Can a Robot Be a Moral Agent?

Valery E. Karpov[1,2(✉)]

[1] National Research Centre "Kurchatov Institute", 1 Ac. Kurchatov Sq., Moscow 123182, Russian Federation
karpov.ve@gmail.com
[2] Moscow Institute of Physics and Technology (State University), MIPT, Institutskiy per. 9, 141701 Dolgoprudny, Moscow Region, Russian Federation

Abstract. Issues of the ethically aligned design of intelligent/autonomous systems have now moved into the fields of normative and technical regulation. If a system must make ethically determined decisions, then it must be recognized as a moral agent. This paper provides a list of the properties of a moral agent and shows not only that an artificial agent can have such properties, but also that they are technically determined as manifestations of adaptive mechanisms. In particular, it is shown that mechanisms such as the presence of the "I" component in the sign-oriented picture of the agent's world, the presence of an emotional-needs architecture, and the mechanism for comparing the observed conspecific with the "I" make it possible to realize the phenomena of social learning and a property such as empathy.

Keywords: Moral agent · Emotional-needs architecture · Empathy · Social learning · Imitative behavior · Ethically aligned design

1 Introduction

The ethical issues of artificial intelligence have long been an actively discussed topic, and, in recent years, these issues have moved from the category of humanitarian considerations into the field of technical regulation. For example, the IEEE has launched a global initiative for research in the field of the ethics of AI. The results of such studies should be technical regulations governing the development and implementation of AI systems, with requirements for their ethical behavior. The title of the document is noteworthy: "Ethically Aligned Design". Another illustrative example is UNESCO's report on ethics of robots, entitled "Report of COMEST on Robotics Ethics" (authored by COMEST—the World Commission on the Ethics of Scientific Knowledge and Technology) [1].

Most discussions about the ethics of intelligent/autonomous systems (I/AS) concern various kinds of threat, the social and economic consequences of their use, the ethics of the developers themselves, etc. In this work, we are interested in a different aspect of the ethics of I/AS: we are interested in systems that autonomously make decisions critically important for humans. The method of application of ethical mechanisms in decision making is not significant. For example, ethical considerations may apply to evaluating a

S. O. Kuznetsov et al. (Eds.): RCAI 2020, LNAI 12412, pp. 61–70, 2020.
https://doi.org/10.1007/978-3-030-59535-7_5

particular decision or action. Evaluation of an action D can be determined by technical, legal, and moral considerations:

$$\text{Eval(D)} = \text{technical\_evaluation(D)} + \text{legal\_evaluation(D)} + \text{moral\_evaluation(D)} \quad (1)$$

Variations of the notorious trolley problem can be used as an illustration of such reasoning. and moral considerations can be presented as a kind of filter. The task of the latter is to make a choice among many alternatives. If the decision cannot be determined on the basis of technical and legal requirements, then some additional heuristics should be applied. These heuristics are ethical rules, which comprise the ethical behavior of I/AS. Conventionally, this can be represented as follows (Fig. 1):

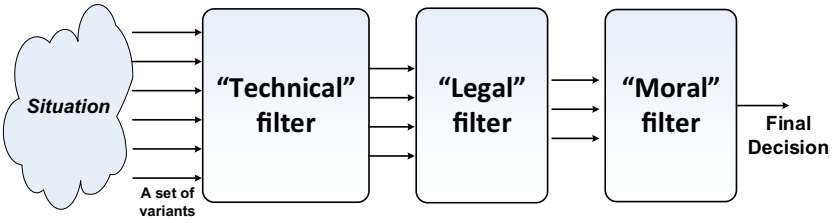

Fig. 1. Moral choice as a way to resolve ambiguity

Suppose that we can formalize the provisions of moral philosophy so that they can be represented by a certain system of rules (although this is a difficult task that requires a separate discussion). The problem then inevitably arises that, if the decision is based on moral principles (it is subjective, poorly verified, vague, etc.), then there is only one way to increase confidence in it: acceptance that the decision was made by a so-called moral agent, i.e., some entity to which we have delegated the right to apply ethical considerations. The question is then raised of whether there are prerequisites for I/AS to become a moral agent.

2 A Moral Agent

The basic definitions of the essence of a moral agent are usually anthropocentric and concise; as Parthemore and Whitby write, "by 'moral agent', we mean any agent that is appropriately held responsible for its actions" [2]. The reasoning is usually added that a moral agent acts in accordance with its role (see Mayo's work [3]), which speaks to freedom and is regarded as a necessary condition of a man being a moral agent, knowing that certain things are "right" or "wrong". Parthemore and Whitby state that a moral agent is necessarily a conceptual agent, i.e., an agent that possesses and employs concepts (units of structured thought), including the concept of "self".

We do not undertake to discuss the full list of properties that a moral agent should have, which is a purely philosophical problem that is compounded by the lack of consistent, constructive definitions of the basic concepts of ethics. Instead, we are interested

in the purely applied aspect of creating I/AS, and the behavior (decision-making) of the created I/AS should correspond to our general ideas about the behavior of a moral agent. In this case, we will rely on the assumption that a moral agent can be not only a person, but any entity, including an artificial agent. A human monopoly on moral issues has long been questioned (see, for example, de Waal [4]), and we take the next step, moving away from biological chauvinism altogether.

We postulate that the many manifestations of the properties of a moral agent are determined by three basic mechanisms. These are (1) the agent's possession of a world model in which there is an "I" component (cognizing subject), (2) a mechanism for comparing the observed other agent (conspecific) with the "I", and (3) the presence of an emotional-needs architecture of the lower-level control system. At the same time, we will try to show that all these components have a very practical, real embodiment in technical systems, and we will further consider how these mechanisms allow the realization of a number of behavioral phenomena inherent in a moral agent.

3 Phenomena and Mechanisms

3.1 Emotions and Needs

Let us start with the lower level of the organization of agents. The role of emotions in the formation of ethical norms—and how emotions determine the ethics of human behavior—is being actively explored by both philosophers and sociologists [5–7], and Marvin Minsky [8] suggests treating emotions as another way of thinking.

There is every reason to believe that emotions (on the physiological level) and temperament (on the psychic level) can be inherent in a technical system as purely pragmatic mechanisms that affect the success of an artificial agent in complex nondeterministic environments, see Karpov [9, 10]. In these works, emotions are viewed as a property of the control system that facilitates the realization of functions known in psychology as contrasting perception, behavior stabilization, state indicating, working in conditions of incompleteness of information, and so on [11–13].

We note here that, in the architecture of the control system, the reactions of the system—defined as emotional—are determined by positive feedback loops. These connections are responsible for estimating the situation and determining the magnitude of the emotional state according to Simonov's Information Theory of Emotions [11]:

$$\mathrm{E} = \mathrm{f}(\mathrm{N},\ \mathrm{p}(\mathrm{I}_{\mathrm{need}},\ \mathrm{I}_{\mathrm{has}})) \tag{2}$$

where E is emotion, its magnitude, and sign (quality); N is the strength and quality of the current need; $p(I_{need}, I_{has})$ is the assessment of the ability to satisfy a need on the basis of innate and acquired life experience; I_{need} is information on how to satisfy needs; and I_{has} is information on the means (resources) available to the agent that are required to satisfy actual needs. It is important here that the behavior of the agent (robot) is determined by its needs and emotional state.

Figure 2 illustrates an example of the basic architecture of the emotional-needs control system. An "emotional" agent is equipped with a set of simple sensors and solved a standard behavioral task, using some simple rules such as: "IF (hungry) THEN

(find food)", "IF (detect obstacle) THEN (run away)" etc. The influence of emotions on agent's behavior is realized as a positive form of feedback between the output signals (current actions or procedure) and behavior rules.

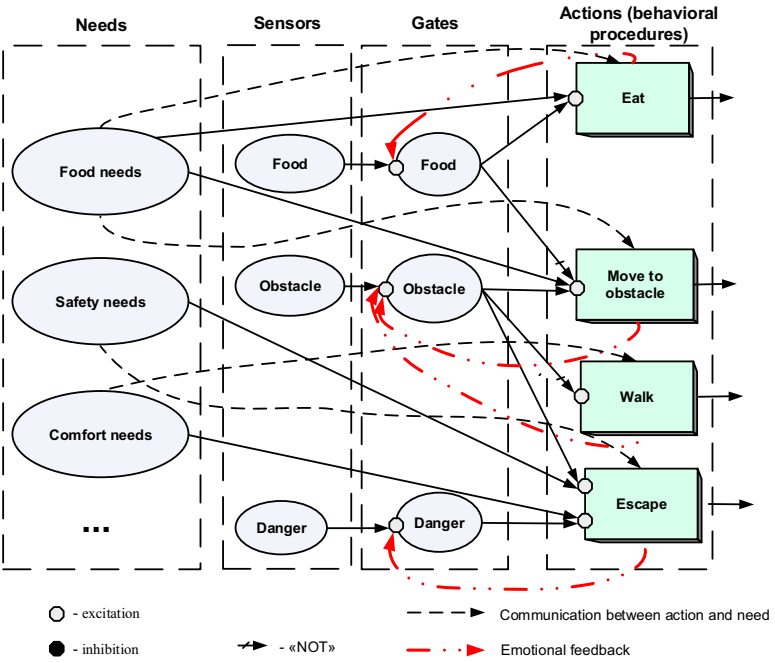

Fig. 2. Emotional-needs architecture.

An "Actions" block is a set of behavioral procedures. Every procedure is activated by signals from a "Needs" block and signals from special "Gate" elements. The "Gate" is an element that accepts direct signals from sensors and feedback signals from output elements. Every output procedure has its own emotional "weight". This signal is an input value for the gate element. It means that the positive emotion, associated with action a_i, ("Eat", "Walk", …) will cause an increase in the activity of this action (manifestation of the positive feedback loop). We emphasize that emotional-needs architecture is the physiological, basic, or reflex level of a control system. Here the main function of emotions is to stabilize behavior.

3.2 Model of the World, "I"

An important attribute of the management system of an intelligent agent is the availability of knowledge about the world around it. If a component called "I" (the subject of activity) is added to this model of the world, then we get what is called a picture of the world (PW) (see Osipov [14]). In a certain sense, PW can be considered as some kind of superstructure over the basic stimulus-reactive level. From an architectural point of view,

this is the component that implements the impact on the sensory system, determines the significance of certain needs, and, thus, changes the nature of the system's behavior, its goal-setting, etc. One of the most effective models for representing knowledge in PW is the symbolic or semiotic model, in which the main essence—the sign—is represented by four of its components: name n, percept p (image, form of expression), value m (method of use), and personal meaning a (goals, motives, personal meaning). In the model, homogeneous components each form a network, i.e., here we are dealing with four networks.

The following assumptions are important. We assume that the elements of the control system are equipped with one more confirming input—in addition to the exciting or initiating input—which is the input for the signal from the top of the "I". Thus, the action will not be activated if there is no confirmation signal from the "I", interpreted as the "belonging" of this action to the agent. In a certain sense, it is a feeling (sensation) of the self, i.e., identification or perception of an object as one's own. Without such a sensation, a mismatch of activity occurs in nature, such as the complex neuropsychiatric disorder called "alien hand syndrome", of which one clinical symptom is the presence of subjective sensations of the foreignness of a limb. From the point of view of semiotics, this means that these actions are the *meaning* of the sign "I", i.e., the question of conditionality is resolved in the most natural way.

The second assumption is that the activation of a component of a sign entails the activation of its other components, and associative connections arise between simultaneously active network nodes.

3.3 Imitative Behavior

This model quite naturally implements such phenomena as imitative behavior and social learning (learning by observing others). For example, let the agent know that the objects α_1 and α_2 are edible, i.e., belonging to the category of stimulus S (food) for action R (eat). Let the agent further observe that someone (conspecific) eats the object x, which was not previously considered by the agent as edible. Then, as a result of this observation, the agent will also classify this object as edible. A diagram for this is shown in Fig. 3.

In Fig. 3a, S_m is the component of the meaning of the "edible object" sign; R_m and R_p are the components of the meaning and perception of the sign "eating", respectively; and $Self_m$ and $Self_p$ are the values and percept of the sign "I", respectively.

Obs(A') and $Obs(R_{A'})$ are the results of observation: the agent sees that conspecific *A'* performs action *R*. Performing action R_m activates some motor function act (actually eating), and the execution of the procedure is accompanied by the issuance of some signal (*Signal*).

So, the animat sees that the agent *A'* performs some action *R* with respect to the object *X*. Moreover, *X* was not previously considered by the subject as a determining factor for the stimulus *S* (the *X*-*S* connection was not part of the subject's personal experience). Observation of the conspecific's actions leads to activation of the R_p sign percept. The presence of the percept-value relationship means the activation of the value element R_m: $Obs(R_{A'}) \rightarrow R_p \rightarrow R_m$. At the same time, the observed conspecific is compared with "I": the $Self_p$ percept is activated, which leads to the activity of $Self_m$: $Obs(A') \rightarrow Self_p \rightarrow Self_m$.

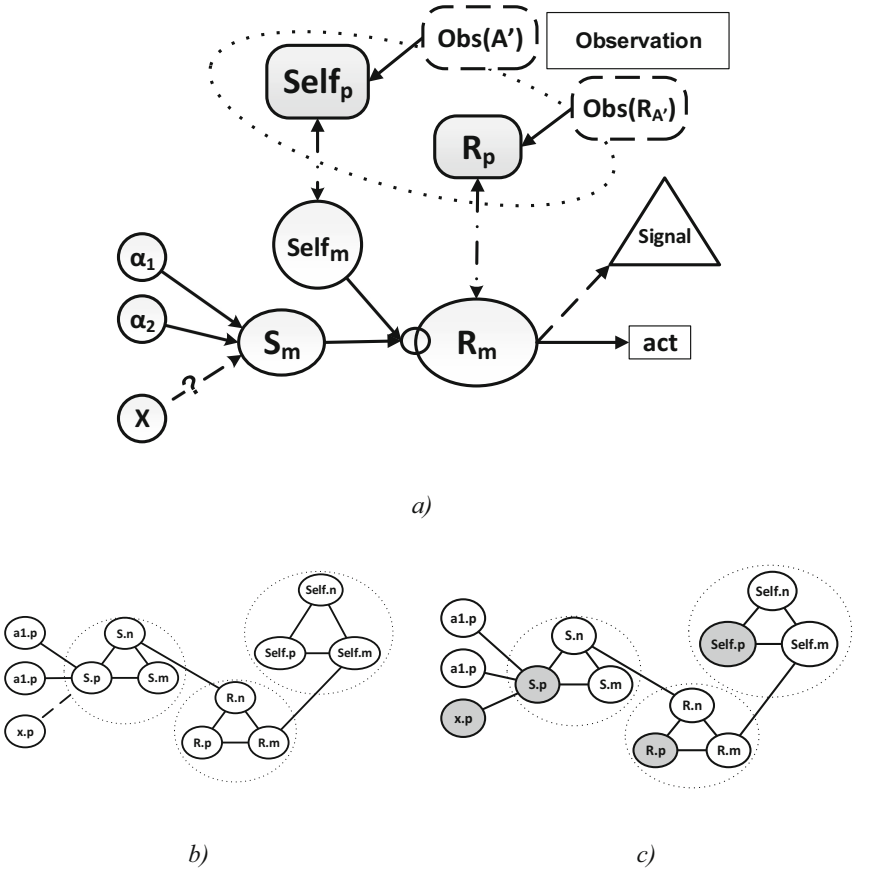

Fig. 3. The schema of imitative behavior. a) The conceptual scheme: solid lines are the relationships between value elements. The dashed-dotted lines are the connections between the components of the sign—the percept-value. b) The initial state of the system: all vertices are inactive, and there are a priori connections between the signs and the components of the sign. Communication S.p-x.p is optional (dotted line). c) The situation of monitoring the actions of the other agent.

This is the result of matching the conspecific with "I". Thus, all components of the circuit are in an excited state—*R*, *S*, and the actual object of observation, *X*. An associative connection is formed between *X* and *S*; that is, during observation, object *X* is included in the animat's behavioral experience, and this is done on the basis of observing the conspecific's behavior. This is imitative behavior; adding an evaluation element to this scheme then allows us to describe another phenomenon—the formation of reflex reactions, which are also formed on the basis of observation.

In this scheme, the most important point is the comparison of the subject with the conspecific ("I" and the other), determining the degree of their proximity. In nature, this identification is probably similar to what is called kin selection, when behavior determined by the degree of kinship of interacting individuals becomes evolutionarily beneficial (see, for example, Wilson [15]).

3.4 Empathy

This term refers to the ability to respond to the emotional states of surrounding individuals of varying degrees of proximity. Empathy is believed to determine the emotional propensity for collaboration and the manifestation of altruism. Of course, an individual's predisposition to empathy is a necessary, but not sufficient, property for the morality of the individual. Moreover, empathy is not a purely human property; in ethology, one of the mandatory mechanisms for the formation of the social interaction of individuals is the so-called sympathetic induction, the definition of which is identical to the definition of empathy. We are interested in two aspects of empathy: the mechanism of its implementation and the object of the empathy.

The realization of the empathy mechanism is possible on the same principle of identification (or determination of the degree of proximity) of the observed agent with the "I". The "formula" for empathy is quite simple and is determined by the components of the I/AS control system:

$$\text{Empathy} = \{\text{Emotions} + \text{Identification of the conspecific} + \text{Imitative behavior}\} \quad (3)$$

With the object of empathy, the situation is somewhat more complicated. In a certain sense, emotions are, first of all, a way of integrally assessing an individual's state (see [11]). Moreover, we assume that there is an external manifestation of the emotional state of agents, and it is significant here that empathy is the basis for a higher level of management related to goal-setting and planning. In terms of moral philosophy, this means the action of the golden rule: either implement a plan of action in which the other will feel good (increase the level of emotional state—a positive wording of the rule: "act in such a way…") or form a plan that does not lead to the appearance of negative conspecific emotions (negative wording: "do no harm"). In any case, the conspecific's emotional state influences the formation of the agent's behavior motive and goal. This is the main role of empathy from a technical point of view.

3.5 Characteristic Properties of Moral Agents

Next, we summarize some characteristic properties attributed to moral agents.

Language. Morality cannot exist without a symbolic language. The animal world does well without language; what is sometimes called a language is signal communication, i.e., the external manifestation of the internal state of the animal (see, for example, the work of Panov [16]). The role of signaling communication is certainly great, and many mechanisms of social behavior and interaction are built on this, including empathy, as discussed above.

Motive. The issues of motivating the behavior of a moral agent are considered so diverse in moral philosophy that we can find any convenient point of view (as, incidentally, on almost all other issues). For example, the view of John Locke is very convenient: personal interest is the only reasonable motive (cited in [17]).

Feelings. If, by feelings, we understand a certain process that reflects a subjective evaluative attitude to some objects, then those processes that occur in an emotional-needs architecture can rightfully be called feelings. A robot can really feel. If it is implied that a moral agent must possess "moral sentiment" (see, for example, [18]), then the task is simplified. So, we can completely abandon the consideration of this property and, at the same time, refer to Kant, who removed feelings from the realm of morality, considering their participation in the motivation of acts to be a prerequisite for the moral inferiority of the latter [19].

Alternatively, consider that, according to Hutcheson, moral feeling cannot directly motivate but is a response to a motive (cited in [17]). This also speaks to David Hume's assertion that moral feeling is the result of the action of simpler psychological principles—sympathy and the association of ideas. Any motives leading to human happiness are approved, and good consequences are indirectly experienced through sympathy, leading to a positive feeling about such motives.

Sympathy. This phenomenon is also the result of a comparison of the observed other agent (not even necessarily conspecific) with the "I". Naturally, the strength of sympathy depends on the degree of proximity of the observed agent. It should be noted that this can be considered as a direct consequence of the organization of a sign-oriented picture of the world. Excitation of some components of the sign (percept, value or personal meaning) leads to the excitation of its other components, including the name "I".

Responsibility. The moral agent is held responsible for its actions. Here, everything can turn out to be very simple. According to Parthemore and Whitby, a moral agent must possess certain key concepts and have the ability, over an extended period of interactions between the agent and its social and physical environment, to deploy those concepts appropriately [2]. This view of responsibility does not help either: "Thus, to be morally responsible for something, say an action, is to be worthy of a particular kind of reaction — praise, blame, or something akin to these — for having performed it", see M. Talbert [20].

Sometimes, the requirement for the independence of decisions expressed in judgments and actions is added to a personal responsibility for the consequences of decisions, where independence means that the subject should not act in accordance with a program laid down by someone else. However, this kind of reasoning usually—and quickly—becomes speculative. It is interesting that such a statement of the question of the boundary between what is laid down by nature and what is free will and independence is very rarely posed in a technical interpretation. There is usually a clear understanding that, on the one hand, there is some fixed, a priori specified part of the control system, and, on the other, that there are dynamically changing components. Consider the animat architecture; it clearly distinguishes the lower physiological, fixed, reflex level (on which, by the way, the emotional part of the control system works) and a superstructure thereof—the cognitive level, which is represented, for example, by a semiotic system.

So, we can at least state that many of the properties of a moral subject can be inherent in an artificial agent—and here, we are not talking about simulating mental or cognitive processes, properties of consciousness, and so on, but about dealing with purely technical solutions that are designed to increase the adaptive capabilities of a technical device. These decisions (models and mechanisms) can also be interpreted in humanitarian terms.

We have carefully avoided issues of moral philosophy; discussions about utilitarianism, evolutionary ethics, and even pragmatism are not within our competence. We have only tried to ask the question: if there is a certain list of properties that a moral agent should possess, are there reasons why we cannot recognize such an artificial agent—a robot?

4 Conclusion

The mankind comes to the idea that we are delegating intelligent/autonomous systems making independent decisions that are critical for people. If, in a situation of choice, both technical and legal arguments are exhausted, then moral criteria remain, and trust in such an "ethical" decision is possible only when the decision-making entity is a moral agent.

We emphasize once again that all the mechanisms described above were introduced exclusively for reasons of technical expediency, in order to solve the problem of creating effective adaptive mechanisms in three stages, by solving three classes of problems. At the first stage, these mechanisms should allow the technical device to act expediently in a complex, nondeterministic, dynamic environment. At the second stage, the task of organizing interaction within a group of agents was solved until the appearance of forms of social organization, and the formation of agent societies was also considered as an adaptation mechanism. The third stage is the task of purposefully managing social behavior, and, again, additional adaptation mechanisms were needed here, allowing society to maintain its stability. One of the most important factors of stabilization is the existence of mechanisms for resolving conflicts within society, and this is the main task and essence of morality.

These questions are not new to moral philosophy. For example, according to Drobnitsky, the essence of normative regulation is that "the action of social laws passes into the actions of individual agents" and thus "the social whole reproduces itself through individual mass behavior," and morality is a special case of this process [21, 19].

Today, there is intensive and fairly successful development of cognitive and social abilities of intelligent autonomous systems. However, in the field of ethics of I/AS behavior, promotion is fraught with a number of difficulties, and the main problem is the lack of constructive models that researchers expect from moral philosophy. Their absence often leads to models and methods remaining at the level of amateur understanding of moral problems.

Acknowledgments. This work was partially supported by the RFBR grant 17-29-07083-ofi_m.

References

1. UNESCO, "Report of COMEST on Robotics Ethics." UNESCO, COMEST, Paris, p. 64 (2017)
2. Parthemore, J., Whitby, B.: What makes any agent a moral agent? Reflections on machine consciousness and moral agency. Int. J. Mach. Conscious. **05**(2), 105–129 (2013)

3. Mayo, B.: The moral agent, in Royal Institute of Philosophy Lectures, pp. 47–63 (1968)
4. de Waal, F.: The Bonobo and the Atheist: In Search of Humanism Among the Primates. W W Norton & Co, New York (2013)
5. Neu, J.: An Ethics of Emotion? Oxford University Press, Oxford (2009)
6. Callahan, S.: The role of emotion in ethical decision making. Hastings Cent. Rep. **18**(3), 9 (1988)
7. Connelly, J.E.: Emotions and the process of ethical decision-making. J. South Carolina Med. Assoc. **86**(12), 621–623 (1990)
8. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and The Future of the Human Mind. Simon & Schuster, New York City (2006)
9. Karpov, V.E.: Emotions and temperament of Robots: Behavioral Aspects, Journal of Computer and Systems Sciences International, ISSN 1064-2307, Original Russian Text © V.E. Karpov, 2014, published in Izvestiya Akademii Nauk. Teoriya i Sistemy Upravleniya, 2014, No. 5, pp. 126–145., vol. 53, no. 5. Pleiades Publishing, Ltd., pp. 743–760, 2014
10. Karpov, V.: Robot's temperament. Biol. Inspired Cogn. Archit. **7**, 76–86 (2014)
11. Simonov, V.P.: Thwarted action and need – informational theories of emotions. Int. J. Comp. Psychol. **5**(2), 103–107 (1991)
12. Ilyin, E.P.: Emotions and Feelings. Piter, Saint-Petersburg (2001). (in Russian)
13. Rai, M., Yadav, R.K., Husain, A.A., Maity, T., Yadav, D.K.: Extraction of facial features for detection of human emotions under noisy condition, no. September, pp. 49–62 (2018)
14. Osipov, G.S., Panov, A.I., Chudova, N.V., Kuznecova, J.M.: Znakovaja kartina mira sub'ekta povedenija (Semiotic view of world for behavior subject). Fizmatlit, Moscow (2018)
15. Wilson, E.O.: Genesis: The Deep Origin of Societies (2019)
16. Panov, E.: Evolution of dialogue. Communication in development: From microorganisms to humans. Moscow (2014). (in Russian)
17. Darwall, S.: The foundations of morality: virtue, law, and obligation. In: Rutherford, D. (ed.) The Cambridge Companion to Early Modern Philosophy, pp. 221–249. Cambridge University Press, Cambridge (2007)
18. Smith, A.A.: The Theory of Moral Santiments, 6th ed. MetaLibri (2006)
19. Apresyan, R.G., Artemyeva, O.V., Prokofiev, A.V.: The phenomenon of moral imperative. Critical essays. Institute of Philosophy, RAS, Moscow (2018). (in Russian)
20. Drobnitsky, O.: The concept of morality. Historical and critical essay. Science, Moscow (2007). (in Russisn)