Int J Soc Robot (2009) 1: 223–232
DOI 10.1007/s12369-009-0025-3

ORIGINAL PAPER

Evolutionary Ethics in Agent Societies

Pieter Spronck · Berend Berendsen

Accepted: 20 May 2009 / Published online: 2 June 2009

Abstract To investigate whether the theory of evolutionary ethics is a valid explanation for the existence of ethical behaviour, we approach the ethical issues of morality from a computational point-of-view. We define a model of multi-agent societies, in which agents are able to evolve a moral sense. In the model, moral sense is defined by every agent's personal set of rules, which determines its interaction with other agents. By performing simulations of the model we investigate under what circumstances the agents in the societies develop a moral sense which allows the society to thrive. We use four conceptually different agent designs: agents with a minimum of attributes, agents with family relations, agents with a memory, and agents with reputation in the society. From our results we conclude that there are circumstances under which agents evolve ethical behaviour, those circumstances being specific settings of family relations and reputation.

Keywords Ethics · Morality · Evolutionary algorithms · Multi-agent societies · Simulation

1 Introduction

What is morality, and how did the human race become instilled with a moral sense? Philosophers such as Plato, Kant [14], Spencer [20] and Huxley [12] sought to answer such questions through the formulation of ethical theories. Those theories debate how individuals should behave and what actions should be considered ethical. The definition of morality plays a crucial role in the formulation of ethical theories.

The challenge of ethical theories formulated by the great philosophers was that, whatever theory one devised, any attempt to validate it was restricted to pure reasoning. Empirical experiments could not be executed because experiments on ethics with people were (and are) considered unethical themselves. Nowadays, the research domains of evolutionary computing and artificial life provide tools to execute empirical experiments and provide arguments for and against ethical theories.

Evolutionary ethics aims to bridge the gap between philosophy and the natural sciences by arguing that natural selection instilled human beings with a moral sense [9]. The theory states that individuals in a society will have a higher survival rate when their actions are based upon good morals. Therefore, they have a higher chance to be selected in the evolutionary cycle, which causes their moral behaviour to spread to future generations.

Is the theory of evolutionary ethics a valid explanation for the existence of ethical behaviour? Modern computers provide powerful tools to execute empirical experiments in this respect. In this paper, we present an artificial life model of multi-agent societies, wherein the individual agents have the possibility to evolve a moral sense. In the model, a moral sense is defined by every agent's personal set of behavioural rules. The model is inspired by the previous work in this field of research done by Epstein and Axtell [8], and Mascaro [16].

With our model we aim to investigate under what conditions agents in a society will develop a common moral sense that allows the society to thrive. We are particularly interested in the evolution of ethical behaviour. In our research,

P. Spronck (✉) · B. Berendsen
Tilburg Centre of Creative Computing (TiCC), Tilburg University,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: p.spronck@uvt.nl

B. Berendsen
e-mail: berend.berendsen@quintiq.com

we defined different agent attributes and agent society configurations. We allowed the agents to evolve a moral sense. Whenever a thriving society was evolved, the nature of the resulting set of behavioural rules was examined. The set can have, for instance, an egoistic or altruistic nature, and there might be specific rules which allow a society to thrive. We interpreted the different ethical concepts by using the definitions given by Jaffe [13].

The outline of this paper is as follows: Background material for our research is given in Sect. 2. The society model that we use is explained in Sect. 3. We performed four experiments, namely with (1) baseline behaviour, described in Sect. 4, (2) family relations, described in Sect. 5, (3) memory, described in Sect. 6, and (4) reputation, described in Sect. 7. Our results are discussed in Sect. 8. Section 9 concludes and looks at future work.

2 Background

In this section we provide background information on our research. Section 2.1 describes the philosophy of evolutionary ethics. Section 2.2 describes in what sense our model can be considered an artificial life model. Section 2.3 indicates the little research that has been done in the area of ethics in agent societies. Section 2.4 defines the behaviours that are distinguished in our research.

2.1 Evolutionary Ethics

The term evolutionary ethics denotes an approach to naturalistic moral philosophy which seeks to explain how moral traits and behaviour evolved. Evolutionary ethics tries to bridge the gap between philosophy and the natural sciences by arguing that natural selection has instilled human beings with a moral sense [9]. The theory entails that morality can be understood as a phenomenon that arises during the evolution of sociable, intelligent beings, and not as the result of our rational faculties or of divine revelation. The theory states that a group of individuals which base their actions upon a moral sense will have a higher survival rate. Thus, individuals with a moral sense that stimulates them to behave ethically would have a selective advantage.

According to the theory of evolution, natural selection entails that, in general, only the fittest individuals in any given population will survive and reproduce [6]. An organism's evolutionary goal seems to be to promote its own fitness in order to survive long enough to reproduce. In situations where an organism has to choose between enhancing its own fitness and enhancing the fitness of others, the organism is expected to choose to enhance its own fitness. However, individuals frequently behave in ways that increase the fitness of a group, where the members of the groups are usually of their own kind. This paradox of altruism, which is supported by empirical facts [19], is one of the problems evolutionary ethics are confronted with.

The paradox has implications for the theory of evolutionary ethics. For instance, if some moral traits are altruistic in the evolutionary sense, then the evolutionary explanation of altruism will be a part of the explanation of morality [5]. Darwin circumvented this problem for the theory of evolution by stating that it was a mistake to think that natural selection only operated on individuals [6].

Interdisciplinary approaches between scientists and philosophers have the potential to generate important new ideas, but the theory is confronted with a number of problems, such as the altruism-paradox. Empirical results from the domain of computer science, for example the research discussed in this paper, can help evolutionary ethics to overcome some of these problems.

2.2 Artificial Life

Artificial life is a domain that studies natural life by attempting to recreate biological phenomena using computer simulations which resemble those phenomena [15]. Artificial life is often described as an attempt to understand high-level behaviour which is the result of low-level rules. Artificial life complements the analytic approach of traditional biology with a synthetic approach in which, rather than studying biological phenomena by taking apart living organisms to see how they work, one attempts to put together systems that behave as living organisms.

Behaviour which is not directly programmed in individuals, but which does emerge from the interaction of individuals with each other and with their environment, is called "emergent behaviour" [17]. The simulation of our model, described in Sect. 3, is an artificial life simulation. The difference between "regular" artificial life simulations and the simulation of our model, is that in the first simple rules are defined in individuals of the simulation to create emergent behaviour of the entire system, while in our model we want to *evolve* those behavioural rules. We use genetic algorithms [10, 11] for this evolution process.

2.3 Ethics in Agent Societies

Epstein and Axtell [8] attempted to grow an artificial society from the bottom up by defining simple local rules, which were able to develop complex behaviour of the agents. An example of such a simple rule for an agent is: "Move towards the nearest spot with the most food available and collect all the food from that spot." When the individuals were placed in a world with 'food-mountains,' the individuals showed 'hiving' behaviour. These experiments show that using simple local rules in agents can lead to interesting emergent behaviour within multi-agent societies.

Mascaro [16] suggested using genetic algorithms to evolve behaviours, and then analysing to what extent these behaviours are ethical as defined by Utilitarianism [2]. The simulation by Mascaro [16] allows the creation of basic environments, containing agents that are able to perform a finite number of actions dependent upon their own state and the state of their immediate environment. Suicide, which is the most extreme example of altruism, is one of the possible actions of the agents. Including suicide in the agents' behaviour may result in a stable society, which can cope with variance in the availability of food by making the number of agents correspond to the availability of food.

2.4 Social Behaviour

For our research, it is necessary to classify an evolved moral sense of a society and compare classified moral senses with each other to determine if there are significant differences. For this classification, we need definitions of behaviour and social interaction. The definitions introduced by Jaffe [13] form a system that can be used in concrete situations such as multi-agent societies. Jaffe's system uses values for two terms, namely (1) S, which is the benefit for the society, and (2) I, which is the benefit for the individual. To classify actions performed by agents, four different valuations are distinguished:

Social investment ($S \geq 0$, $I \geq 0$): the action is useful for both the society and the agent;
Destructive egoism ($S < 0$, $I \geq 0$): the action harms society but benefits the agent;
True altruism ($S \geq 0$, $I < 0$): the agent makes a personal sacrifice for the good of society; and
Destructive behaviour ($S < 0$, $I < 0$): the action harms both society and the agent.

These definitions are straightforward and create a framework which can be used when doing research which involves moral behaviour. We chose to classify actions of agents according to these definitions. Moreover, we define a *common moral sense of a society* as the average of executed actions after a certain period in a society.

When do we consider a common moral sense to correspond to ethical behaviour? Naturally, that depends on the system of ethics that we prefer the agent society to subscribe to. In most ethical systems, ethical behaviour has some consideration for the good of the society as a whole, i.e., individuals should be concerned with more than only their own welfare. Therefore, within the confines of Jaffe's system, ethical behaviour means that the common moral sense should avoid destructive egoism and destructive behaviour. Furthermore, we are particularly interested in discovering whether it is possible for an agent society to evolve support for some true altruism, which would indicate that the good of the society is an integral aspect of the common moral sense.

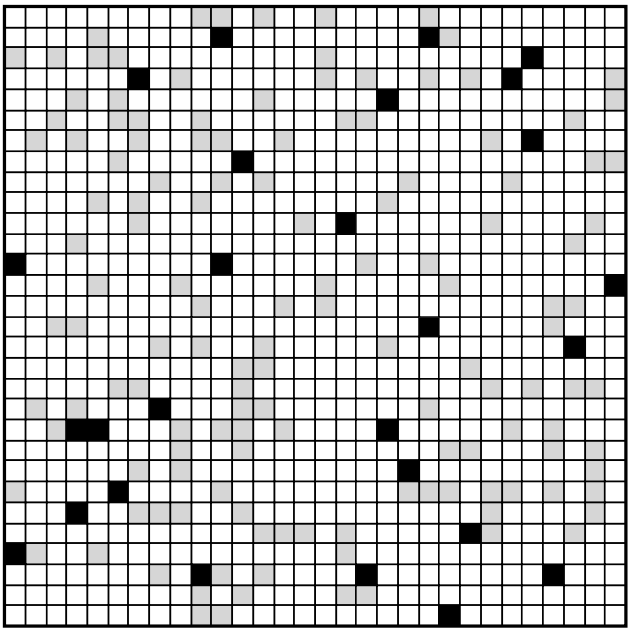

Fig. 1 The environment, 30 × 30 cells

3 Society Model

In this section we describe the design of our model, and the simulation of this model which allows us to investigate multi-agent societies that are capable of evolving a common moral sense. The model consists of agents, their moral sense, and the environment in which the agents live. The simulation of the model is responsible for running the model and collecting statistics of executed experiments.

In Sect. 3.1 the general setup of the model is explained. Section 3.2 describes the attributes of the agents, and Sect. 3.3 describes their behavioural rules. Section 3.4 gives the details of the mating process. Section 3.5 discusses how a common moral sense is ascribed to an agent society. A detailed description of the model's implementation is given by Berendsen [3].

3.1 The Environment

Our society model is situated in an environment, that consists of a two-dimensional 30 × 30 grid. Each of the grid's cells may contain food and/or an agent. When a cell contains food, it will only disappear when it is completely collected by an agent. The environment is bounded by the edges, so there is no wrap-around. A graphical example of the environment can be seen in Fig. 1. At the start of an experiment, the environment contains 200 agents, and 1% of the cells contain food.

If an agent moves towards a cell where food is available then the agent collects the food. The quantity of food that an agent can collect from one cell is a maximum of 20 units.

The food is uniformly distributed over the environment. The factor of cells which receive new food every round are specified by a normal distribution, with a mean of 0.06 and a standard deviation of 0.001.

The unit of time used in the society model is the 'round.' In our experiments, every run in the model will last 20,000 rounds. Each round every agent is allowed to execute one action; this can be an action which is the result of a rule, of the mating process, or of instinct. The order of agents which are allowed to execute actions is randomly determined.

When an agent is selected, it automatically consumes part of its health and ages one round. The agent will then test whether it is capable of starting the mating process. If the agent succeeds to mate with another agent, both agents have finished their action for the round. When an agent is not able to start a mating process, it will try to execute one of its rules. An agent is only able to execute one rule, and when it has done so, the agent has finished its round. If the agent fails to execute any of its rules, it will follow its instinct, which is to forage food.

3.2 Agent Attributes

An agent distinguishes itself from other agents by its characteristics and its behavioural rules. The characteristics are based on the work of Epstein and Axtell [8] and Mascaro [16]. The distinguishing characteristics of the agents are age and health.

Age: The age of an agent is defined by the number of rounds the agent has been alive. An agent can live for a maximum of 100 rounds. The age attribute is introduced to set a natural cap on the agents' lifetimes.

Health: Each agent has an associated number called "health." An agent loses 10 units of health each round. Health is replenished by consuming food, whether foraged, stolen, or gifted. If the health of an agent reaches zero, the agent dies. An agent which is created at the beginning of a run in the model will have 100 units of health. An agent which is created through the mating process inherits 25% of the health of each of its parents. There is no upper limit on the health for an agent, thus health is an element of $\mathbb{N}$. The health attribute is used as the fitness measure for agents. If an agent has more health, it is considered to be more successful. This measure is commonly used in artificial life and multi-agent society studies [8, 16, 18].

An agent's field of vision determines how many cells the agent can see across the environment. An agent can see in the four principal lattice directions: North, South, East, and West. An example of an agent with a vision of 3 can be seen in Fig. 2. All agents in the simulation have a vision of 3, which enables the agents to explore their near surroundings and find food and partners to mate with.

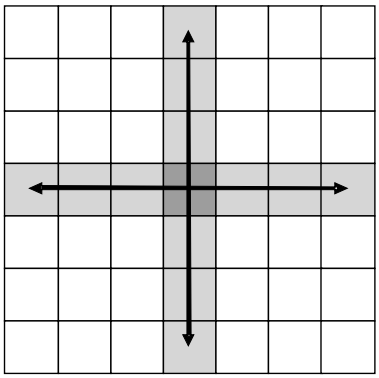

Fig. 2 An agent in a part of the environment with a vision of 3; the *grayed area* is visible to the agent

3.3 Agent Behaviour

An agent is purely reactive, which means that the actions it can execute depend on its characteristics, the characteristics of neighbouring agents, and the nearby environment. Every agent has between 5 and 7 behavioural rules, which are ordered by specificity. These rules are responsible for the actions of the agent and define its personal moral sense. The rules consist of tests and actions which can be represented as: "$TEST \wedge TEST \rightarrow ACTION$."

The first part of a rule is the condition under which the action of the rule will be executed. A condition consists of one or more tests. All tests must be satisfied for the condition to be satisfied. The tests can refer to the age and health attributes of a certain agent, and the availability of food in the neighbourhood. There are five possibilities to refer to an agent within the vision range of the owner of the rule, namely (1) the owner of the rule itself, (2) the closest agent, (3) the furthest agent, (4) the strongest agent, and (5) the weakest agent. The tests can also contain mathematical and logical operators, wildcards, and numbers. An example of a condition of a rule is: "if closest-agent age $<$ 10 AND self health $<$ 20." This condition contains two test clauses and will only succeed when both of them are true.

The second part of the rule is an action, which is the result of successful tests. There are four different actions: wander, forage, steal, and share. Share and steal are actions which involve interaction between two agents. The agent that is stolen from or shared with, is selected by the condition of the rule which has steal or share as action. When a condition consists of multiple test clauses, the agent from the last test clause is used as target for the steal or share action. If the target agent is the owner of the rule itself, a random agent in the vision range is selected. The four actions are defined as follows.

Wander: The wander action of an agent implements a move of random length. This position can be as far as the agent's vision permits. The agent can move in any of the

four lattice directions. If an agent moves to a cell where food is positioned, the agent will collect food, up to a maximum of 20 units (which translates to 20 units of health). Wander is a neutral action; only when an agent encounters food by chance it is positive, otherwise the action has no benefit. The wander action is characterised as *social investment*.

Forage: The forage action simulates the harvesting behaviour of an agent. The agent looks for cells in the environment where food is available, as far as the agent's vision permits. The agent then moves to a randomly-chosen cell with food, and collects food from that cell, up to a maximum of 20 units. Forage is an egoistic, but positive action. The forage action is characterised as *social investment*.

Steal: When an agent executes the steal action it tries to steal health from a neighbouring agent specified in the test of the rule. The amount of health which is stolen is 25 units. Steal is typically an asocial and egoistic action; the agent who executes the steal action is the only one who benefits from it and the victim agent is harmed. The steal action is characterised as *destructive egoism*.

Share: The share action allows an agent to share its wealth (measured in health) with another agent. When an agent shares, it gives away 25% its health to the agent specified in the test of the rule. Share is an altruistic action; the sharing agent gives away his own health for the benefit of another agent. The share action is characterised as *true altruism*.

Note that there is no action which can be classified as destructive behaviour. We excluded this possibility intentionally, because its destructive nature would obstruct the evolution of societies. Moreover, arguably it is a kind of behaviour that has no survival ability at all.

3.4 Mating

The mating behaviour is the same for all agents, although the characteristics and rules of the resulting child-agent of the mating process depend on its parents. An agent tries to mate with another agent before it tries to execute its behavioural rules. The mating process is divided into three parts.

The first part checks whether the agent which starts the mating process satisfies the mating requirements. These requirements consist of tests against the age and the health of the agent. An agent has to be alive for at least 18 rounds and must have at least 50 units of health. The idea is that an agent must be of a certain age and have a certain level of health to mate, because we are only interested in agents who can survive to reproduce.

The second part of the mating process is finding a suitable partner. Partners have to be located in one of the four lattice directions of the agent and have to be within its range of vision. The agent randomly chooses one of its neighbours to be its partner for the mating process. The partner has to meet the same age and health requirements as the initiating agent.

The last part of the mating process is responsible for the creation of the child agent. The parents donate 25% of their health to the child. The child is placed on an empty cell near one of the two parents. It is possible that the child is placed on a cell where food is available, which results in a higher initial health for the agent.

The behavioural rules of the child are determined by the uniform crossover operator [10]. For each part within the chromosome that defines the rules of the child, one of the parents is chosen. In the model this means that for all parts of the rules there is a 50% chance that the part is chosen from parent 1 and 50% chance that it is chosen from parent 2. An example of the uniform crossover operator in the simulation can be seen in Fig. 3, where one rule of each parent is combined to create one rule for a child. The rules of the child also have a 2% chance of being *mutated*. The mutate operator changes a part of the rule into something else. It is also possible that the condition of a rule is expanded or contracted with a test clause, but there is a minimum of zero and a maximum of two test clauses.

3.5 Classification of Moral Sense

The statistics that are gathered during a run are used to investigate the moral sense of the agents within a society. There are three kinds of statistics gathered: (1) statistics on the demographics of the society, (2) statistics on the actions of agents that were executed during a run, and (3) statistics on the rule sets of the agents.

We classify a rule set by looking at the executed actions by all agents of a society in the last 5,000 rounds of a run. Only actions executed by agents which survived at least to the mating age are taken into account, because we are only interested in adult agents which are strong enough to survive and mate. The common moral sense of the society consists of the averages of the moral evaluation of the actions in the rule set. For example, a common moral sense may consist of 30% social investment, 25% destructive egoism, and 45% true altruism, when in the last rounds of the run the agents executed an average of 10% wander, 20% forage, 25% steal, and 45% share actions. We discount any actions which are executed less than 1% of the time, as, due to the mutation factor, it is impossible that specific actions are completely removed from a rule set. We therefore assume that such rare action are not a significant part of the common moral sense.

4 Baseline Behaviour

The goal of the first experiment was to set a baseline for the results for the subsequent experiments. We were particularly interested to see whether the society would evolve to rule out destructive egoism, and perhaps would even allow

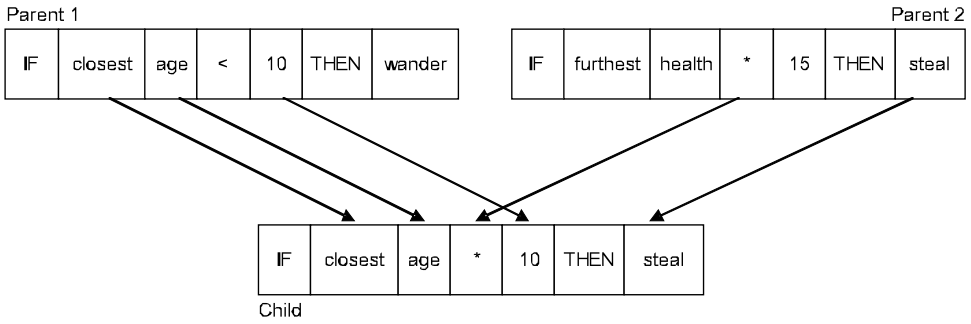

Fig. 3 Uniform crossover operator applied to behavioural rules

Table 1 Percentages of executed actions, averaged over 20 runs

	Average	St. dev
Wander	0.36	0.18
Forage	53.57	6.25
Steal	45.85	6.22
Share	0.23	0.28

some true altruism. We ran 20 repetitions of an experiment with the basic settings as explained in Sect. 3. Table 1 shows the average percentage of actions executed, as well as the standard deviation.

From Table 1 we conclude that there are only two behavioural actions of the agents which are responsible for the common moral sense of the agents, namely forage and steal. The common moral sense is egoistic, as agents are only looking after themselves. The single goal of an agent is to gather as much food as possible, doing that by searching for food, or by stealing food from neighbours.

Foraging is an easy way to obtain health and to ensure survival, which does not depend on the cooperation with other individuals. Stealing is another easy way to obtain health and to ensure survival, but at the cost of a different agent. Because there is no direct punishment on stealing from other agents, there is no difference for an agent between stealing and gathering food, except for the execution of the action itself. Obviously, the society as a whole cannot support only stealing agents, lest it would die out. But a large amount of stealing is certainly viable, and is therefore evolved. Sharing food, within such a society, will probably lead to early starvation.

The evolved societies are stable. Fluctuations in numbers of agents and food availability are small, and the average health and age of the agents remains the same throughout a run. Figure 4 shows the food and population size of the first 5,000 rounds of a typical run. The upper line is the availability of food in every round, counted as the number of cells in the environment which have food available and are not occupied by an agent. The lower line is the population size in every round. As can be seen in the figure, the society reaches its fairly stable state within the first 1,000 rounds of a run.

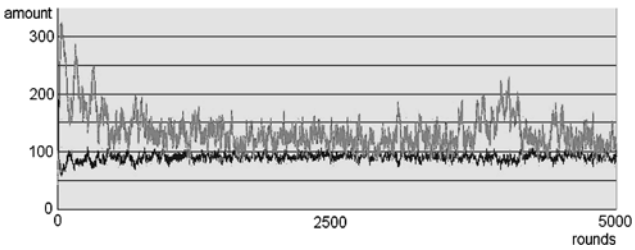

Fig. 4 The available food (*upper line*) and the population size (*lower line*) in the first 5,000 rounds of a run

To investigate whether the availability of food would change the common moral sense of the societies, we varied the mean food distribution within the interval [0.02, 0.09], with steps of 0.01 (0.06 being the default setting). While this significantly influenced the percentages of social investment and destructive egoism (with increasing availability of food, the agents' tendency to steal increased also), this did not introduce true altruistic behaviour. Changing other parameters, such as the number of initial agents, the availability of initial food, the initial health of the agents, and changes to size of the environment, did not result in a significant change of the results.

Why are the agents not behaving altruistically? A possible explanation is that altruistic agents lose their own health by helping others. There is no direct or indirect benefit for an altruistic agent, since it loses health and the receiving agent will not remember that it received health from another agent. Furthermore, the society does not benefit from altruistic behaviour, as agents are able to live and survive without the help of other agents.

5 Family Relations

In the second experiment, the agents in our model are extended with parent-child relations to examine if such relations will enable societies to develop a more altruistic moral sense. The idea behind family relations is that children cannot survive without the help of the parents. Incorporating this idea in our model will bind parents to their offspring

Table 2 Percentages of executed actions, averaged over 20 runs, with parental instinct

	Average	St. dev
Wander	0.38	0.38
Forage	69.52	29.31
Steal	29.31	5.74
Share	0.79	1.32

Table 3 Percentages of executed actions, averaged over 20 runs, with 900 initial agents and no stealing

	Average	St.dev
Wander	2.34	3.99
Forage	82.14	5.29
Share	15.51	3.20

and may possibly result in family structures in which altruistic behaviour is executed.

To accomplish that parents are able to help their children, it is necessary for the parents to be able to identify their children. We extended the agents with knowledge of its children. The language of the rules was extended by adding "child" to possible agents used in the tests. When a test of a rule includes "child," the tests are executed on a neighbouring child of the owner of the rule.

An agent is a child in the first 5 rounds of its lifetime. Children are characterised by the following four features: (1) a child has a vision of 1, which means it will only move to a cell adjacent to itself, (2) a child is only able to steal 3 health from another agent, (3) a child is only able to share 3 health with other agents, and (4) a child can only consume 3 health from a space with food on it. The mating process is altered such that children have 40 health when they are created. These alterations are responsible for simulating weak children and exclude the possibility that agents survive childhood without the help of their parents.

To our initial dismay, we found out that running a simulation of our model with these settings for 20 runs, resulted in 20 extinct societies. Each of the societies died out within the first 150 rounds of its run. There are at least two reasons why the societies do not survive. First, children are an easy victim for agents who steal, and a child is unlikely to survive the theft. Second, the initial behaviour of agents must already include sharing with children, and this trait must be transferred to children, both of which are not statistically likely.

Thus, we decided to also extend the agents with a parental instinct. This means that if the agent does not find an action to execute in its behavioural rules, it will share with its child. Parental instinct is not counted as an executed share action, because it is not part of the evolved rule set. Parental instinct enabled the societies to survive and the resulting average percentage of executed actions can be found in Table 2. Although surviving societies are found, the common moral sense did not include true altruism. The children survived because of the parental instinct, making sharing unnecessary for the survival of the societies. This results in the same situation as the baseline model from Sect. 4.

We then continued the experiment by removing the parental instinct, but increasing the number of initial agents to 900, resulting in a fully-occupied initial environment. This could be the solution to the problem of not having a sufficient initial number of sharing agents in a run. We also removed the steal action from the agents in the model, which could solve the problem of children being the victim of stealing agents.

Of 2,000 executed runs, only 20 resulted in a surviving society. Table 3 shows the average of executed actions in the last 5,000 of 20,000 rounds of those 20 surviving runs. Running the simulation of our model with the above described settings does result in surviving societies, but the probability that a society survives is very small. The results from Table 3 show that the resulting common moral sense consists of about 15% true altruism. Examining the rule set, we found that, indeed, parents behaved altruistically towards their children.

From the results of the second experiment we make three observations:

1. Weak children do not survive, because parents do not behave altruistically towards their children. This results in the extinction of the entire society;
2. When we extend the agents with a parental instinct, the societies are able to survive, but the common moral sense does not include altruistic behaviour; and
3. Large populations with weak children and family relations do not survive, except when altruistic behaviour evolves in the common moral sense and when stealing behaviour is excluded.

The results of the experiments with family relations without stealing behaviour show that a common moral sense is evolved in which agents decide to help other agents by sharing their belongings. The results provide arguments to believe that the theory of evolutionary ethics is a valid explanation for the existence of a moral sense, but in our model only with highly specific parameters.

6 Memory

In the third experiment the agents from the basic settings discussed in Sect. 4 were extended with memory. The reason why we implemented memory is that it could lead to a more altruistic common moral sense, because agents will be able to remember whether other agents interacted positively

Table 4 Percentages of executed actions, averaged over 20 runs, with agents with memory

	Average	St. dev
Wander	0.75	1.64
Forage	62.39	7.50
Steal	36.50	7.50
Share	0.36	0.41

Table 5 Percentages of executed actions, averaged over 20 runs, with reputation with the basic setting

	Average	St. dev
Wander	2.94	3.63
Forage	62.27	6.89
Steal	35.02	7.86
Share	0.22	0.12

or negatively with them. To extend the agents with a memory, they need a list of agents who interacted with them. For each other agent they keep a balance of how much these other agents shared with them, and how much they stole from them. A positive balance indicates more sharing than stealing, and a negative balance more stealing than sharing. To enable agents to identify friendly and hostile agents in their rules, the language of the rules was altered. Instead of using the closest, furthest, weakest, and strongest agent, we used friendly, hostile, and neutral agents to test characteristics against. These agents are specified as follows.

Friendly agent: The neighbouring agent with the most positive balance between the owner of the rule and the neighbouring agent, is defined as the friendly agent. If there is no neighbouring agent with a positive balance, the test will fail.

Hostile agent: The neighbouring agent with the most negative balance between the owner of the rule and the neighbouring agent, is defined as the hostile agent. If there is no neighbouring agent with a negative balance, the test will fail.

Neutral agent: The neighbouring agent with a balance of zero between the owner of the rule and the neighbouring agent, is defined as the neutral agent. When there are more agents which satisfy these requirements, one is randomly chosen as the target. If there are no neighbouring agents with a balance of zero, the test will fail.

Table 4 gives the results for the experiment with agents with memory. These results show that, while the amount of stealing that occurs is reduced compared to the baseline model (Sect. 4), individual memory does not lead to a more altruistic common moral sense. An explanation for this is that agents, due to their relatively short lifetimes and ability to move, do not interact often enough with each other to build up a memory which could influence the common moral sense.

7 Reputation

The memory of agents introduced in Sect. 6 did not result in an altruistic common moral sense, arguably because of a lack of interaction. We therefore introduced reputation in the agents. Reputation can be seen as a global memory, every agent having a friendly, hostile, or neutral reputation to all other agents. Reputation is incorporated in the model, based on the model used in Sect. 6, by giving the basic agents a friendly, hostile, or neutral reputation. We experimented with three different settings of reputation, which are defined as follows.

Basic setting: The reputation of an agent is increased when the agent decides to share with another agent, and decreased when the agent decides to steal from an other agent. A friendly agents is an agent which shared more than it stole in its lifetime and a hostile agent is an agent which stole more than it shared in its lifetime.

Friendly forever: When an agent decides to share with another agent, it receives a friendly reputation which will last until the agent dies.

Hostile forever: When an agent decides to steal from another agent, it receives a hostile reputation which will last until the agent dies.

Table 5 lists the results of the basic setting of reputation. We examined only societies in which instinct no longer plays a role in the gathering of food. The results show that, while the amount of stealing is reduced compared to the baseline model, there is no significant altruistic behaviour in the evolved societies. Actually, the results are quite similar to those achieved with the experiment with memory (Sect. 6).

Table 6 lists the results of the friendly forever setting of reputation. Again, we used only societies which evolved to a state in which the instinct is no longer used by the agents. The results show us that altruistic behaviour becomes a small but significant part of the common moral sense of societies. Investigating the rule sets of evolved societies shows that agents have rules which prescribe to share with friendly or neutral agents.

Table 7 lists the results of the hostile forever setting of reputation. Once more, we used only societies which evolved to a state in which the instinct is no longer used by the agents. The results show that, compared to all other models tested, there is significantly less destructive egoism behaviour. However, altruistic behaviour is not included in the common moral sense. Investigating the rule sets from evolved societies shows that agents have rules which prescribe to steal from hostile agents. This leads to a situation in

Table 6 Percentages of executed actions, averaged over 20 runs, with reputation and the friendly forever setting

	Average	St. dev
Wander	2.81	4.52
Forage	55.68	21.01
Steal	37.81	19.83
Share	3.69	4.17

Table 7 Percentages of executed actions, averaged over 20 runs with reputation and the hostile forever setting.

	Average	St. dev
Wander	9.47	9.86
Forage	70.34	7.84
Steal	20.02	9.50
Share	0.17	0.12

which hostile agents steal from other hostile agents, which has as a result that hostile agents have a high chance of not surviving to the mating age.

8 Discussion

Can our results contribute to the discussion of the theory of evolutionary ethics? From our personal view they can: even though our model is a simplistic representation of the real world, it can simulate the evolution of a moral sense in a society. Complex interaction between individuals within a society are supported, and through reproduction and death, a moral sense is generated.

Our results show that adding even a little bit of complexity to the agents in the society reduces the amount of destructive egoism professed by the agents. However, the results show also that evolving an altruistic common moral sense is not as evident as expected. In a number of situations an altruistic common moral sense did indeed evolve, but in other situations where an altruistic moral sense was expected it did not evolve. It is possible that this is the result of the simplicity of the agents in our model, or because of assumptions made when we defined our model. However, we got the impression that evolving ethical behaviour in general (such as the behaviour we see in human societies and many animal societies) is not as obvious as it seems.

Computer science and ethics are two research domains which are rarely combined, even though computer science can offer a wide range of new insights into problems which are considered to be solved by philosophers. When computer science, and especially social simulations, are combined with ethics, this is usually done with the iterated prisoner's dilemma [1, 4, 7, 21]. The iterated prisoner's dilemma can offer interesting insights into ethics, but the problem itself is rather simple. The simplicity of the iterated prisoner's dilemma is also its power, because it has only a limited number of outcomes, thus conclusions and explanations are relatively easy to provide. The danger of drawing conclusions from research done with the iterated prisoner's dilemma is that it is hard to say something about ethics in general or ethical behaviour within humans. The iterated prisoner's dilemma is just a model of a simple problem with a limited number of choices.

We presented a more complex model, which incorporates interaction between agents that are free to move around and reproduce. Our model also offers a very large number of possible experiments for evolving a moral sense. We should add that, naturally, even with this more complex model we can not draw hard conclusions about moral sense in the "real world."

9 Conclusion

We experimented with four different conceptual designs of the agents in a model of agents societies: agents with a minimum of attributes, agents with family relations, agents with a memory, and agents with reputation in the society. The baseline experiments with agents with a minimum of attributes resulted in egoistic societies. The introduction of family relations resulted in societies which were unable to survive, except when parental instinct was added or highly specific parameters were used. With those specific parameters a common moral sense was found which consisted for 15% of altruistic behaviour. The introduction of memory resulted in the same egoistic societies as with the agents with a minimum of attributes. The results of experiments with reputation depend on the setting of reputation which is used; in certain settings an altruistic common moral sense was found.

We answer the research question that we posed in Sect. 1 as follows: In every experiment, except for the test with family relations and weak children, the agents developed a common moral sense which allowed their society to thrive. The evolved common moral sense was almost always a mixture of destructive egoism and social investment, which means it was decidedly egoistic. However, compared to the baseline society that used the simplest agents, all societies had a significantly reduced amount of destructive egoism. Furthermore, a more altruistic moral sense was found in two society configurations, namely (1) with family relations, a large initial number of agents, and no stealing, and (2) with a friendly-forever reputation setting.

In future work, more actions for the agents can be incorporated in the model. Suggestions for such actions are 'kill' and 'suicide.' Mascaro [16] suggests 'rape,' 'abortion,' and 'racism.' Another possible extension is the introduction of tribes.

Is the theory of evolutionary ethics a valid explanation for the existence of ethical behaviour? In very simple societies in nature, for example bacterial colonies, we do not observe altruistic behaviour. However, altruistic behaviour often occurs in more complex societies, even relatively simple ones such as ant colonies. Our experiments show that ethical behaviour can evolve even in a simple society, given the right circumstances. This lends credibility to the idea that in a highly complex society, such as our human one, good morals may very well be the result of evolutionary ethics.

References

1. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
2. Bentham J (1780) An introduction to the principles of morals and legislation. Clarendon Press, Oxford
3. Berendsen B (2005) Evolutionary ethics in agent societies. Maastricht University, Maastricht
4. Bradish S, O'Riordan C (2000) The voter's paradox? Evolution of cooperation in N-player games. Technical Report, Department of IT, NUI, Galway, Ireland
5. Byron M (1999) Evolutionary ethics and biologically supportable morality. In: Proceedings of twentieth world congress of philosophy: philosophy educating humanity
6. Darwin C (1871) The descent of man. Watts & Co., London
7. Delahaye J, Mathieu P (1996) Random strategies in a two levels iterated prisoner's dilemma: How to avoid conflicts? In: Miller H, Dieng R (eds) Proceedings of the ECAI 96 workshop: modelling conflicts in AI, pp 68–72
8. Epstein J, Axtell R (1996) Growing artificial societies. MIT Press, Cambridge
9. Fieser J, Dowden B (2000) Internet Encyclopedia of Philosophy. http://www.utm.edu/research/iep/
10. Goldberg D (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading
11. Holland J (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge
12. Huxley T (1898) Evolution and ethics. Appleton, New York
13. Jaffe K (2003) Altruism, altruistic punishment and social investment. Acta Biotheor 52(3):155–172
14. Kant I (1797) Groundwork of the metaphysics of morals. Barnes & Noble, New York
15. Langton C (1997) Artificial life, an overview. MIT Press, Cambridge
16. Mascaro S (2001) Evolutionary ethics. Master's Thesis, School of Computer Science and Software Engineering, Monash University, Victoria, Australia
17. Pfeiffer R, Scheier C (1999) Understanding intelligence. MIT Press, Cambridge
18. Resnick M, Begel A (1991) Starlogo, artificial life simulation. http://education.mit.edu/starlogo/
19. Sober E (1993) Evolutionary altruism, psychological egoism, and morality: disentangling the phenotypes. In: Nitecki M, Nitecki D (eds) Evolutionary ethics, pp 199–216
20. Spencer H (1874) The study of sociology. Williams & Nordgate, London
21. Yao X (1999) How important is your reputation in a multi-agent environment. In: Proceedings of the 1999 IEEE international conference on systems, man, and cybernetics, pp 575–580

Pieter Spronck is an Associate Professor at the Tilburg centre for Creative Computing (TiCC) of Tilburg University, The Netherlands, and also at the Dutch Open University, Heerlen, The Netherlands. His research, publications, and teachings are focused on artificial intelligence in games, in particular on the application of machine learning techniques to make game artificial intelligence more effective and/or more entertaining.

Berend Berendsen received his MSc in Artificial Intelligence from Maastricht University, The Netherlands. He is currently employed as a specialist in Scheduling and Planning software with Quintiq, 's Hertogenbosch, The Netherlands.