Artificial Intelligence and Law **2:** 51–63, 1994.

© 1994 *Kluwer Academic Publishers. Printed in the Netherlands.*

Artificial Morality and Artificial Law

LOTHAR PHILIPPS
Institute of Philosophy of Law and Computers and Law, University of Munich, Germany

Key words: artificial morality, prisoner's dilemma, game of Chicken, contractual situation, competitive situation, common good, conditional cooperator, reciprocal cooperator, level of reflection, genetic algorithm.

Abstract. The article investigates the interplay of moral rules in computer simulation. The investigation is based on two situations which are well-known to game theory: the prisoner's dilemma and the game of Chicken. The prisoner's dilemma can be taken to represent contractual situations, the game of Chicken represents a competitive situation on the one hand and the provision for a common good on the other. Unlike the rules usually used in game theory, each player knows the other's strategy. In that way, ever higher levels of reflection are reached reciprocally. Such strategies can be interpreted as 'moral' rules.

Artificial morality is related to the discipline of 'Artificial Life'. As in artificial life, the use of genetic algorithms suggests itself. Rules of behaviour split and reunite as chromosome strings do.

1. Introduction

Artificial Intelligence and the law: this combination suggests a 'liaison': Artificial Law [Philipps 1989]. The discipline of Artificial Life exists already [Levy 1992]: artificial life forms, simulated by computer, adapt to an artificial environment, struggle for life, diversify, combine their natural assets and evolve. Could not the same be possible for morality and the law? Rules of behaviour, which guide imaginary people, struggle, cooperate, diversify and unify.

The Canadian philosopher Peter Danielson has recently published a book in which he explores these possibilities using computer programs: 'Artificial Morality – Virtuous Robots for Virtual Games' [Danielson 1992]. Some of the principles investigated by Danielson already extend into the realm of legal discourse.

2. The Prisoner's Dilemma and Contractual Situations

One of the most important insights of modern moral philosophy is that many contractual relations share the structure of the prisoner's dilemma. This insight is linked to the use of the computer as an instrument of philosophy.

The prisoner's dilemma describes a scene set in the USA: After a robbery, two vagabonds are apprehended near the site of the crime. The sheriff is convinced that he has caught the perpetrators but he cannot prove it. He locks the suspects in separate cells and makes clear to them their situation:

1. Should one of them plead guilty, but not the other, the one who confesses will be released for giving State's evidence. The other will face a long term of imprisonment.

2. Should both plead guilty, there will be no need for State's evidence. Both will be sentenced to prison, but only to medium terms, since the confessions will constitute mitigating circumstances.
3. If neither of them should plead guilty, the court will have no choice but to sentence both to only short terms of imprisonment – for vagrancy.

The prisoner's situation can be translated to a game theory matrix (a high number does not imply a long term in prison, on the contrary: the higher one's number the better one's situation):

	Denial (cooperation)	Confession (defection)
Denial (cooperation)	2,2	0,3
Confession (defection)	3,0	1,1

This matrix shows what is likely to happen: both will confess. For each of them will say to himself: If my partner should confess, it will be better for me to confess as well, or my prison term will be long. If the other does not confess I will profit all the more: I will be released.

In terms of game theory this means that the strategy of confession dominates denial. Or perhaps the cautious maximin principle was applied. It states that the course of action should be chosen which offers in its worst case the comparatively best result. The worst result of confession is a medium prison term; in the case of denial it is a long prison term.

It is remarkable that the rational course of action for each individual prisoner is not prudent for both together. It would be better for both to keep silent; that would mean only short imprisonment for both.

The fundamental philosophical importance of the prisoner's dilemma has long been realized. According to my knowledge, it was Canadian philosopher David Gauthier who first called attention to the fact that the situation is the same for contracts [Gauthier 1969]. This is definitely true for contracts in the 'State of Nature' where state power is not available for enforcement. Each party might reason this way: I would like to perform my part of the deal, but how can I be sure that the other will do the same? After all, he will consider the possibility that I might not perform. Therefore, to minimize the potential damage, neither will honour the agreement. This is true not only for outright breach of contract, but also for insufficient performance, which seems even more realistic.

Such contracts in a state of nature still exist in our present society. For example plea bargaining, the deal for the sentence struck between judge, defence, and state's attorney. In Germany plea bargaining is not considered permissable but takes place again and again [Schünemann 1990]. It is possible for the judge to ignore the deal, or for a defendant to make a false confession and incriminate an innocent third party.

We all know the legal instruments of enforcement for contractual obligations. But perhaps many people honour their agreements regardless of the state's threatening shadow. This is the case in long standing business relations, which can be simulated if the prisoner's dilemma occurs repeatedly ('iteratedly'). Each party performs truthfully for fear that a lucrative connection might break.

The American political scientist Robert Axelrod challenged scientists from all over the world to a tournament – political scientists, psychologists, biologists, and game theoreticians [Axelrod 1984]. Each contestant developed a computer program to master the iterated prisoner's dilemma. Each program was matched against every other program, including itself, and also against a program that executed random moves. The remarkable outcome of two tournaments was that each time the most simple program won: TIT FOR TAT. This program starts by cooperating and then continues by imitating its opponent's move. Should the opponent cooperate, it will keep cooperating. If its partner defects, it will also defect – until the partner once again switches to cooperation: then it will follow suit immediately.

The principle of this success is easily understood by comparing TIT FOR TAT to the behaviour of a program which only acts defectively (which is individually rational in a single prisoner's dilemma). If both programs play against each other, TIT FOR TAT will lose due to its initial show of trust. This defeat, however, only occurs in the first round and does not bring the opponent many points. TIT FOR TAT on the other hand will accumulate point after point when playing against itself. The morally satisfying result of this is the development of a stable population of TIT FOR TAT players, from which notorious defectors will be excluded.[1]

The principle is plausible; but still it is remarkable that no other program, no matter how shrewd or sophisticated, has managed to outwit TIT FOR TAT. Peter Danielson admits that he spent the 'better part of a weekend' developing a strategy he was sure would defeat TIT FOR TAT [Danielson 1992, p. 47]. It proved an illusion. This seems to show the truth in a statement by anthropologist Levi-Strauss: That the law of reciprocity is as fundamental to social interaction as the law of gravity is to physics.

In fact, the phenomenon of reciprocity is also encountered among animals. Biologists distinguish two variants of animal altruism [Danielson 1992, pp. 39–51]: Kinship altruism signifies that individual advantages are sacrifices to proliferate the genes. The other variant is reciprocal altruism: present gain is given up in favour of future advantages from longstanding relations.

This reciprocal altruism will only work if the partner is recognized, but there are instructive examples for surrogates for recognition: In their 'home port', some large predatory fish have their teeth cleaned by small fish (which provides sustenance to these). Outside the 'home port', the predatory fish would devour its friends because it does not recognize them. Both species of fish cannot recognize each other; but they recognize the location in which a peaceful and useful encounter takes place.

[1] Those who are glad about the results of Axelrod's tournaments should remember two things. (1) Besides the prisoner's dilemma there is the game of Chicken which is also of fundamental philosophical importance but – as will be seen – renders morally dubious results. (2) In Axelrod's tournaments players meet in changing pairs of two players. In n-person prisoner's dilemma results might be completely different, especially if participants remain anonymous. While parasitical behaviour is bound to fail in the long run in a two-player game, it seems plausible in a n-player game that single parasites will attach to cooperative groups. [Schüßler 1990] has softened Axelrod's rigid rules and performed such computer simulations. However, he has reached remarkably 'positive' results.

3. The Prisoner's Dilemma as a Game of Moral Philosophy

Should the behavioural patterns I have described be called moral? Perhaps one could speak of quasi-morality or proto-morality for animal behaviour, and 'artificial morality' for computer programs. Peter Danielson, the author of Artificial Morality, is against these labels: All these phenomena are finally mechanisms of selfishness, and nothing more. If the prisoner's dilemma is appropriate for the description of moral behaviour, it must prove so in a single dilemma situation where no future gain is on the horizon. I would like to follow Danielson's philosophic approach but not his taste in terminology. 'Artificial morality' – analogous to 'artificial life' – is by far too elegant and general a term to be confined to a limited meaning. Selfishness and morality are connected in so many areas that an encompassing term is helpful. The principle of reciprocity is one of the biological roots of morality and it would be wrong to exclude it from moral discourses, despite the fact that morality in a modern sense can no longer be reduced to such origins.

Danielson's 'moral prisoner's dilemma' differs from the standard prisoner's dilemma in that the participants know their counterpart's strategy, and are thus able to predict their moves. This might seem to take the interest out of it in terms of game theory; but in terms of philosophy it does not. Let us assume that I perceive my partner to be cooperative. As an egoist I would defect in order to profit the most. As an altruist I would probably cooperate. It is plausible that my actions might be considered 'immoral' in the first instance, but 'moral' in the other. Of course it is not very likely that my partner, perceiving my egoistical strategy will continue to cooperate, but it is not impossible – and moral values are rarely a matter of likelihood.

Danielson has written PROLOG programs allowing interactive computers to recognize their strategies. Within this given context – single encounters with mutual recognition – strategies evolve according to Darwinian principles. The results are differentiated and surprising.

Perhaps the most plausible strategy is that of a 'conditional cooperator' introduced by David Gauthier. The conditional cooperator will cooperate only with those who will cooperate themselves. He or she refuses to cooperate with 'straightforward maximizers' who will exploit a partner's willingness to cooperate. By seeking his own profit, the conditional cooperator also furthers the common good. Gauthier believed to have thus found a point at which rationality and morality coincide.

Danielson, however, found a rationality gap in Gauthier's morality. Why, says Danielson, not exploit those partners who will always and unconditionally cooperate? Reason only demands cooperating with those who will cooperate conditionally only – under the condition that they in turn will be cooperated with. Thus Danielson introduced the 'reciprocal cooperator'. An example might show the difference: A reciprocally cooperative merchant would only sell at an adequate price if the buyer bargains for it; a conditionally cooperative merchant will always sell at the adequate price (but not give away goods, of course).

Gauthier has replied to Danielson that his reciprocal cooperator makes a 'moral monster'. Danielson has not withdrawn to a strictly rational and formal position, but has followed Gauthier onto the field of substantive arguments [Danielson 1992, pp. 61–123]. He has suggested an evolutionary test in which an influence of Axelrod's investigations can be felt. What would a society be like in which Gauthier's conditional cooperators are represented? These would not only support each other but also those good-natured members who will cooperate in any case ('unconditional cooperators'). That in itself is not bad, but by supporting unconditional cooperators the immoral straightforward maximizers will indirectly be supported as well, since they naturally prey on unconditional cooperators. If on the other hand a position of reciprocal cooperation is adapted, not only unconditional cooperators, but also straightforward maximizers will be pushed out of the society. An example of this process [Gauthier 1988]: A farmer kills all the rabbits in his fields to destroy the foxes' primary source of food. Translated to practical rules this could mean: Eliminate the weak in order take away from the evil their natural prey! More specifically: Eliminate the careless to diminish the profiteers' source of income! Or more pertinent: Get rid of the asylum-seekers to remove the neo-Nazis' primary target of attack!

Interestingly enough, a very similar game concept and closely related argumentation can lead to a different point of view. In the late 18th century, Jeremy Bentham wrote 'In Defence of Ursury.' Bentham argues that the existence of ursury is beneficial to a society. Ursurers are pikes among carps: They shake people up from their sluggishness and create a sense of self-responsibility in the population. The book was very influential, also in Germany. The liberals succeeded in abolishing criminal punishment for ursury. Thus, our criminal code (Strafgesetzbuch) did not mention ursury in its original 1871 version. Soon afterward this changed, however.

Bentham's argument is structured like a theodicy: the Evil is considered a tool toward the Good. This type of argument is important today as well: many despise neo-Nazis, but are happy that they are doing the dirty work.

As soon as abstract subjects are no longer used, or rabbits and foxes for examples, as soon as real people in real roles and situations are pictured, it becomes apparent that this attempt to deduce morality from rationality has failed. What would we think of a State, a law-giving body, which willingly or unwillingly considered itself bound by principles of pure reason? It would have given up as a moral entity.

On the other hand we see that Danielson's investigations are not an empty intellectual game, but something to be taken seriously. We realize that the argumentational patterns in question are realistic and that they are used again and again in arguments on morality, law or politics.

4. The Game of Chicken – Competition and Common Goods

For a long time, the prisoner's dilemma was considered the only game type which is fundamental to law and morality. Danielson is right to insist, however, that the 'Chicken' type is no less important. Chicken is played among American youths in several variants; this is one: Two adolescents are speeding toward each other in cars. The one who first veers off the collision course is a chicken, and loses. Of course it is possible that both drivers chicken out, or neither of them.

Chicken differs from the prisoner's dilemma in that the worst possible result for both participants occurs after mutual 'defection' (both stay in one and the same lane). In the original game, you can expect to be dead. If you evade the other car, you may have lost face, but you are alive – the second to worst result. Conversely, mutual defection in the prisoner's dilemma leads to the second to worst result, and the worst case is suffered by the single cooperator. In books on game theory it has (under the influence of the prisoner's dilemma) become a custom, to mark the result of one-sided defection with the letter T (for temptation) and the result of one-sided cooperation S (sucker's payoff); R (reward) represents mutual cooperation and P (penalty) mutual defection. For the prisoner's dilemma the preference scale of results would be: T > R > P > S. For Chicken P and S are switched, the preference scale is: T > R > S > P. Of course the letter codes have meaning only for the prisoner's dilemma, but they are maintained in the description of other games for comparability's sake.

If the prisoner's dilemma is especially well suited to describe situations that might be governed by contracts, Chicken will fit two different types of situations. The first is a competitive situation. Two competitors can ruin each other – but the one who quits before it comes to that will be at a disadvantage.

The second situation that can be described using Chicken is the preservation of a common good. A good example is this [Taylor and Ward 1982]: Two Dutch farmers' fields lie behind a dike. When a storm tide is expected, the dike needs to be reinforced. Each farmer could work on the dike (C) or not (D). If neither of them does anything (D/D), the dike will break and the catastrophe ensue. If both put in work (C/C), the dike will hold and neither will lose much time. If only one works on the dike (D/C or C/D) the dike will also hold, but that farmer will lose much working time, while his neighbour will make money on the time he saved. There is a mutual interest in the preservation of a common good, but each private interest is in reaching this goal at the other's expense.

The fact that both competitive situations and common good provisions can be represented by the Chicken game show the game's high level of abstraction. A difference can be seen in that in competitive situations defection occurs in the form of action, whereas defection in a situation of common good provision is constituted by omission. The following example will show, however, that competition and common good are interrelated: During rush-hour the main streets of a certain city are too clogged to allow anybody to get through. This is a situation of paralyzing competition. Now an efficient public transportation system is established which leads to a common good situation. But many dodge the fares and consequently prices have to be raised. If the number of people who ride without a ticket increases, and the others will not be willing to pay the resulting high prices, the system will collapse.

It is an essential difference between the prisoner's dilemma and Chicken that the best way to induce the partner to cooperate in a prisoner's dilemma is by promises, but in a game of Chicken it is by threats. When somebody threatens to defect in a prisoner's dilemma the other will defect as well, unless he is a saint or out of his mind. But given a credible promise he will cooperate if he acts in a morally sound way. In a game of Chicken on the other hand it is rational to cooperate under a credible threat of defection from the other person. (Whether promise and threat are meant seriously or are a bluff is, of course, another question.)

A promise is a reference mainly to oneself (I will do something useful). A threat refers mainly to the other (you will suffer damage). This thought can be generalized by distinguishing four basic types of protagonists according to who cooperates with cooperators or defectors, and who defects from cooperators or defectors. The prisoner's dilemma emphazises 'ego', Chicken emphazises 'alter'. The following matrix attempts such a typology.

	Prisoner's Dilemma (ego-based types)	Chicken (alter-based types)
(C/C, C/D)	The Naive	The Soft
(C/C, D/D)	The Reliable	The Righteous
(D/C, C/D)	The Inconsistent	The Bully
(D/C, D/D)	The Cautious	The Tough

From the point of view of moral playing, Chicken is much more problematic than the prisoner's dilemma. The 'Reliable', who deserves our esteem, is replaced by the 'Righteous' who will enter an intersection on a green light even though he sees another car approaching from the side. I have mixed feelings about the Righteous. It is also disturbing that in Chicken-type games those will prosper who are soft in dealing with the tough, and tough in dealing with the soft. In a prisoner's dilemma such behaviour would be irrational, but being the 'Bully' is a successful strategy for Chicken.

Compared to the prisoner's dilemma, Chicken creates fewer types of behaviour that are moral and rational at the same time. It might be possible to conclude that the law-giver has more reason to interfere in social situations which resemble Chicken-type games, than in prisoner's dilemma-type situations. Computer simulations and computer games might be helpful in shedding some light on these questions.

5. The Make-Up of the Universe of Moral Discourse

Let us attempt to reconstruct the world of moral behaviour, by basing it on the interaction of moral rules.

On the lowest level, someone might act without reflecting, either cooperatively or defectively (C or D). Cooperative behaviour might be viewed as naively moral, defection as naively immoral.

The state of absence of reflection will not be likely to endure, especially as soon as anyone is given reason to be upset at a defective partner. He or she will develop a rule to govern his or her own behaviour. This rule will be based on the other player's expected actions. Perhaps one will decide to cooperate with those who cooperate, and defect from those who defect (Gauthier).

Since the other player has two options (to defect and to cooperate), and since one's reaction may in turn be to defect or to cooperate, four possible rules exist. (Corresponding to the four basic types of behaviour mentioned before. – In the following expressions, the second letter represents the partner's expected behaviour, while the first letter gives one's own reaction. DC, for instance, means: I will defect from cooperating partners.)

1. CC, CD
2. CC, DD
3. DC, CD
4. DC, DD

First, it is possible to still cooperate, but now upon reflection, no matter if the other defects or cooperates. Second, the rule might be to cooperate with the cooperative and defect from the defective. Third, one might – inversely – defect from the cooperative and cooperate with the defective. And fourth, one could always defect – in grim determination – no matter whether the partner cooperates or defects.

The first and second rules are clearly to be considered moral, the first one in an irrational, maybe saintly fashion, whereas the second rule of cooperating, yet not wanting to be cheated, is a rational one. As mentioned before, Gauthier has suggested this rule.

But the reflection can be taken a step further: the decision between cooperation and defection can be based on the partner's rule of behaviour, instead of being based simply upon his expected behaviour as such.

As there are four first level rules, a second level rule might look like this (a tabular representation was chosen to demonstrate the structural connection to the first level rules, which are either cooperated with or defected from):

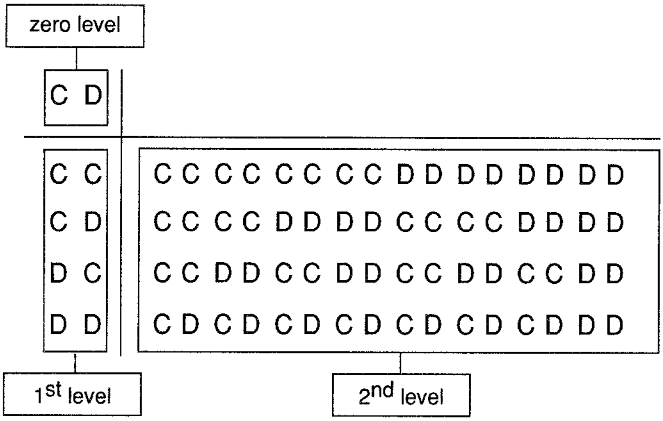
D(*CC, CD*),
C(*CC, DD*),
D(*DC, CD*),
D(*DC, DD*)

This is the controversial rule Danielson has suggested as an improvement over Gauthier's rule. (The crucial first two sections of the rule, which constitute the difference from Gauthier's rule, are printed in italics). The player cooperates only with those who would otherwise defect. As has already been explained, the clue to this solution is the attempt to take rationality a step further, to find the point where morality and rationality coincide and to thus give a rational explanation for moral behaviour.

In the lively argument between Gauthier and Danielson (The saintly and the naive who are unconditionally moral would fall victim to Danielson's moral), a structural difference has not been given much attention: Gauthier's rule is a first level rule, based simply on the other's behaviour. Danielson's rule, however, is a second level rule and thus based on the other's first level rule. Danielson's rule cannot be expressed as a first level rule, but Gauthier's rule can be written as a second level rule: C(CC, CD), C(CC, DD), D(DC, CD), D(DC, DD). Generally speaking, any rule can be expressed in terms of a higher level of reflection.

The underlying structural concept can be summed up in this way: zero level rules express simple behaviour: C or D. First level rules base behaviour on zero level rules,

TABLE 1.



C	D

C	C
C	D
D	C
D	D

C	C	C	C	C	C	C	C	D	D	D	D	D	D	D	D
C	C	C	C	D	D	D	D	C	C	C	C	D	D	D	D
C	C	D	D	C	C	D	D	C	C	D	D	C	C	D	D
C	D	C	D	C	D	C	D	C	D	C	D	C	D	D	D

that is the partner's expected simple C or D. Second level rules are accordingly based on first level rules, third level rules on second level rules and so forth. With each level of reflection, the subject of the rule changes back and forth.

Zero level rules are single expressions and there are two of them (2^1); first level rules have two terms and there are four of them (2^2); there are 16 (2^4) four-term second level rules. Then the number of combinations explodes: There are 65 536 (2^{16}) possible third level rules with sixteen terms each.

That cannot be imagined, but the four possible first level rules and maybe even the sixteen second level rules can be pictured mentally. All that is required is familiarity with the truth tables of propositional logic.

Top left the two elementary zero level expressions are given: simple cooperation and simple defection. Below the first level rules, based on the elementary behavioural patterns: four possibilities to react to cooperation or defection by cooperating or defecting. On the right, the 16 second level possibilities of reacting cooperatively or defectively to first level rules are given.[2]

The letters given in the chart only represent the top part of the rules (written on the very left in the formal expression). That to which the letters refer can be reconstructed from the chart, however. Abbreviations are used for the sake of clarity. The logical expressions of 'and', 'or', 'if... then' which are contained in the chart, can be used for mnemonic representation of the rules. For example, the first of the 16 columns means that one will cooperate on the second level of reflection, no matter what rule the partner follows. This column equals tautology in logic.

A rule of a higher level becomes manageable if it is possible to separate a few crucial terms and to group all others together. In this way Danielson's second level rule with its four terms can be written as; C(CC, DD), else D.

[2] I would like to mention a comparable concept from the fifties: the philosopher [Günther 1963] attempted to develop a logic of reflection which would oscillate between the poles 'I' and 'You' – and a third pole 'It'. His aim was to create a basis for 'conscious machines' – very much as Danielson wants to create a basis for 'moral machines'.

But can higher level rules even be considered realistic? At least for third level rules the answer is definitely 'yes', for the simple reason that it must be possible to react cooperatively or defectively to second level rules such as Danielson's. Some might reject Danielson's rule because they perceive in it a 'moral monster'. The third level reply would be: D(C(CC, DD), else D), else C[3] Such a rule seems unreasonable in that it makes such a sweeping statement, but it need not be so. It is conceivable that rules specialize as the model becomes more differentiated. Rules might for instance be limited to the repression of specific other rules, in which case they would have to be supported by other rules in order to survive. So a division of labour takes place.

Each new level of reflection can be linked to a gain in differentiation: in terms of rationality and the division of labour. For the qualities of those rules that have proven successful can be combined in a new rule. That is no different from a breeder cross-breeding specimens with superior characteristics.

In a way, Gauthier's rule CC, DD can be regarded as a successful cross between the elemental rules C and D. A genetic algorithm [Goldberg 1989] could manage this as follows: First, C and D are lifted to the first level of reflection. The new expressions can be looked upon as strings of 'chromosomes'. They can be split in two parts and rearranged again so that a cross-over takes part. Two rules which are more differentiated result: Gauthier's rule and the bully rule in the game of Chicken.

C => CC, CD =>CC... CD => CC, DD (Gauthier)

D => DC, DD =>DC... DD => DC, CD (Bully in the game of Chicken)

We can continue the process of breeding: lifting the rules (Gauthier and Bully) to the next higher level and crossing them again. If the 'string of chromosomes' is split at the first 'breaking-point' these rules will be generated: the rule of Danielson and a second one which probably will not be successful:

C..., C..., D..., D...=> C... D.., D..., => C... D.., C..., C...
(probably not successful)

D..., D..., C..., C...,=> D... D..., C..., C...=> D..., C..., D..., D...
(Danielson)

Once new rules have been designed-mechanically or intellectually-, they should be tested in competition. They should play against each other, as in Axelrod's experiments. It will become clear which rules cooperate (possibly at the expense of third parties), which intrude into populations of other rules, and which are excluded from such populations. The results are not to be left to themselves according to Darwin's principle of fitness. In my opinion, it is necessary to evaluate the rules according to moral concepts (that are not derived from the computer stimulation) and to 'breed' them. But still it is important to take care that even where the moral quality is emphasized, the rules will be sufficiently

[3] The 'morally judgmental' impression such a rejection gives is probably no coincidence: Once rules are no longer based on the other's actual behaviour, but on the other's rules (second level and up), they will seem 'moral' (or for that matter, 'immoral'). 'Moral' in this context is taken in the sense the philosophical tradition of Thomasius and Kant has given it, that 'inner' behaviour is opposed to 'the law' as an 'outer' behaviour.

robust to prevail in the 'daily hassle' and first in the computer simulation. Otherwise morality will bear the paleness of the Platonic.

The approach (combinatorial, graded in levels, genetically, and in tabular form) suggested here is, as far as I can see, new. Danielson uses a different approach: He first describes a behavioural rule informally without giving its level. In order to formalize the rules he writes a PROLOG-program. The program will run on the computer but it cannot be grasped intuitively. Our method has the advantage of transparency: both a rule's reflective level and complexity are readily apparent. Also one can see whether the rule is complete, i.e., whether it gives all actions and rules possible at the given level of reflection, and whether the options are enumerated or some given as a remainder (...else...). Besides, at first glance it can be seen whether two rules are related, and the degree of relationship can be precisely determined.

But this approach reveals another problem: the amount of undecidable cases. The impression would be that if two subjects are linked in a game situation, one must be on a higher level of reflection than the other if there should be a determinable decision. This impression is not quite correct; but it is true that the amount of undecidable cases increases with the mutual level of reflection.

However, the existence of such undecidable cases does not imply a fault in our approach, but rather problems of the matter itself. It would be wrong to try to avoid them by using an elegant computer program. An example:

Two players are facing each other using Gauthier's rule (first level of reflection):

CC, DD / CC, DD

There is no decisional problem: Each can afford to cooperate first; the other will follow suit. No problem occurs if one advances to the second level of reflection and adapts Danielson's rule:

C(CC,DD), else D / CC, DD.

Formally the left player is forced to cooperate. But also in substance: he need not fear to cooperate first because he knows that the right player will cooperate.

But what happens if two players meet on the second, on Danielson's level?

C(CC, DD), else D / C (CC, DD), else D.

Neither player can afford to cooperate first (an advance payment) because the other would exploit it. The other's readiness to exploit justifies distrust.

This leads us back to the question of morality and rationality. Can we really say the readiness to exploit, turning into justified distrust, is more rational than to consciously miss some chances for exploitation and thereby establish trust? I believe that it is not possible to match rationality and morality once and for all (as Danielson would like to do) because it is always possible to withdraw to a higher level of reflection where both will diverge again. If this hypothesis of only transitory congruity of morality and rationality could be formally proven, that would be an important philosophical insight.

So we are not only faced with the formal problem of undecidability, but also with the substantive problem of rationality. And again we see that the assumption that Danielson's rule may be immoral, but in any case more rational than Gauthier's is not beyond doubt, at least not in general. There are several good reasons to support this claim; I will only mention one of them: Bargaining leads to inertia and time loss.

Max Weber pointed out that the introduction of mass department stores in the USA has given strong leverage to modern capitalism. Buyers could rely on the fact that goods would be offered at the lowest feasible price. There was neither reason nor room for bargaining. For our purposes it is important to note that, according to Max Weber, the reasons for this remarkably rational and successful invention were originally of a religious and moral nature.

Another example, but this time taken from Chicken, not from the prisoner's dilemma: Two bullies are facing each other:

CD, DC / CD, DC.

In this situation the first level of reflection already fails to provide determination. Each bully makes his action contingent on the other's action: he will be soft with the tough and tough with the soft – but whether the other bully is soft or tough remains open.

But only until one bully rises to the next level of reflection. A 'second degree bully' will be soft with the tough and tough with the soft, but in addition he will now be tough with a simple fellow bully (and soft, by the way, with a 'righteous' (the second term in the parentheses) who does not yield to another's bullying).

D(CC, CD), C(CC, DD), D(DC, CD), C(DC, DD).

If the other bully should follow the first onto the second level of reflection, a new stalemate ensues.

Remarkably the second degree bully is guided by nearly the same rule as the 'reciprocal cooperator' (Danielson's moral subject) in the prisoner's dilemma; only the last terms of the rules are different: D..., C..., D..., C... (2nd degree bully), D..., C..., D..., D... (Danielson). This resemblance can be explained 'genetically', as we have seen before.

Perhaps it would make sense to resolve such stalemates using general principles of law, moral, or common sense as default rules. The first such principle to suggest itself would be one of generalisation, perhaps Kant's categorical imperative. However, to use this principle would take extensive operations research, which on the other hand might be supplied by Artificial Morality techniques. The categorical imperative is by no means as clear as it seems plausible.

If, for instance, the categorical imperative were to be interpreted as the demand of practical consistency (a view which is suggested by Kant at times), the question would be: What makes it less consistent for two defective prisoners to have to spend medium terms in prison, while cooperative prisoners spend short terms in prison? It is less pleasant – that is all.

Remarkably, we have here the game theoretical version of an argument which Hegel used to refute Kant: 'And if there were no deposit – what contradiction would that constitute?' (Kant had argued that if every trustee would embezzle the money entrusted to him, the institute of deposit could not exist.)

In the face of such difficulties, I suggest beginning with the morally ambiguous but easily formalized principle which is typical for professional ethics: 'Birds of a feather flock together!' (in German more drastic: 'Eine Krähe hackt der anderen kein Auge aus!'). Such a default rule would not be unrealistic even with mostly defective players. After all, solidarity is not only found among the decent, but even crooks have some honour. In a more restricted way, such a reciprocity rule has already been proposed by Danielson [1992, pp. 79–81].

The default rule would thus be that, if two players are facing each other under the same rule of behaviour, and their decision is not determined by this rule, they should cooperate.

Once this simple default rule is tested, we can go on from there....

References

Axelrod, R. 1984. *The Evolution of Cooperation.* New York: Basic Books

Danielson, P. 1992. *Artificial Mortality: Virtuous Robots for Virtual Games* London: Routledge

Gauthier, D. P. 1969. *The Logic of Leviathan: the Moral and Political theory of Thomas Hobbes.* Oxford: University Press

Gauthier, D. P. 1988. Moral Artifice, *Canadian Journal of Philosophy* 18: 385–418.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*, 2nd ed., New York: Eddison-Wesley

Günther, G. 1963. *Das Bewußtsein der Maschinen. Baden-Baden*: Agis.

Levy, St. 1992. *Artificial Life: The Quest for a New Creation.* New York: Pantheon

Philipps, L. 1989. Gibt es ein Recht auch für ein Volk von künstlichen Wesen, wenn sie nur Verstand haben? In *Jenseits des Funktionalismus: Arthur Kaufmann zum 65. Geburtstag*, ed. L. Phillipps & H. Scholler, 119–126. Heidelberg: Decker & Müller.

Schünemann, B. 1990. *Absprachen im Strafverfahren? Grundlagen, Gegenstände und Grenzen. Gutachten zum 58. Deutschen Juristentag.* München: C.H. Beck

Schüßler, R. 1990. *Kooperation unter Egoisten: Vier Dilemmata.* München: Oldenbourg

M. Taylor, H & Ward, H. 1982. Chickens, Whales and Lumpy Goods: Alternative Models of Public-Good Provision, *Political Studies* 30:350–370.