds behave *as though* they had rationally chosen to maximize their fitness by operating a Nash equilibrium of their game of life.

Equilibrium selection problem

When a particular game is played many times in a society, people get accustomed to playing it in a particular way. For example, each time we get into our car in the morning we are playing a Driving Game with all the other people driving to work. In Britain, we keep the accident rate down by all driving on the left. In France, the convention is to drive on the right. The side of the road on which it is conventional to drive is the archetypal example of a *social norm*.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

The feature of a social norm that needs to be emphasized here is that it won't survive if its use fails to coordinate the players' behaviour on a Nash equilibrium of whatever game is being played. Immanuel Kant's categorical imperative is therefore a non-starter as a stable social norm because it calls for Adam and Eve to play *dove* in the Prisoners' Dilemma, but the only Nash equilibrium in the Prisoners' Dilemma requires that Adam and Eve both play *hawk*.

The Prisoners' Dilemma is unusual in having only one Nash equilibrium. Nearly all games that arise in real life have many Nash equilibria. A society that operates a particular social norm must therefore have solved an equilibrium selection problem. Societies sometimes choose their social norms consciously and deliberately, as when Sweden switched from driving on the left to driving on the right in the early hours of 1st September, 1967. But many social norms are the end-product of an uncontrolled process of cultural drift.

The Stag Hunt Game is used to illustrate one of the many difficulties that the equilibrium selection problem creates. In Rousseau's story, Adam and Eve agree to cooperate in hunting a stag, but when they separate to put their plan into action, each may be tempted to abandon the joint enterprise by the prospect of bagging a hare for themselves.

The circled payoffs in the payoff table show that there are two Nash equilibria in pure strategies, one in which the players cooperate by both playing *dove*, and one in which they defect by both playing *hawk*. The situation differs from the Prisoners' Dilemma in that both of these patterns of behaviour are viable candidates for social norms because both correspond to Nash equilibria.

Although both Nash equilibria in the Stag Hunt Game are candidates for a social norm, the efficient (waste-free) equilibrium in which Adam and Eve both play *dove* assigns both players a higher payoff. So why don't Adam and Eve agree to make the efficient equilibrium their social norm?

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

However, moving from an inefficient social norm to a new social norm isn't necessarily easy. In Naples, the inefficient social norm in which traffic signals are ignored is so ingrained that progress seems impossible.

The Stag Hunt Game shows how hard it can be even for fully rational players to move from an inefficient equilibrium to an efficient equilibrium. Assuming that the current social norm is for Adam and Eve both to play *hawk*, then Adam may seek to persuade Eve that he plans to play *dove* in the future, and so she should follow suit. But she will remain unconvinced, because whatever Adam may actually be planning to play, it is in his interests to persuade her to play *dove*. If he succeeds, he will get 4 rather than 0 if he is planning to play *dove*, and 3 rather than 2 if he is planning to play *hawk*.

Rationality alone therefore doesn't allow Eve to deduce anything about his plan of action from what he says, because he is going to say the same thing no matter what his real plan may be!

In spite of these problems, I think we should sometimes expect biological and cultural evolution acting in tandem to select an efficient equilibrium in the long run. To see why, suppose that many identical small societies are operating one of two social contracts, *a* and *b*. If *a* makes each member of a society that operates it fitter than the corresponding member of a society that operates *b*, then here is an argument which says that *a* will eventually come to predominate.

To say that a citizen is fitter in this context means that the citizen has a larger number of children on average. Societies operating social contract *a* will therefore grow faster. Assuming societies cope with population growth by splitting off colonies which inherit the social contract of the parent society, we will then eventually observe large numbers of copies of societies operating social contract *a* compared with those operating contract *b*. But this is how evolution works. In this case, the efficient social contract *a* will have proved fitter than its inefficient rival *b.*

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

Coordination games

I think that fairness evolved as Nature's answer to the equilibrium selection problem in human coordination games. As in the Driving Game of Figure 3, such games have more than one efficient equilibrium. However, the Battle of the Sexes is more typical of such games, because Adam and Eve don't agree about which of the two efficient Nash equilibria is preferable as a social norm.

The politically incorrect story that accompanies the Battle of the Sexes makes Adam and Eve a newly married couple on their honeymoon in New York City. At breakfast, they discussed whether to attend a boxing match or the ballet in the evening, but without reaching an agreement. During the day they got separated in the crowds, and they must now choose where to go in the evening independently.

As with the Stag Hunt Game, the purpose of the Battle of the Sexes is to illustrate how hard it can be to solve equilibrium selection problems. But not all equilibrium selection problems are so difficult. We commonly solve coordination problems by appealing to an appropriate fairness norm without any thought or discussion. Who goes through that door first? How long does Adam get to speak before it is Eve's turn? Who moves how much in a narrow corridor when a fat lady burdened with shopping passes a teenage boy with a ring through his nose? Who should take how much of a popular dish of which there isn't enough to go around? Who gives way to whom when cars are maneouvring in heavy traffic? Who gets that parking

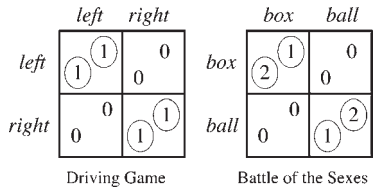

Figure 3: Coordination games.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

space? Whose turn is it to wash the dishes tonight? These are picayune problems, but if conflict arose every time they needed to be solved, our societies would fall apart.

Most people are surprised at the suggestion that there might be something problematic about how two people pass each other in a corridor. When interacting with people from our own culture, we commonly solve such coordination problems so effortlessly that we don't even think of them as problems. Our fairness program then runs well below the level of consciousness, like our internal routines for driving cars or tying shoelaces. As with Moliere's Monsieur Jourdain, who was delighted to discover that he had been speaking prose all his life, we are fair in small-scale situations without knowing that we are fair.

Nash Demand Game

The coordination games that have been discussed so far are too simple to serve as a model of the coordination games that gave rise to the human sense of fairness. For this purpose, we need to consider games with a continuum of efficient equilibria. Sharing food is the paradigmatic example. If Adam and Eve have the time and opportunity, they may negotiate some compromise split of the available pie, but even in the absence of any explicit communication, it still often makes sense to regard their dilemma as a bargaining problem.

The simplest coordination game that meets our requirements is called the Nash Demand Game. Adam and Eve simultaneously choose strategies that translate into payoff demands. If the pair of demands they make lies in the shaded set of Figure 4, then both players receive their demands. If not, then the result is the payoff pair called the state of nature in Figure 4.

The location of the Nash bargaining solution is determined entirely by the shape of the shaded set and the location of the state of nature. The location of the utilitarian

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

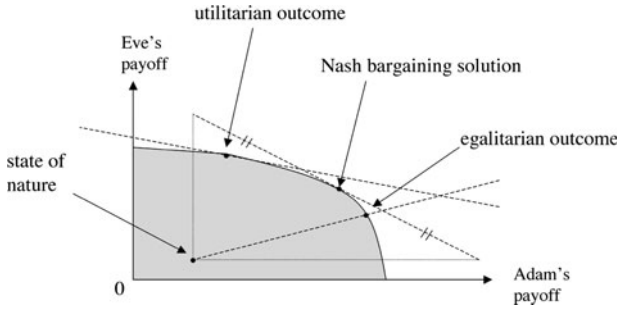

Figure 4: Three possible social norms.

and egalitarian solutions depends on the slope of the lines that determine them. These slopes are determined in turn by the standard of interpersonal comparison of utility that society currently operates.

Any payoff pair on the curved boundary of the shaded set of Figure 4 that is better for both players than the state of nature corresponds to an efficient equilibrium of the game.

Neither player can gain by asking for more, given the demand made by the other. It is also an (inefficient) Nash equilibrium for both players to make huge demands that cannot be met, in which case the outcome is the state of nature. We therefore have a game that shares the properties of both the Stag Hunt Game and the Battle of the Sexes, but differs from the Battle of the Sexes in that compromise is possible.

It may be that Adam and Eve have been operating an inefficient social norm that would land them at the state of nature in the Nash Demand Game that they now have to play. They would do better to adopt a new social norm that coordinates their behaviour on one of the game's many efficient Nash equilibria. Figure 4 shows three possible social norms that might be used to solve their equilibrium selection problem, each of which has been touted as the fair social norm by various authors.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

The Nash bargaining solution has no virtues at all as a fairness norm, having been explicitly constructed by Nash to predict the outcome when Adam and Eve ignore fairness considerations in favour of bringing to bear whatever bargaining power they may have at their disposal. However, arguments can be marshalled for both the utilitarian and the egalitarian social norms. How do we distinguish between them? John Mackie told us to look to anthropology if we want such questions answered.

Anthropology

Ten thousand years or more ago, humans lived in smallish groups of up to 120 or so people, who survived by hunting and gathering. No such groups exist any more, but anthropologists were able to study their way of life before it became extinct. Two features are universal across the world. African pygmies, Andaman islanders, Greenland eskimos, Australian aborigines, Paraguayan indians, and Siberian nomads all operated societies without bosses or social distinctions. They also shared food – especially meat – on a relatively equal basis.

Folk theorem

As in small villages today, it must have been hard to keep any secrets in an ancestral foraging group. This fact is helpful when applying game theory to such a society, because it allows the folk theorem of repeated game theory to be applied. This is perhaps the most significant result that game theory has to offer to political philosophy. It tells us what Nash equilibria are available when a group of sufficiently forward-looking people play the same game every day for an indefinite period.

To understand the result, imagine a benign social planner assisted by an all-powerful police force, who is able to enforce any outcome he chooses in the daily game

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

the group play. The folk theorem says that the same outcome is available as the outcome of a Nash equilibrium in the repeated game. That is to say, the players don't need to be forced to behave well by some external agency. Nor do they need to share the altruistic leanings of Dr Jekyll. A society of selfish Mr Hydes can be just as successful in coordinating their behaviour on a cooperative outcome, because any such outcome is available as a Nash equilibrium in the repeated game. Nobody can profit by cheating, because the strategies required by the Nash equilbrium call on the other players to punish anyone who deviates. Or to say the same thing another way, our ancestors needed no police force to enforce their good behaviour, because they policed each other.

The folk theorem allows us to use Figure 4 as a representation of the repeated game of life played by an ancestral hunter-gather group. If the state of nature represents the outcome of one Nash equilibrium of the game, then any outcome in the shaded set that everybody in the group prefers to the state of nature is also a Nash equilibrium outcome of the repeated game – and therefore available as an alternative solution to their equilibrium selection problem.

Leadership?

Modern societies mostly resolve their equilibrium selection problems by delegating the choice of equilibria to leaders at various levels. However, if the foraging societies studied in modern times are any guide, protohuman foraging societies had no leaders. I shan't repeat the speculations from my *Natural Justice* on why such an anarchic a form of social organization should have conferred an evolutionary advantage as compared with the dominance hierarchies that are normal among the other species of great apes, but simply accept that the leadership solution to the equilibrium selection problem was unavailable to our foraging

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

ancestors. So how did they solve their equilibrium selection problems? In particular, how did they decide who should get how much meat after a successful cooperative hunt?

Fairness

A social species that doesn't have leaders to nominate an equilibrium in a society's game of life must use some other equilibrium selection device. In the case of our own species, the device that evolved was fairness. If some other social norm had evolved to solve the equilibrium selection problem, philosophers who believe in moral absolutes would nowadays be singing the virtues of the properties of this other norm, rather than the properties of the norm that actually did evolve. But what are these properties? How do our fairness norms work?

An Origin for the Golden Rule?

Westermarck was notorious as an unrepentant moral relativist, but even he recognized that the Golden Rule – do as you would be done by – seems to be universal in human societies. Is there any reason why evolution should have written such a principle into our genes? If the Golden Rule is understood as a simplified version of the device of the original position, I think an answer to this question can be found by asking why social animals evolved in the first place. This is generally thought to have been because food-sharing has survival value.

The original position

John Rawls made the original position famous in his celebrated *Theory of Justice.* John Harsanyi independently invented the original position at around the same time. Both give credit to earlier scholars who toyed with the same idea.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

Rawls uses the original position as a hypothetical standpoint from which to make judgments about how a just society would be organized. Members of a society are asked to envisage the social contract to which they would agree *if* their current roles were concealed from them behind a 'veil of ignorance'. Behind this veil of ignorance, the distribution of advantage in the planned society would seem as though determined by a lottery. Devil take the hindmost then becomes an unattractive principle for those bargaining in the original position, since you yourself might end up with the lottery ticket that assigns you to the rear.

Rawls defends the device of the original position as an operationalization of Immanuel Kant's categorical imperative, but I think this is just window-dressing. The idea certainly hits the spot with most people when they hear it for the first time, but I don't believe this is because they have a natural bent for metaphysics. I think it is because they recognize the deep structure of the fairness norms that they actually use every day in solving the equilibrium selection problem in the myriads of small coordination games of which daily life largely consists.

Implicit insurance contracts

How and why might the neuronal wiring necessary to operate the original position have evolved? Imagine a time before cooperative hunting had evolved, in which Adam and Eve foraged separately for food. They would sometimes come home lucky and sometimes unlucky. An insurance pact between them would specify how to share the available food on days when one was lucky and the other unlucky.

If Adam and Eve were rational players negotiating an insurance contract, they wouldn't know in advance who was going to be lucky and who unlucky on any given day on which the contract would be invoked. They would then be bargaining behind a veil of uncertainty that conceals who is going to turn out to be Ms Lucky or Mr Unlucky.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

Both players then bargain on the assumption that they are as likely to end up holding the share assigned to Mr Unlucky as they are to end up holding the share assigned to Ms Lucky.

In the original position, Adam and Eve are no longer uncertain about whether they will turn out to be Ms Lucky or Mr Unlucky. The setup requires instead that they put themselves in the positions of two new players who behave as though they are uncertain whether they will turn out to be Adam and Eve. That is to say, they must imagine themselves in the shoes of somebody else – either Adam or Eve – rather than in the shoes of one of their own possible future selves. Space does not allow me to expand on the parallels between the two situations, but I hope it is clear that if Nature wired us up to solve the simple insurance problems that arise in food-sharing, then she also simultaneously provided much of the wiring necessary to operate the original position.

Utilitarianism or Egalitarianism?

In analyzing the bargaining problem faced by Adam and Eve behind the veil of ignorance, Rawls was led to an egalitarian outcome. Harsanyi argued that the outcome would be utilitarian. Who was right? With the assumptions they both made, I think the answer is Harsanyi. Rawls finds his way to an egalitarian conclusion only by the iconoclastic expedient of denying orthodox decision theory. However, I think that it was Rawls who had the better intuition.

Both Harsanyi and Rawls postulate an external enforcement agency that polices the hypothetical deal reached by Adam and Eve in the original position. Harsanyi invents an enforcement agency called “moral commitment’’. Rawls’ agency is called “natural duty”. But a naturalist like myself has little patience with such metaphysical fancies. If we accept that any potential policemen must themselves be players in the game of life, any social norm must necessarily

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

be self-policing in order to survive. That is to say, it must coordinate our behaviour on a Nash equilibrium of the game.

If we give up the idea of external enforcement altogether, and insist that all aspects of the original position must be self-policing, I prove in my book *Just Playing* that the final outcome of rational bargaining in the original position must be egalitarian in the sense illustrated in Figure 4. The proof simply consists of applying the Nash bargaining solution to the problem faced by Adam and Eve in the original position, and following the logic wherever it goes.

Interpersonal comparison

The egalitarian result is consistent with Aristotle's intuition that "what is just is what is proportional". There is a body of psychological evidence which suggests that laboratory subjects also see things this way. But what should be regarded as proportional to what?

This question raises the issue of interpersonal comparison. At what rate do Adam's utils get traded off against Eve's utils? Answering this question is my main contribution to the debate, but all I can say here is that my theory assumes that the appropriate standard of interpersonal comparison is determined by cultural evolution, and so depends on the history of experience of a particular society. For example, my theory suggests that it will always be regarded as fair for a person with high social status to get a smaller share than a less exalted individual, but the exact amount by which their shares differ will depend on the cultural idiosyncracies of the society in which they live.

Moral Relativism

My theory is analogous to current beliefs about the nature of language. The original position corresponds to Chomsky's deep structure of language. If I am right, this deep structure of fairness is written in our genes, and

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

hence is universal in the human species. However, critics prefer to focus on the fact that the appropriate standard of interpersonal comparison varies with the culture in which a fairness norm operates – just as the details of the particular language we learn depend on the culture in which we were brought up.

My theory of fairness is an attempt at a descriptive theory; it seeks to explain how and why fairness norms evolved. Karl Marx might respond that it is all very well seeking to understand society, but the point is to change it, and I don't disagree. I hope very much that the scientific study of how societies really work will eventually make the world a better place for our children's children to live in, by clarifying what kind of reforms are compatible with human nature, and which are doomed to fail because they aren't.

As an example, consider the pragmatic suggestion that we might seek to adapt the fairness norms that we use on a daily basis for settling small-scale coordinating problems to large-scale problems of social reform. This is one of the few things I have to say that traditional moralists find halfway acceptable. But they want to run with this idea without first thinking hard about the realities of the way that fairness norms are actually used in solving small-scale problems. In particular, they are unwilling to face up to the fact that fairness norms didn't evolve as a substitute for the exercise of power, but as a means of coordinating on one of the many ways of balancing power.

This refusal to engage with reality becomes manifest when traditionalists start telling everybody how they 'ought' to make interpersonal comparisons when employing the device of the original position. But if I am right that the standards of interpersonal comparison we actually use as inputs when making small-scale fairness judgments are culturally determined, then these attitudes will necessarily reflect the underlying power structure of a society. One might wish, for whatever reason, that these attitudes were different. But the peddling of metaphysical arguments about what would be regarded as fair in some invented

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press

ideal world can only muddy the waters for practical reformers who actually have some hope of reaching peoples' hearts. Nobody is going to consent to a reform on fairness grounds if the resulting distribution of costs and benefits seems to them unfair according to established habit and custom, whatever may be preached from the pulpit.

This pragmatic attitude mystifies traditional moralists, who pretend not to understand how a naturalist like myself can talk about optimality at all. How do I know what is best for society? From whence do I derive my moral authority? Where are my equivalents of the burning bush and the tablets of stone?

The answer is that I have no absolute source of moral authority to which to appeal – but nor does anyone else. I know that my aspirations for what seems a better society are just accidents of my personal history, and that of the culture in which I grew up. If my life had gone differently or if I had been brought up in another culture, I would have different aspirations. But I nevertheless have the aspirations that I have – and so does everyone else.

The only difference between naturalists and traditionalists on this score is that naturalists don't try to force their aspirations on others by appealing to some invented source of absolute authority. The reality is that if enough people with similar aspirations are sufficiently close to the levers of power, they shift the social contract because that is what they want to do. Reforms never get implemented in any other way. As General Napier said when asked to tolerate the Hindu practice of suttee:

> You say that it is your custom to burn widows. Very well. We also have a custom: when men burn a woman alive, we tie a rope around their necks and we hang them. Build your funeral pyre; beside it, my carpenters will build a gallows. You may follow your custom. And then we will follow ours.

Ken Binmore is a visiting Professor of Philosophy at the London School of Economics.

https://doi.org/10.1017/S1477175609000025 Published online by Cambridge University Press