Application of Moral Norms in Behavior Modelling of Artificial Agents

Valery Karpov[1], Petr Sorokoumov[1]

[1] *National Research Centre "Kurchatov Insitute", 1, Ac.Kurchatov Sq., Moscow, 123182, Russia*

Abstract

This article describes a model of a social agent, whose behavior can be stated in terms of basic moral mechanisms and norms. Morality is considered here as a flexible adaptive mechanism that allows agents to vary behavior depending on the environment conditions. The control system of the social agent is based on the emotional-needs architecture. This architecture allows to interpret the agent`s behavior in terms of empathy, sympathy and friend-foe relationships together with the mechanisms of imitative behavior and the identification of other observable agents with the subjective "I" concept. Experiments with this model are described, the main variable parameter of which was the tendency to sympathy. The objective of the experiments was to determine the dependence of the group "well-being" indicators on their altruism. The results obtained are quite consistent with the well-known sociological conclusions, which made it possible to say that the proposed behavioral models and architecture of agents are adequate to intuitive ideas about the role and essence of morality. Thus, the possibility of transition in this area from abstract humanitarian reasoning to constructive schemes and models of adaptive behavior of artificial agents was demonstrated.

Keywords

Social behavior models; moral agent, animat, emotional-needs architecture, empathy, sympathy

1. Introduction

The study of the social behavior of agents is one of the most important areas of artificial intelligence, because, as in natural societies, it allows a group to achieve a synergistic effect i.e. increase the capabilities of the group compared to the sum of the individual capabilities of its members. From a technical point of view, the formation of social communities is one of the most effective methods for adapting groups of agents. An important place in such studies is occupied by models in which the purposeful behavior of the entire society is determined by models of individual behavior and local interaction of group members.

The applications of such models can be very diverse, but they are mainly divided by the solved tasks into two groups: the creation of stable societies from individuals and the destruction of already existing societies. The tasks of the first group include, for example, modeling natural eusocial communities with the transfer of the obtained models to groups of robots [1,2]; the second task is the modeling of bacterial communities (biofilms) to destroy them [3]. Then the stability of society is equivalent to the ability to maintain homeostasis i.e. to maintain the character of functioning even with changes in the details of behavior.

Thus, it is obvious that modeling of society requires the use of adequate models of individual behavior of individuals, in which methods of maintaining homeostasis are used. One of the areas of

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL: karpov_ve@mail.ru (A. 1); petr.sorokoumov@gmail.com (A. 2)
ORCID: 0000-0002-9364-1223 (A. 1); 0000-0002-2930-3860 (A. 2)

research in group interaction is the paradigm of social behavior modelling in groups of agents or robots. Within this paradigm, the tasks of creating models of individuals capable of social (inter-agent) interaction can be solved. It also can build mechanisms that determine the conditions for the formation of sustainable societies. The purposeful management of the social behavior of artificial agents can be studied also. Within the framework of the society management the study of the regulatory mechanism, which is inherent in highly developed societies and is called morality, is of particular interest. Here morality is considered in its direct sense, as a kind of regulatory mechanism that determines the ways of resolving conflicts within a group.

The proposed work considers the issue of applicability of the emotional-need control system of an individual-agent to solving the problem of organizing the interaction of agents and resolving conflict situations on the basis of mechanisms that are commonly referred to as moral regulations.

2. Emotional-needs architecture

The individual behavior of agents (both living and artificial) is determined by the structural features of their basic, "physiological" level, which operates with such entities as stimuli, needs and emotions. The role of emotions in the formation of ethical norms and how emotions determine the ethics of human behavior are actively studied by both philosophers and sociologists [4, 5]. Moreover, Marvin Minsky in his work [6] suggests considering emotions as a different way of thinking. Below, emotions are used simply as a mechanism that affects the success of the functioning of an artificial agent in complex non-deterministic environments.

One of the most constructive models is P.V. Simonov's need-information theory of emotions [7]. According to it the integral assessment of the situation is determined by the assessment of the balance between the necessary and available means of meeting urgent needs. The ratio proposed by Simonov can be written as:

$$E = f(N, p(I_{need}, I_{has})) \quad (1)$$

where E is the estimation of the emotion (its magnitude and sign); N is the strength of the current need; $p(I_{need}, I_{has})$ is an assessment of the ability to satisfy a need based on innate and acquired life experience; I_{need} is information about the way to satisfy the need; I_{has} is information about the funds available to the agent, required to meet current needs. Obviously, such an approach makes it possible to calculate the emotional reaction of an agent based on assessments of needs and means of satisfying them. The formalization of this qualitative, evaluative ratio for determining emotions is as follows.

Let the agent's behavior be determined by a triple:

$$Agent = \langle S, N, R \rangle \quad (2)$$

where R is a set of production rules, N is a set of needs, S is a set of sensors. Let's represent R rules in MYCIN-like form [8], i.e. as a set of products with confidence coefficients:

$$R_i: (Cond_1 \,\&\, \ldots\, Cond_k) \rightarrow f \mid w_i$$

Here R_i is a rule, $Cond_i$ are conjuncts of a condition, $Cond_i \in [0,1]$; f is a number of the performed action, $f = 1..M$; w_i is the confidence coefficient (CC) of the rule, $w_i \in [0,1]$, M is the number of actions (behavioral procedures) that the agent performs. Each product assesses confidence that the agent will take the specified action f. The confidence factor of the conclusion is calculated according to the usual rules of fuzzy logic. In the developed model, it turned out to be convenient to choose these rules so that all CCs fell into the range [0.1]. In addition, let us indicate for each product the need $Need$ (R_i), the satisfaction of which the action is directed to.

The agent's actions can be complex procedures that implement whole complexes of actions (in biology they are called “fixed complexes of actions”, FCAs) or even behavioral sequences. In other words, it is assumed that we are talking about a high-level architecture that operates precisely on the behavior of the agent.

After calculating all CCs the agent will be able to choose which of the procedures to run. Thus, the set of productions R generates a vector of CCs that each action should be performed. If two different rules require the same action to be performed, their CCs $A1$ and $A2$ are combined according to the rule:

$$A_{sum} = A1 + A2 - A1A2$$

After choosing an action using formula (1), one can evaluate the emotions evoked in the agent after such a choice. In this case, I_{need} is defined as a vector of CCs of actions conclusions ($I_{need} = A$), and I_{has} is defined as a vector of actions actually performed at a given time. Its dimensions are the same as for A; the value "1" is set for the actions actually performed, and 0 for the rest. For simplicity, the function f in formula (1), combining the strength of a need and the ability to satisfy it, will be considered a product. Thus, calculating the CCs A_i for all actions at the moment, we can determine their emotional estimates E_i, which make up the vector of emotions E:

$$E = k\,(I_{has} - I_{need}),\ i = 1..M \qquad (3)$$

where k is a constant coefficient. It can be seen that if some action cannot be performed (since only one action is performed at a time), then the agent has a negative emotion. The sum of the emotions associated with all actions is the estimation of the general emotional state of the agent.

Let us further assume that the agent's emotional state affects his perceptions, but not directly on the sensors, but through intermediate gates G. As a result, the agent perceives from each sensor not the sensory data S itself but some emotionally colored perception: $G_i = G_i\,(S_i, E)$. All gateway functions G_i were defined as linear combinations of S_i and E components in the current model.

It can be seen that, ultimately, emotions form a feedback from the selected actions to the input of the selection mechanism i.e. to the conditions of the *Cond* rules, using not direct S, but emotionally colored sensory data G. As a result, the functioning of the agent's emotional-needs scheme of behavior looks like an endless cyclical process:

1. Formation of a vector of emotionally colored perceptions G.
2. Calculation of the vector of coefficients of confidence in the conclusion of the rules A.
3. Selecting an action to be performed.
4. Formation of the vector of emotions E.

This circuit is shown schematically in Fugure. 1.

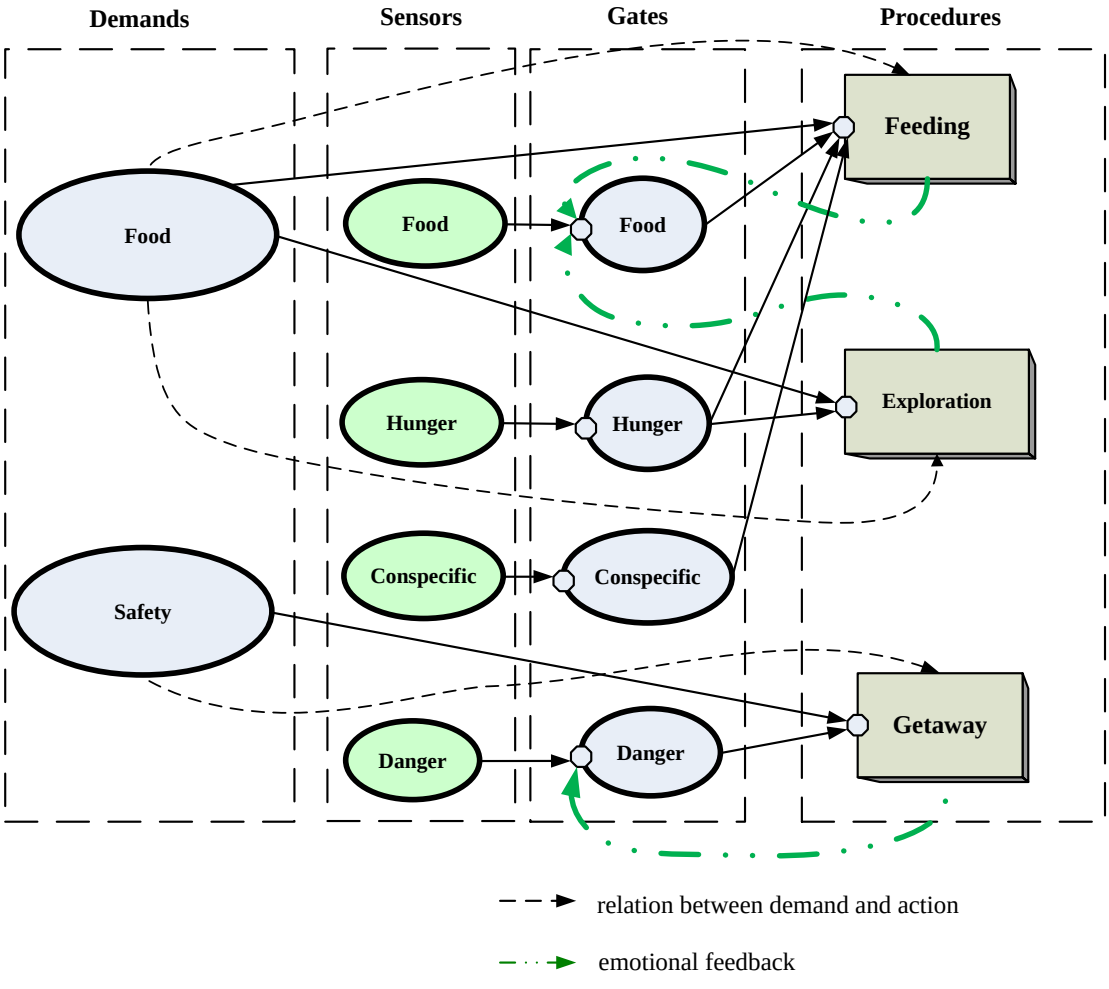

Figure 1: Emotional-needs architecture

3. Moral norms as a system of rules

Morality, if we consider it a set of behavioral rules at the top level, is one of the mechanisms of adaptation. Morality is the most flexible and variable superstructure of the control system. Moral norms can change many times during the life cycle of an individual. Managing the needs of an individual or the nature of his behavior are rather crude methods of intervention, fraught with impairments to performance. It is much more convenient to operate with a superstructure that is responsible for the nature of intrasocial relations, i.e. what we call morality in human society. The question of whether the concept of morality is applicable to artificial agents is decided by the philosophers themselves ambiguously [9].

The most important feature of morality as an adaptive and regulatory mechanism is its flexibility and variability. This superstructure over basic behaviors is extremely "lightweight" and can vary widely throughout the individual's life cycle.

It is believed that the basis of moral behavior is empathy. It is the ability to respond to the emotional state of a conspecific – another one, a representative of the same species. A special case or manifestation of empathy is the so-called “sympathy”, i.e. mutual involvement and other positive reactions. Sympathy-empathy is based on three basic mechanisms: (1) the ability to identify conspecifics with oneself (the “I” component of the sign world picture, if the agent has one), (2) the presence of an emotional-needs control system and (3) the presence of signal communication.

Mechanisms (2) and (3) allows the agent to generate certain signals perceived by others, corresponding to his certain emotional state. Within the framework of the emotional-needs architecture, this means the generation of communication signals associated with the gateways that are most active at a given time (Fig. 1). This is the mechanism that works for animals when signal communication reflects the emotional state.

The matching mechanism is based on the fact that the agent's control system operates with semiotic structures, and the mechanism of internal activation of sign components operates in them. In this case, if the agent perceives a conspecific as a friend (or as a similar individual) this observation excites the percept of the “I” sign, which entails the associative excitation of the “meaning” and “sense” peaks of this sign. Then relevant procedures and other elements of the sign network are activated. In the architecture discussed above, we are talking about the excitation of gateway elements, i.e. the emergence of a situation in which the absence of a clearly observable stimulus (sensor) still forms a response (excitation of the gateway *G* and the corresponding procedure). The need for concepts for the emergence of morality was noted, for example, in [10].

So, let’s introduce a certain meta-parameter, which we will call the inclination to sympathy *S*, $S \in [-1, 1]$. This parameter determines the perception of other agents as friends or foes. An agent with $S > 0$ identifies the observed conspecific with himself with a confidence factor *S*; the semiotic mechanism of such an identification is described above. As a result of this identification, the agent can, for example, share food with this conspecific. An agent with negative sympathy does not identify with other agents, considering them as a source of resources (food) with a confidence coefficient $|S|$. With zero sympathy, the agent is indifferent (tolerant) to other agents. Moreover, if an agent sees a stranger whose resource is greater than that of himself, then this stranger is perceived as a danger.

Interpretation of the influence of "moral norms" on the previously described emotional-need scheme is carried out through the described gates. Let them not only add emotions to sensory perceptions, but also modify them in accordance with some set of fuzzy moral rules, for example:

IF observing_specifics AND S <0, then initiate the "Danger" gateway

IF I observe_specific AND S> 0, THEN initiate the "Friend" gateway

4. Model experiments

To assess the proposed modification of emotional-needs architecture and the empathic mechanisms of the phenomena of moral behavior, the problem of modeling the feeding behavior of a community of agents (we will call them animats wheh their likelihood with living creatures will be important[11]) was considered. The world in which society is modeled is a cellular field in which agents and food particles are located. Each cell can contain a food particle and an arbitrary number of

agents. To exclude edge effects, the field has a toroidal topology, that is, the top and bottom, as well as the left and right sides are connected.

Modeling is carried out step by step. At each step, the agent performs one of the admissible elementary operations: moving to one of eight neighboring cells, eating food, or feeding another agent. Eating replenishes the agent's amount of resources, which is expressed by the value $R \in [0, 1]$. As food, the agent can perceive both food particles and other agents; in this case, it is only permissible to eat a weaker agent by a stronger one (i.e., having a greater value of the resource R).

The animat's sensors determine the perception of the external world, as well as the internal state, and are set by numbers in the range [0,1]: the presence of food, the assessment of the feeling of hunger ($1-R$), the presence of a friendly agent, and the presence of a hostile agent.

Animats have two needs: for food and for safety. They are constant values in the range [0,1], characterizing the significance of the corresponding factors. The agent has three actions available: feeding, getaway and exploration. For each step the agent spends a fixed amount of resource R; if R is zero, the agent cannot perform actions. Altruistic agents are able to feed other agents; their own their food supply is halved after the feeding. Agents are not able to die.

The described environment was implemented as a Python application. An example of the type of medium with agents and food particles in its interface is shown in Fig. 2. At the beginning of work, agents are randomly distributed around the world. Food particles on the map are replenished after each modeling step: in each cell adjacent to the edible particle, a new particle can appear with a certain specified probability. This probability, given a fixed initial amount of food on the map, determines the resource wealth of the environment.

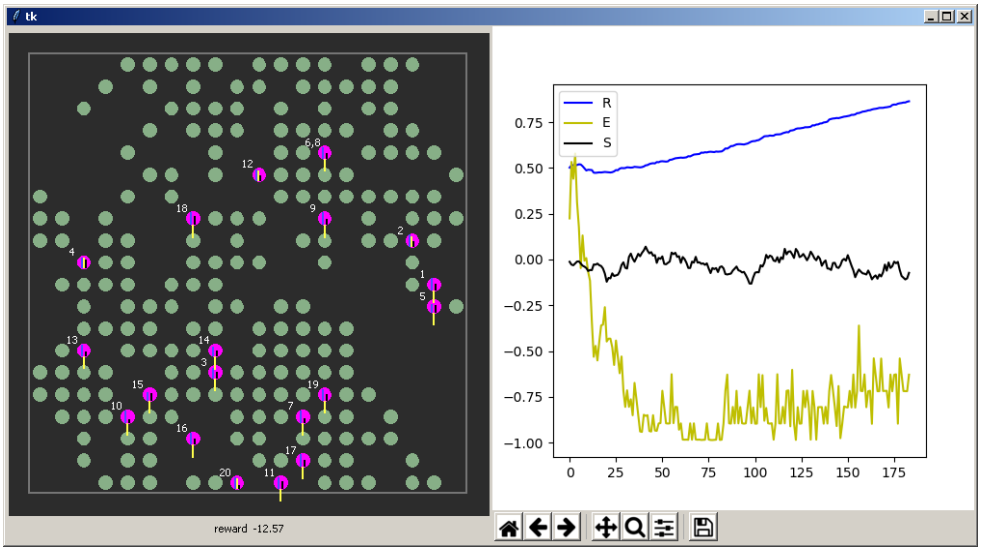

Figure 2: The appearance of the model in the developed modeling environment. Green circles represent food particles, numbered pink circles represent agents. Parameters R (amount of resource), E (emotional level), S (level of sympathy) are shown on each agent by columns. The right side shows the graphs of the dependence of R, E, S on the simulation step

To assess the influence of empathy S on the efficiency of society, a series of experiments was carried out with the developed model. At the same time, the groups of agents had the same number (50 participants), but a different proportion of altruistic agents (from 0 to 100%). At the same time, all selfish agents have a propensity for sympathy S = -0.75, and all altruists have S = 0.75. Average values of R and E at the last step of modeling are shown in Figure 3.

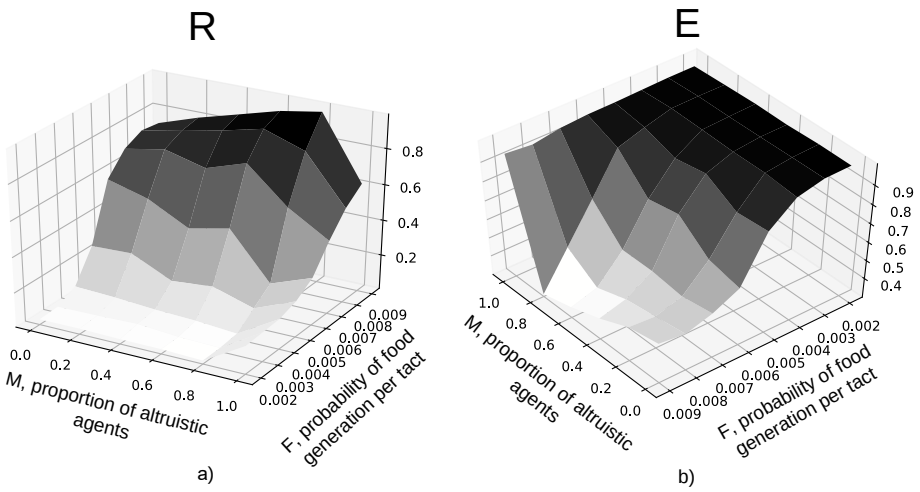

Figure 3: Graphs of the dependences of the group's success on the conditions of existence *F* and the proportion of altruists *M*. The directions of change in *F* on the graphs are opposite to more clearly demonstrate the nature of the dependence

With an increase in the productivity of the environment and constant sympathy *S*, the average values of *R*, as expected, grow monotonically. However, the dependence of *R* on the proportion of altruistic agents turned out to be non-monotonic; in particular, for the most productive environments (when food multiplication probability was 0.009), the highest average *R* value was achieved with 80% altruistic agents. A fully altruistic collective turned out to be generally less productive in rich environments than collectives with aggressive agents, but in poor environments its productivity is higher. The most likely reason for this dependence is the constant redistribution of resources by agents instead of collecting new resources when predators are absent; this reduces both resource losses during predation and the intensity of their gathering.

An analysis of the *E* values in the same results shows that, in general, in the absence of resources, the emotional level is significantly higher than in the presence of them. This is explained by the presence in such a situation of only one option for action - the search for food; if the agent has to choose from a variety of options, some of them remain unused, and this lowers *E*.

The change in moral norms over time can be modeled by modifying *S* according to some law. For example, the effectiveness of society when changing moral norms to more or less altruistic ones was studied (Figure 4).

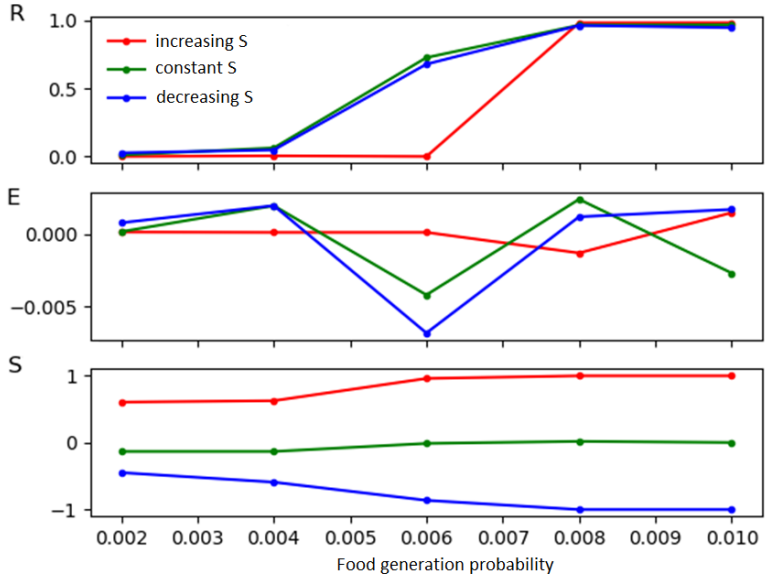

Fig. 4. Dependence of the parameters of the efficiency of society on the direction of change in moral norms

It can be seen that an increase in altruism in an environment of average productivity leads to a drop in *R* (the average amount of a resource for a group at the last step of modeling), while such an increase is not accompanied by significant changes in the emotional level. Here the model reproduces the observed decline in the viability of society with a sharp decrease in the level of danger for individuals (and, as a consequence, an increase in altruism).

5. Conclusions

In this work the task of studying the influence of the number of egoists/altruists on the efficiency of society, how this ratio depends on environmental conditions, etc. was not set. There are a lot of such works both in sociology and in psychology. The goal was only to demonstrate the efficiency of some models of individual organization and mechanisms of inter-agent interaction in solving problems of managing the society's behavior at the highest level, at which aspects of interaction are usually considered and evaluated from the moral point of view.

The developed modification of emotional-needs architecture made it possible to reproduce in the model some characteristic features of the behavior of real societies, showing the potential of including formalized moral norms in the architecture of agents. Experimental testing of this architecture on real groups of robots performing practically important tasks looks promising.

6. References

[1] Burgov E.V., Malyshev A.A. Qualitative and quantitative characteristics of bionispirated models of group robotics (In Russian) // 5th All-Russian scientific and practical seminar “Unmanned vehicles with elements of artificial intelligence”. 2019. P. 139-148.

[2] Vorobiev V.V., Rovbo M.A. Application of learning transfer in semiotic models to the problem of foraging with real robots (In Russian) // International journal "Software products and systems", 2020. Vol. 21. P. 413-419.

[3] Koshy-Chenthittayil S. et al. Agent Based Models of Polymicrobial Biofilms and the Microbiome - A Review. // Microorganisms. 2021. Vol. 9, no. 2.

[4] Callahan S. The Role of Emotion in Ethical Decisionmaking // Hastings Cent. Rep. 1988, Vol. 18, № 3.

[5] Neu J. An Ethics of Emotion? Oxford University Press, 2009.

[6] Minsky M. The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind. Simon & Schuster, 2006.

[7] Simonov P.V. Needs-informational theory of emotions (In Russian) // Questions of psychology. 1982. Vol. 6. P. 44–56.

[8] Jackson P. Introduction to expert systems (3rd edition). Addison-Wesley, 1998. 560 p.

[9] Wallach W., Allen C. Moral Machines: Teaching Robots Right From Wrong // Moral Machines: Teaching Robots Right from Wrong. 2008.

[10] Parthemore J., Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency // Int. J. Mach. Conscious. 2013, Vol. 05.

[11] Wilson S.W. Classifier Systems and the Animat Problem // Mach. Learn. 1987, Vol. 2, № 3.