GAME THEORY AND MORAL NORMS:

AN OVERVIEW AND AN APPLICATION*

Bruno Verbeek

Australian National University

1. Introduction

In 1954, the British philosopher Richard Braithwaite gave his inaugural lecture entitled Theory of Games as a Tool for the Moral Philosopher.¹ In his lecture Braithwaite argued that problems of distributive justice have the same structure as the bargaining problem, which had been analyzed some years before by John Nash, the later Nobel Prize winner, using game theory. Braithwaite predicted that game theory would fundamentally change moral philosophy. His prediction came less than ten years after the publication of John von Neumann and Oskar Morgenstern’s Theory of Games and Economic Behaviour – a book that started a completely new branch of social science and applied mathematics.² In hindsight, this was the moment that modern game theory entered the discipline of ethics.

In a way, the introduction of game theory in ethics was a sort of homecoming. For example, game-theoretic ideas can be found in the work of Thomas Hobbes and David

---

* This article is a translation and adaptation of my essay ‘Functie en Autoriteit van Morele Normen’ Algemeen Nederlands Tijdschrift voor Wijsbegiarte 94, no. 1 (2002).


Hume. In spite of this shared intellectual past, Braithwaite's prediction has not (yet?) come true. Moreover, it took quite some time before other philosophers became as enthusiastic as Braithwaite. Ten years later, in 1964, Brian Barry published *Political Argument* and in 1969 David Lewis’ seminal work *Convention* came out.\(^3\) Notwithstanding this faltering start, the introduction of game theory in moral philosophy has produced a steadily increasing flow of publications.

### 2. Functionalism

Initially game theory was used to analyze the *function* of morality. A good example is Edna Ullman-Margalit’s *The Emergence of Norms*, in which she demonstrated that morality enables agents to cooperate and coordinate their actions in situations where the pursuit of self-interest prevents this.\(^4\) Her by now classic example is that of two artillerymen who face the choice to flee from the advancing enemy or stay and operate their gun. Their gun is located in a strategically important pass. If both stay, they have a chance of getting wounded, but it is certain that the advance of the enemy will be halted. If both flee, the enemy will be able to take the mountain pass, overtake them and make them POW-s. If just one of them stays while the other flees, the brave artillerist will die in battle, but the other gunner will have just enough time to escape safely. Supposing that both try to survive this ordeal, preferably unhurt, each soldier has good reasons to flee.

We can understand this claim using a simple two-by-two matrix (Figure 1). Each gunner has the choice between fleeing and staying and fight. This choice is represented in

the rows for gunner #1 and the columns for gunner #2. Each cell in the matrix represents the outcome of each possible pair of choices. Each cell has a pair of numbers. The number in the lower left corner of each cell represents how gunner #1 values this outcome, relative to the other possible outcomes. It is in the jargon of decision theory, #1’s utility of this outcome. The number in the upper right corner represents the evaluation of this outcome by #2.

<table>
<thead>
<tr>
<th></th>
<th>stay</th>
<th>flee</th>
</tr>
</thead>
<tbody>
<tr>
<td>stay</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>flee</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1

Now we can see why both gunners would choose to flee. Consider the case for #1. Suppose #2 decides to stay and fight. In that case, #1 is best off by fleeing. He will survive without getting hurt. In the formal representation of the matrix, he will secure a utility of 3 rather than 2. Suppose #2 decides to flee. Again, #1 does best by fleeing. He will survive the battle, although he will be imprisoned for the duration of the war. If he were to stay and fight, he would certainly die. That is, by fleeing he will secure a utility of 1 rather than 0. Gunner #2 is in the same position as #1: regardless of what #1 does, he will be better off by fleeing.

---

In short, each individual gunner would be better off fleeing, regardless of what the other does. However, the paradox is that both would be better of if both stood their ground. That is, there is a Pareto efficient outcome that will not be realized if both artillerists choose rationally. Suppose that both understand the structure of their predicament. Since they would see that each has good reasons to flee, they could try to rule out this possibility. For example, they could chain each other to the gun en thus make it impossible to flee. (I am told that 18th century artillerists actually did this!) Ullman-Margalit argued that the situation of the gunners (i.e., the so-called prisoner’s dilemma) is structurally equivalent to many everyday interactions. Paying taxes, keeping certain kinds of promises, honoring contracts, in short, all those situations in which doing the right thing requires a sacrifice of one’s interests, are like the predicament of the gunners. Just like the gunners had to resort to measure to prevent each other from fleeing, there are many policies en institutions that make it impossible, or at least very costly, not to do the right thing.

In addition to these forms of external commitment, there is a form of internal commitment. Conscience, the realization that it is unjust to take a free ride, that it would be unfair not to do the right thing: all motives and dispositions that we associate with morality. According to Ullman-Margalit, the function of morality is to prevent sub-optimality that could result form individually rational behavior. Morality binds the individual internally to her gun, as it were. Morality, therefore, is our answer for the failures of rationality on Ullman-Margalit’s account.

There are several problems with this functional analysis of morality. First, it is unclear whether the function of morality, is also the reason for its emergence and existence. It is one thing to claim that one of the functions of noses is to carry reading glasses. However, nothing follows from this claim when we ask why we have noses. It could be argued that this
is not a relevant question in this connection. We are interested in the justification of moral
prescriptions. Note, however, that the justification of morality is completely analogous to the
emergence of morality in this sense. It is one thing to claim that morality has certain
functions but that does not make morality justified.

A second problem with this functionalist analysis is the following. It is not true that
moral prescriptions always have optimal effects. Suppose that an impoverished person could
improve his circumstances by selling himself into slavery. The intended master would be
better off as well since she would now have additional labor. (Finally, the problem of the
laundry will be solved!) However, there are strong moral reasons against such Pareto-
efficient agreements. It is, therefore, far from obvious that morality has the function of
avoiding sub-optimal results.

Finally, this functional analysis implies an opposition between rationality and morality.
From the perspective of each individual gunner, it is rational to flee. The only thing stopping
them is something non-rational, i.e., morality. In other words, according to the functional
analysis, it is not rational to comply with moral prescriptions and it is a (happy) coincidence
when people do.

3. Bargaining theory and the social contract

These problems have led to a further integration of game theory in ethics. Instead of seeing
morality as an external constraint on rational behavior, several authors have argued that
morality should be seen as the result of individually rational behavior. Roughly speaking
there are two ways in which this idea is articulated. Some authors regard morality as the
unintended by-product of a 'quasi-evolutionary' process of interactions. I discuss this idea in
the next section. In addition, there is a tradition that sees morality as the intentional result of the interactions between ideally rational agents under ideal circumstances. This is an old idea in moral and political philosophy: it is the doctrine of the social contract. The introduction of game theory, as well as insights from decision theory, has contributed to the huge interest in this idea over the last decades. John Rawls, John Harsanyi, Richard Braithwaite, Brian Barry and much later David Gauthier have used the tools of game and decision theory to formulate their versions of the theory. They attempted to show that rational agents in an suitably idealized bargaining situation (1) will always agree on a specific distribution of the benefits of cooperation; (2) that this distribution is just; and (3) – in the case of Gauthier – that rational agents will actually comply with the terms of the bargain.

It is crucial for these theories exactly how the bargaining problem is characterized. Gauthier, as well as many others, thinks that it is a prisoner’s dilemma. That is, the predicament of the parties in the ideal bargaining position is structurally equivalent to the situation of the artillerists, as I described it above. Without any cooperation, the gunners are doomed to flee and spend the remainder of the war in captivity. Suppose that it is possible to make binding agreements in this situation. Does this solve the problem? It does not because it is not self-evident how the benefits of cooperation will be distributed. It might seem that in this case there is only one way in which these can be distributed, but appearances deceive. The artillerists could decide to follow a mixed strategy. A mixed strategy is a lottery over the available strategies of each individual. For example, the gunners could decide to flee with a probability of – say – \( \frac{1}{3} \) and stay and fight with a probability of \( \frac{2}{3} \).

---

The gunners realize that they each individually can realize at least the one but worst outcome of non-cooperation. This means that the outcome of their agreement should be at least as good as the non-cooperative outcome. Therefore, the distribution that they will agree to should at least offer them 1. Suppose that the gunners have a pair of dice. Now they can realize cooperative distributions that offer them a whole range of utilities between 1 and 3. For example, if they agree to throw both dice and if a total of six or less comes up #1 will flee (thus realizing a utility score of 3). However, if the total of both dice is more than six, #1 will stay and fight the enemy (realizing his worst outcome of 0). The expected utility of this deal for #1 is $\frac{5}{12} \cdot 3 + \frac{7}{12} \cdot 0 = 1.25$, while #2 can expect 1.75 from this deal. In this way, the gunners can realize a whole range of outcomes by varying the chances that improves on the non-cooperative outcome. These outcomes form the bargaining area (see figure 2).

![Figure 2](image)

Intuitively it seems straightforward that the outcome of the agreement between #1 and #2 should be (2, 2). However, considered formally this is anything but straightforward. Every outcome that gives each gunner an expected utility of more than 1 seems rationally acceptable. Which one will rational gunners select? Within bargaining theory, the part of
game theory that deals with these problems, there are two approaches to answer this question. First, there is the traditional axiomatic approach. On this approach, the theorist considers the bargaining area and asks herself which outcome(s) satisfy a number of reasonable requirements of a rational outcome of the negotiations. For example, things such as the names of the parties concerned should not matter for the result, whereas their preferences should matter. This approach has been very influential in game-theoretic social contract theory. Rawls, Harsanyi, Barry and Gauthier all have used axiomatic approaches to justify their favorite version. Their verdict in the case of the gunners is the same: the rational thing to agree to is a distribution that gives each gunner an expected utility of 2. (Note that this verdict does not tell the gunners how they should realize this outcome. There are two ways in which they could secure an expected outcome of (2, 2). They could both stay and fight or they could flip a fair coin to decide who gets to stay and who is allowed to flee.)

The axiomatic approach pays no attention to the structure of the process of negotiation. All it requires as input is information about the pay-offs of the parties. Whereas it is true that sometimes it does not matter how the negotiation process is structured, sometimes it is very important. For example, if it is the case that #1 can make a claim and all #2 can do is to accept or refuse, #1 does best by offering #2 an expected utility of 1.00001 and claim 2.99999 for himself. Given the rules of the negotiation process #2 will have to accept this since the alternative is (slightly) worse. On the other hand, if the rules allow for exchanges of claims and offers the situation is quite different. Therefore, if you want to predict what the result of the negotiation process between rational agents will be, it is crucial to know the rules of negotiation in detail as well as the bargaining area. In addition, it is important to know whether the parties will keep to the agreement. For if this is not the case,

---

it is unlikely that the parties concerned will accept the agreement instead of an agreement that will turn out to be binding.

Therefore, it is better to think of the bargaining process as a series of possible moves in a game that precede the game that the gunners face. This is the second approach, which regards bargaining processes as non-cooperative games. The solution to such a game then corresponds to the solution to the bargaining process. On this approach, one needs to pay a lot of attention to detail. Consequently, the analysis of such is complicated and often messy. This is another reason why the axiomatic approach was so attractive for moral philosophers. However, it is very well possible that the solution to the game and the solution based on the axiomatic approach are identical. In fact, this is what you would expect if the axiomatic solution is any good as a rational solution. This intuition is the driving force of the so-called Nash program. This program aims at evaluating axiomatic solutions by checking whether the outcome of a negotiation game leads to the same outcome. The success of the Nash program is crucial for the plausibility of the classic axiomatic theories of the social contract. For if you regard morality as the result of (hypothetical) negotiations between ideally rational agents then you have a fundamental result. It ensures the moral philosopher not only that there is a unique rational outcome of the negotiations, it also tells the philosopher what the result will be.

Does this answer the objections to the functional approach? It does so in part only. Let us start with the good news. Our first objection against the functional approach was that there is no connection between the function of morality and the justification of morality. The game-theoretic theories of the social contract can meet this objection as follows. According to these theories, moral rules are justified if and only if one can regard them as
the outcome of a rational process of negotiations between rational agents. Since all these theories assume that the result of such processes will be Pareto-efficient, the alleged function of morality – ensuring efficient outcomes is part of their justification. If a norm is not efficient, it cannot be justified. Function and justification are the same on this approach. This is surely a huge improvement on the functional approach.

However, the good news ends there. The second objection against the functionalist approach was that some basic moral norms often are not necessarily Pareto-efficient. This is clearly incompatible with the basic premise of this approach. (Incidentally, this is one of the reasons why the game-theoretic social contract theories often are revisionist.) The objections against this type of theory are not limited to this. For instance, there is a standard objection against most, if not all, contract theory that it is unclear why the agreements between ideally rational agents would tell us anything about the morality of us, not-so-ideally-rational creatures. The plurality of solution concepts is a bad omen in this regard. In addition, there are reasons to doubt that the game-theoretic approach to bargaining can really help us predict the outcome of the negotiations of rational agents. Both the axiomatic approach and the non-cooperative game approach proceed from the assumption that there is a unique, rational outcome of such negotiations. While that may be plausible in some situations, it is far from obvious that this is always the case. That is, the outcome of negotiations often is rationally underdetermined. Non-rational factors, such as salience, precedence, etc., are far more

---

important for determining the result of such negotiations than standard bargaining and game theory lead us to believe.

In addition to this specific problem, there still is a very general worry with the game-theoretic treatment of the social contract tradition. If it is correct that the basic bargaining problem is a prisoner’s dilemma, it is hard to understand how the result of the negotiations could be binding for rational agents. Even if we could predict what the gunners would agree to if they were ideally rational, we cannot demonstrate that they would stick to their agreement given their rationality. Therefore, our last objection against the functional approach – the opposition between rationality and morality – is not met in this approach.

4. Evolutionary game theory and ethics

This brings us to the third way in which game theory has been applied in moral philosophy. This is the evolutionary approach.\(^\text{10}\) The term ‘evolution’ is unfortunate for two reasons. First, because it suggests that morality should be seen as a biological phenomenon. However, that would be too strong. The term ‘evolution’ should be understood as the characterization of the process of social change through which a norm emerges in the behavior of a group or population. Secondly, it is often unclear why the process of social change, as it is often described in this literature, is an evolutionary process. Some argue that a process of social change is evolutionary if the agents regard their environment as fixed and adapt to this environment. Agents adapting their behavior in this way commit a parametric fallacy: they fail to realize that others respond to the environment as well and thus cause the environment to change. If this is the correct way to understand the term ‘evolution’ then the example we
discuss in the next sections is not truly evolutionary, but ‘quasi-evolutionary’ at best. This poses some difficulties for the methodological interpretation of this approach, but I will not discuss those here.

Within the (quasi-) evolutionary approach morality is the result of the interactions of rational individuals, just like the social contract tradition. However, contrary to the social contract tradition there are no negotiations, whether real, implicit or hypothetical. Morality is the unintended by-product of social interaction. One of the most important results in the ever-increasing literature on the evolution of cooperation and altruism is that very sophisticated forms of solidarity can develop in groups of self-interested individuals.

However, the introduction of evolutionary game theory did not only lead to a new understanding as to how a complex system of moral norms could emerge, it also gave a new way of understanding traditional questions within moral philosophy. In what follows I give one example of this development based on work by Robert Sugden, Brian Skyrms, and myself.\textsuperscript{11}

5. The dilemma of moral authority

One of the central problems in moral philosophy is the question how it can be the case that a moral norm imposes obligations on an agent, such that she has to obey these. Obedience is not the same thing as conformity. One can conform to a norm for all kinds of reasons. I conform to the norm against speeding at the moment I type these words. However, this has

nothing to do with the norm against speeding. I am writing this article and speeding does not seem a particular good way to get this task done. Similarly, I would be conforming, rather than complying, with this norm if fear of consequences stops us from speeding. The real possibility of fines or prison sentences, social ostracism, etc., etc., can all be reasons for me not to speed beyond the legal limit. In none of these cases do I comply with the norm against speeding. Only if one abstains from speeding because it is the norm, one complies with the norm. Another way of expressing this is to say that norms have authority.

How is this possible? How could it be that norms – at least sometimes – have authority? I can illustrate this problem as a dilemma. An action is rational if there are good reasons for the action. A rational person will therefore act on such reasons. Suppose then that there is a norm that prescribes an action A. If there are good reasons for A a rational person will do A. However, in that case the fact that there is a norm which tells the agent to A is irrelevant for the agent. It does not make any difference for the agent. Suppose that there is a norm that forbids this rational action A. Then acting as the norm requires will be irrational: it goes against the balance of reasons. Therefore, norms are irrelevant or irrational. One can avoid this dilemma only if one can show that norms in and of themselves can give reasons for action. That is, the dilemma can be avoided only if norms can have authority. One of the attractive and exciting features of the (quasi-) evolutionary approach is that it can show that norms are both rational and relevant to the rational agent.


6. Property and convention

I will illustrate this solution to the dilemma using the example of a very simple property norm. Property can be understood as the distribution of scarce resources. The fact that resources are scarce tends to generate conflict between individual agents. Suppose that there are two individual agents, 1 and 2, who both want to use a resource. Both claim the resource in question. Each agent can choose one of two strategies. They can decide to insist on their claim and if necessary fight the other. Alternatively, they can give in to the claim of the other. Their conflict is not an all-or-nothing affair. It is not a zero-sum game. Each agent has the following ordering of the possible outcomes. The best possible outcome for me is that you give in to my insistence and leave the resource to me. That way we avoid a costly fight and I have the entire resource at my disposal (expected utility 3). Second best is the situation in which we are both prepared to give in. In that case we will share the resource (expected utility 2). Third best is where I give in to your insistence. I will not have access to the resource, but a costly fight has been avoided (expected utility 1). Absolutely disastrous is the outcome where we both insist and fight. The damage that we are likely to inflict on each other is such that neither of us will be able to use the good (expected utility 0). If we combine all this information in a 2 x 2 matrix, we get the so-called hawk-dove game.\textsuperscript{13}

Figure 3

What should 1 and 2 do? Suppose that 2 decides to give in. In that case, the best 1 can do is to insist on his claim. If 1 insists, 2 does best to give in. This outcome is in equilibrium: it is an outcome where none of the agents can do better given the choices of the other agent. Note that in this situation, the reasons for 1 to insist depend on the reasons of 2 to give in. 1 has a reason to insist only in so far as 2 has reasons to give in. However, 2’s reason to give in fully depends on 1’s reason to fight. The reasons of 1 and 2 to converge on this outcome are, therefore, interdependent. Does this mean that 1 should go ahead and insist and 2 should give in? Of course not! There is another outcome that is in equilibrium. If 1 decides to give in then the best 2 can do is to fight; and if 2 will fight, 1 is best off to give in. Again, their reasons are interdependent.

Which of these outcomes will rational agents such as 1 and 2 select? If one of them had a reason to fight or to give in that was independent of the reasons of the other, they might be able to escape their predicament. However, neither has such independent reasons. The evolutionary approach makes three additional assumptions. First, it is assumed that this kind of conflict is unique or limited to 1 and 2 only. There are many agents like 1 and 2 and there are many occasions where there is a potential conflict over scarce resources. In the jargon of

<table>
<thead>
<tr>
<th></th>
<th>insist and if necessary fight</th>
</tr>
</thead>
<tbody>
<tr>
<td>give in</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
evolutionary game theory, it is a repeated game in which anonymous agents from a finite population are randomly paired in each round. If this were the only additional assumption we would make, the agents would not be any closer to a solution. Now 1 and 2 cannot differentiate between the upper right outcome and the lower left outcome. The information that makes this possible is excluded from the formal description of the game. In such a population, one can expect that the agents will end up fighting in a substantial number of cases rather than respecting the claims of one of them.

Here is why. Suppose the population consists of 100 individuals. Each decides to follow a strategy in the encounters with the other members of the group. Some will be prepared to give in; others will want to fight at all times. Some will fight sometimes and give in at other times. Let $p$ be the chance with which a typical individual in this population will insist. Suppose that 50% of the agents always fights, 20% always gives in and the remaining 30% fights in $2/3$ of her encounters. In this population, $p$ will equal 0.7. Therefore, any agent in this population can expect that the other will be prepared to fight with a probability of 0.7. If one would decide to fight, one can expect a pay-off of 1. Suppose that one decides to give in, one can expect a pay-off of 1.3. Those who decide to give in, will do slightly better than those who decide to fight in this population.

At this point, the second, crucial, assumption of the evolutionary approach plays an important role. The assumption is that this relatively larger pay-off of those that give in has consequences for the composition of the population. Under biological evolutionary selection, this is because successful individuals will have relatively more descendants with similar traits. In social evolution another feedback mechanism is postulated. There the typical assumption is that relatively less successful agents will be disposed to copy the more successful strategies. With this assumption in place, one can expect that the fraction of
agents who will give in in this population will increase. Thus the value of \( p \) (the probability that a typical individual in this population will fight) will decrease. However, \( p \) will not approach 0, because at some point the chance that the other will give in will become so high that those who insist and fight will be more successful. If \( p = 0.5 \) the attractiveness of insisting and giving in will be in equilibrium. Such a population will be stable. (I omit the formal proof to this effect).

In such a population one cannot maintain that the agents follow a norm. Sometimes individuals will try to solve their conflict by force, sometimes by sharing the resource and sometimes, rather unpredictably, one of the two contestants will respect the claim of the other agent. There is no systematic way in which the claims of these individuals are coordinated. How could a norm emerge in such a situation?

At this point, I have to introduce the third assumption of the evolutionary approach. Standard game theory assumes that the knowledge of the players is limited to the possible strategies of each player, the possible outcomes and the utility that each individual agent attaches to them. The evolutionary approach assumes that the agents have common knowledge of things that are strictly speaking not part of the mathematical description of the game. For example, the agents could observe who is the first to claim the resource. This makes it possible to follow an asymmetrical strategy. Suppose in this population there is an individual agent who starts to follow the strategy ‘if I am the first to claim the resource, I will insist and – if necessary – fight; if not, I'll give in’. Such an agent is a mutant in the jargon of evolutionary game theory. The expected pay-off for this mutant for an encounter with an agent who follows the symmetrical strategy is 1.5. Suppose, however, that there is one similar mutant who will fight if she is the first to claim the resource and give in if she comes second. If these two mutants encounter each other, their expected pay-off is 2.5. The average pay-off
per encounter for an agent in this population who follows a symmetrical strategy is 1.5. The average pay-off for our mutant, however, is 1.505. This mutant will do slightly better on average than those agents who follow a symmetrical strategy. If there are agents notice this increase in pay-off, we can expect that they will copy this successful strategy. As the number of agents following this strategy in the population increases, the relative success will increase. If 10% of the population follows the asymmetrical strategy, their expected pay-off will be 1.55. It is therefore safe to assume that because of this success more and more agents in this population will copy this strategy at future occasions, until the entire population has switched to this strategy.

This asymmetrical strategy is stable. Once it is there, it is best for each individual agent to follow it. There are no mutations possible that would fare better in this population. Note that this is not the only possible strategy. The population could have settled on the strategy ‘if one is first, give in; if second, insist’. A stable strategy where there are more stable strategies possible is a convention. This does not mean that this strategy is arbitrary; all it means is that because others follow it, it is rational to do so as well.

The suggestion is that moral norms are conventions in this technical sense. Conventions like this can be found in all kinds of interaction patterns, not just in hawk-dove games like this. It depends on the nature of the situation which game model is adequate. If in such an adequate game a convention could emerge that is sufficiently like the moral norms that seem appropriate for such situations, this suggestion is plausible. In the example we gave above it seems that there are many norms of property that are like this ‘first come, first served’ rule. Norms of original acquisition, norms in patent law as well as many norms in international law are like this convention.
7. Interdependent reasons

The conventionalist model for property norms that I described above is a typical example of the evolutionary game-theoretic approach to morality. On this view, norms are stable patterns of interdependent expectations of behavior. Suppose that an agent in the situations described by the model in section 6 is the first to claim a resource. Since the convention is to give in to such claims, this agent has good reasons to expect that others will respect her claim. Because other agents know that this is the convention and know that the first agent knows this, they can expect that this first agent will insist and is prepared to fight. Therefore, the best all other agents can do is to give in. We can predict that in such populations agents will follow the norm to respect such claims.

Is this enough to avoid the dilemma described in section 5? As we saw conformity in itself is not enough to claim that a norm has authority. This claim is warranted only if the reason that agents conform is the fact that it is the norm. The conventionalist analysis of norms has shown that the reasons agents have to make such claims and respect them are interdependent. Given the structure of the situation it is not possible to determine what to do independent of the other members of the population. It depends on the accepted convention what is the rational thing to do. That determines what one can expect and, consequently, what one has reason to do. Since norms are conventions according to the evolutionary approach, we have found the way to avoid the dilemma. It is rational to follow the convention, the norm, and the reason this is so, is that it is the convention. It is, therefore, the norm that determines what a rational agent should do in this population. In other words, norms are both rational and relevant. This is an important result, because it provides an analysis in the authority of moral norms. The evolutionary game-theoretic
approach to moral norms shows that the authority of morality is not something mysterious or *sui generis*. The approach makes further conceptual analysis of this authority possible.

8. Critical remarks

Is there only good news? Is the introduction of game theory in ethics a complete success story? I am convinced of the usefulness of the approach, but I am reluctant to follow Braithwaite’s suggestion and claim a complete victory. I will mention some points of critique.

First, there has been a lot of criticism of the idea that our moral norms are conventions. This criticism has two objects. First, there is the criticism that ‘real’ moral norms are not arbitrary, contrary to the suggestion of the theory. It is conventional to drive on the right side of the road in the USA. It could have been the left side of the road; it just so happens that we do not. However, so the critic claims, there is no similar arbitrariness when it comes down to our most fundamental norms? Take the norm against killing innocent people. In what sense is this norm arbitrary? What is the alternative to this norm that we can intuitively agree could just as well be the accepted conventions? In short, on this line of criticism our most fundamental basic norms are not conventional at all.

This particular criticism is not very relevant. First, it is simply not true that the established convention is arbitrary on the evolutionary game-theoretic approach. Although more than one stable pattern of interdependent expectations could have emerged, it is not the case that just any pattern could have emerged. There are clear limits to the range of possibilities. Therefore, the model itself puts constraints on possible norms.

Secondly, if we look at existing ‘basic’ moral norms, we discover that many of their components are conventional. Take the norm against murder. Murder is the intentional
killing of innocent people. Who exactly fall under the extension of ‘innocent people’? In
most Western societies adulterous spouses are included in this category, but in other cultures
they are excluded. There it is open to the cheated spouses or their relatives to kill the
adulterers. The same holds for the extension of ‘killing’. Does the omission to provide
cytostates to an elderly cancer patient at his request amount to killing him or not? So even if
one insists that the basic moral norm against murder is not conventional, there are many
conventional elements in its interpretation and application. Consequently, the claim that it is
the same norm under all these circumstances is dubious at best.

However, the most powerful argument against this type of criticism we have already
indicated. This type of criticism equivocates our everyday term of ‘conventional’ in the sense
of ‘morally neutral’ with ‘conventional’ in the technical sense that we identified in the
previous section. Norms are only ‘conventional’ in the latter sense: they are stable patterns of
interdependent behavioral expectations.

This brings us to the second type of objection to the evolutionary game-theoretic
analysis of moral norms. According to the analysis, there is a moral norm when all, or a
sufficiently large number of agents behave in a certain way and mutually expects this
behavior of each other, where both the behavior and the expectations are stable. Suppose
that everybody is prepared to kill innocent human beings. Suppose moreover that everybody
knows this to be the case, expects this and that the practice and the expectations are stable.
Does this make justify killing innocent people? ‘Of course not!’ the critics claim. The
evolutionary game-theoretic approach looks in the wrong place for the justification of a
norm. If we want to determine whether it is justifiable to kill innocent human beings, we
need to look at things other than the behavior of people or their expectations.
This criticism is less impressive than it might seem at first sight. First, we should make a distinction between the question whether a particular social practice (e.g., murder) is morally permissible and the question how an individual agent who finds herself within such a practice should act. It is very well possible that it is morally permissible (and sometimes even required) to comply with morally objectionable norms. Suppose that you do find yourself in Hobbes’ state of nature. Everybody around you is prepared to resort to death and destruction in the pursuit of their goals. It is hypocritical to insist under such circumstances that the agents should abstain from these things. It would be too demanding. The situation is regrettable and it is required to try to change it, but one cannot be required to be the only one to comply with a norm to refrain from murder.

Against this line of reasoning, it might be argued that the norm against murder is insufficiently specified. It seems as if this norm applies unconditionally, but perhaps it applies only when a sufficient number of people comply with it. In other words, the correct characterization of this norm is not ‘under no circumstance you may kill an innocent person’, but ‘do not kill an innocent person if \(x\%\) of the population does not kill innocent persons’. This norm is still valid in the (hopefully) hypothetical situation where nobody abstains from murder. If this is a consistent response then the reason that this norm is valid is not determined by the fact whether or not it is applicable. Then the norm against murder is not based on a stable social practice, but it is based on fundamental moral considerations.

This is an important criticism of the evolutionary game-theoretic approach to moral norms. However, it does not show that this approach is completely mistaken. It does show that in extreme and rare cases certain norms have authority on other reasons than those that are generated by stable patterns of interaction. If we give up the notion that game theory provides a ‘theory of everything’ in ethics, we have a sober but correct view of this theory. It is a powerful instrument in the analysis of moral phenomena and concepts. However, morality is so diverse that it is unlikely that there is one correct universal moral theory, whether this is a game theoretic one or not.