# EMOTIONS AND A PRIOR KNOWLEDGE REPRESENTATION IN ARTIFICIAL GENERAL INTELLIGENCE

Andrey Gavrilov

*Abstract: In this paper a prior knowledge representation for Artificial General Intelligence is proposed based on fuzzy rules using linguistic variables. These linguistic variables may be produced by neural network. Rules may be used for generation of basic emotions – positive and negative, which influence on planning and execution of behavior. The representation of Three Laws of Robotics as such prior knowledge is suggested as highest level of motivation in AGI.*

*Keywords: Emotions, neural networks, knowledge representation, hybrid intelligent systems.*

*ACM Classification Keywords: I.2 Artificial Intelligence – General – Cognitive simulation*

*Conference: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

## Introduction

Most sufficient problem in Artificial Intelligence is development of AI functionally similar to human mind. In last time this problem is especially actual in accordance with growth of research and development in intelligent robotics, in particular, humanoid robots. Artificial Intelligence oriented on solving of all tasks by human-like way is named as Artificial General Intelligence [Goertzel and Pennachin, 2007].

Most important issues for development of Artificial General Intelligence are:

- to figure out role and mechanism of introducing of emotions in process of thinking,
- to figure out is it needed to use a prior knowledge and if yes then how to implement one,
- how to implement confabulation, i.e. thinking about future, and connection it with planning and actions,
- how to realize consciousness and thinking about itself,
- how to provide friendly thoughts and behavior of AGI with respect to human beings.

In last time there were published many hypothesis and theories about these [Arbib, 1972], [Pribram, 1971], [Wermter and Sun, 2000], [Hawkins and Blakeslee, 2004], [Hoya, 2005], [Wang, 2006], [Wang and Wang, 2006], [Minsky, 2006], [Hecht-Nielsen, 2006]. Unlike these works we try to connect in one system concepts of emotions, a prior knowledge and friendly to human being behavior.

In this paper we propose the model of AGI based on performance about basic emotions (positive and negative), hybrid architecture consisting of rule-based a prior knowledge representation with linguistic variable (LV) and neural network producing values of LV. In section 1 we are talking about role of emotions in AGI. In section 2 we propose language for representation of a prior knowledge. In section 3 we suggest implementation of Three Laws of Robotics of A. Asimov in proposed language as highest level of motivation in AGI. This a prior knowledge may be expanded by adding rules for logical recognition of “good” or “bad” situations with inputs as result of preprocessing and classification by neural network.

In contrast to another attempts to use the emotions for simulation of mind (for example, in [Goerke, 2006]) we suppose that we have just two main emotions (positive and negative). Many different more specific emotions known from psychology are definition of combination of basic emotions with expression of ones by mimics and other elements of behavior oriented on communication between human beings or with state of organism. Also we suppose that these basic emotions must be produced by a priory determined rules on high level and preprocessing by neural networks on low level. Unlike to these

## Role of emotions

A role of emotions in our mind is very wide. There is not definite performance about emotions. It is explained probably by different view on emotions in different sciences. For example, psychologists basically are interesting external expression of emotions in communications between human beings, but it is not enough for developers of AI. For them it is most interesting implementation of emotions as internal states and its influence on behavior. It is possible to look on emotions from different points of view, e.g. influence of emotions on attention, acceleration of decision making, connection between emotions and metabolism, usage of emotion for communications and so on. In this paper we will focus on connection of emotions with motivation or planning.

We suppose that emotions and motivations are very close. Moreover motivation is based on emotions and most sufficient reason of any activity is attraction to positive emotions and avoidance of negative emotions. Thus we have just two general emotions – positive and negative. All other emotions are kinds of these basic emotions with any nuances as result of influence of state of organism (system) and features of interaction with another person.

We believe emotions influence on both selection of goal and achievement of it. The successful process of achievement of goal is reason of positive emotions and present of any unexpected obstacles is the reason of negative emotions.

In figure 1 our performance about connection between perception and generation of emotion is shown. We especially did not write in details unit “Perception and decision making” because this one may be implemented by different ways with hybridization of different paradigms, e.g. neural networks, reinforcement learning, fuzzy logics and so on, and this problem exceeds the bounds of this paper.

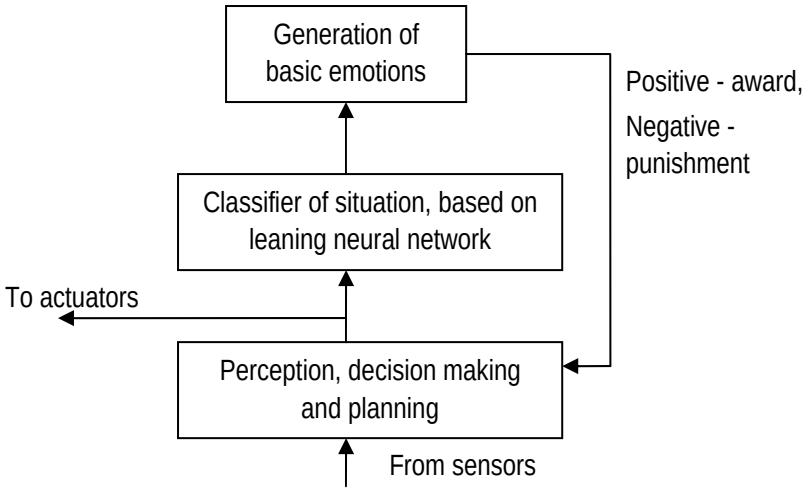


Figure 1. The scheme of connection between perception and emotions

Generation of basic emotions (positive or negative) is based on recognition of “good” and “bad” situations, which may be realized by learning neural network. We suppose that the negative emotions are stimulus for changing of behavior or searching of new plan, whereas the positive emotions are not such crucial for execution of planned behavior. The positive emotions are used as award in any kind of reinforcement learning for storing successful behavior. Produced by such way basic emotions may be stored in associative memory together with any pattern describing an image or situation and may be used for building of plan to avoid or attract to this pattern.

Thus in our performance the highest level of motivation for behavior and control of behavior is the generation of basic emotions - negative and positive.

## Representation of a prior knowledge in AGI

It is interesting question how are related in mind learning and a prior knowledge. In beginning of AI this question was not enough actual because developers are dealing with simulation of symbolic intelligence as a prior knowledge and usage of this knowledge for solving of tasks. But later when it was clear that learning by interaction with environment is general property of real intelligence, another extreme approach sometimes is occurred that the mind has not a prior knowledge. But we think that together with development of learning with producing of categories, concepts, rules from learnt neural networks it is needed to provide an opportunity to introduce in AGI a prior knowledge as rules for control of robot at determined situations, for example, like in experiments with simulated mobile robot controlled by hybrid neural network [Gavrilov and Lee, 2007]. In this model of robot the behavior is determined by rules in confusion with obstacle, whereas the movement was controlled by neural network in situations when robot is enough far from obstacles in front. However this is enough primitive example of combination of prior knowledge and learning. For implementation of AGI it is needed more complex representation of a prior knowledge with uncertainty. And a measure of uncertainty of concepts used in rules may be produced by neural networks as result of learning.

We propose follow simple language for representation of a prior knowledge. It is described lower by grammar in BNF notation.

< Rule > ::= If < Antecedent > then < Consequent >

< Antecedent > ::= < Fuzzy symbolic value > | < Function > (<Fuzzy symbolic value >) | < Condition >
| < Antecedent > and | not < Antecedent >

< Condition > ::= < Fuzzy symbolic value > < Relation > < Operand >

< Relation > ::= = | < | > | ≤ | ≥ | ≠

< Operand > ::= < Fuzzy symbolic value > | < Constant >

< Consequent > ::= < Action >

< Function > ::= Increase | Decrease

Here “Fuzzy symbolic value” means value of linguistic variable [Zadeh, 1975]. In this grammar we did not describe the terminal symbols < Fuzzy symbolic value > and < Action > (names of fuzzy variables and actions) depending on concrete implementation of intelligent system. In next section we suggest example with concrete terminal symbols.

Semantics of this grammar is partially determined by fuzzy logics. It means that every fuzzy variable is determined by confidence factor (member function) between 0 and 1, consequent inherits confidence factor of antecedent. But confidence factor of conjunction is calculated as sigmoid function from sum of confidence factors of members of conjunction. Such model of rule is similar to model of neuron in feed forward neural network. Thus if we will not use <condition> as antecedents we can obtain rules very easy extracted from trained neural network where any symbolic value of fuzzy variable is corresponding to output of any neuron and value of output is equal the value of membership function for this symbolic value of fuzzy variable.

We suppose that AGI works in discrete time with any step. Functions “increase” and “decrease” means corresponding updating of value of membership function in compare with previous step.

## Three laws of Robotics

There is well known problem to provide friendly to human beings behavior of AGI in robotics. We propose to overcome this problem using implementation of Three Laws of Robotics of A. Asimov as prior knowledge producing positive and negative emotions based on recognition of fuzzy concepts “what is bad” and “what is good” for human beings and robot. These original laws are following:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Let any classifier of situations provides recognition of follow classes: *Danger_for_man, Danger_for_myself, Command_executed.*

Every of these classes may be viewed as variable with confidence factor or value of linguistic variable with any corresponding value of membership function. It means that if we use, for example, multilayer perceptron as classifier then it has three outputs and value of these outputs may be used as confidence factors.

Variable *Command_executed* shows success of execution of last command got from human being. May be supposed that confidence factor of it is rising during execution of plan.

Let use as actions (< Action > in grammar) the generation of positive and negative emotions “Negative_emotion” and “Positive_emotion”. Then Three Laws of Robotics may be written by rules as:

If *Danger_for_man* then Negative_emotion;

If *Danger_for_myself* and not *Danger_for_man* then Negative_emotion;

If not *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then

Negative_emotion;

If *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then

Positive_emotion;

Note that same outputs of these rules may be accumulated in accordance with rules of fuzzy logics.

If we want to get more strong statements of these laws we can append the rules simulating processing of derivative of input variables:

If Decrease(*Danger_for_man*) then Positive_emotion;

If Decrease(*Danger_for_myself*) and not *Danger_for_man* then Positive_emotion;

If Increase(*Command_executed*) and not *Danger_for_man* and not *Danger_for_myself* then Positive_emotion;

Of course this mechanism of implementation of Laws of Robotics does not guarantee friendly behavior of AGI because it is possible incorrect learning of robot to classify situations. However, it is impossible to implement these Laws appropriately and surely for every case in life. For more reliability it is possible to formulate by experts a prior high level knowledge for classification in proposed language. In this case rules based representation of Laws is just minimal knowledge base which may be expanded by rules for estimation of different situations and cases. In this case we can speak about implementation of rules-based logical level of thinking and implementation of associative thinking by neural networks.

Basic positive and negative emotions produced by rules as described above may be used for decision making (selection of action) and planning as award and punishment respectively.

Such implementation of Laws can provide human-like behavior unlike fully determined rules in which it is impossible to clear definite such concepts as “harm”, “danger” and so on. It means that we have not guarantee that robot always will be demonstrate friendly behavior like we see for human beings. His behavior basically depends on careful training.

## Conclusion

In this paper we suggest novel performance about implementation of Laws of Robotics as representation of a prior knowledge producing positive and negative emotions influencing on planning and execution of behavior. The language for this representation based on rules and linguistic variables is proposed. The source of information (values of linguistic variables) for rules describing of Laws is classification of situations as result of perception by neural network. The feature of proposed language is an easiness to connect rules with neural networks and to extract rules from trained neural network. It provides the opportunity to grow the rule-based part of AI if it is needed by introducing of expert knowledge or fixing of knowledge obtained by previous training as a prior unchangeable knowledge, in particular, highest knowledge about classification "what is good and what is bad" needed for Laws of Robotics.

Positive and negative emotions produced by rules may be used for planning as awards and punishments respectively for reinforcement learning [Sutton and Barto, 1998].

In future work we plan to implement the architecture of AGI in simulated mobile robots and in characters in game based on proposed in this paper mechanism of representation of prior knowledge and generation of basic emotions.

## Bibliography

[Arbib, 1972] M.A. Arbib. The Metaphorical Brain. Wiley-Interscience, 1972.

[Gavrilov and Lee, 2007] A.V. Gavrilov, S,-Y. Lee. Usage of Hybrid Neural Network Model MLP-ART for Navigation of Mobile Robot. Proceedings of International Conference on Intelligent Computing ICIC'07, China, August, 2007, LNAI 4682. Springer-Verlag, Berlin, Heiderberg, 2007, 182-191.

[Goerke, 2006] N. Goerke. EMOBOT: A robot Control Architecture based on Emotion-like Internal Values. In book: Mobile Robots, Moving Intelligence (eds. J. Buchli). ARS/pIV, Germany, 75-94.

[Goertzel, 2006] Ben Goertzel. Patterns, hypergraphs and embodied general intelligence. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 451-458.

[Goertzel and Pennachin, 2007] Ben Goertzel and Cassio Pennachin. Artificial General Intelligence. Springer-Verlag, Berlin, Heidelberg, 2007.

[Hawkins and Blakeslee, 2004] J. Hawkins and S. Blakeslee, On Intelligence, Times Books, 2004.

[Hecht-Nielsen, 2006] R. Hecht-Nielsen. The mechanism of thought. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 1146-1153.

[Hoya, 2005] Tetsuya Hoya. Artificial Mind System. Kernel Memory Approach. Springer, Berlin, 2005.

[Minsky, 2006] M. Minsky. The emotion machine. Simon&Shuster, New York, 2006.

[Pribram, 1971] K.H. Pribram Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology: Prentice Hall/Brandon House, N.Y., 1971.

[Sutton and Barto, 1998] Sutton R.S., Barto A.G. Reinforcement learning: An introduction. – MIT Press, Cambridge, MA, 1998.

[Wang, 2006] P. Wang, Rigid Flexibility: The Logic of Intelligence, Springer, 2006.

[Wang and Wang, 2006] Yingxu Wang, Ying Wang. Cognitive Informatics Models of the Brain. IEEE Trans. on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 36(2), 2006, 203-207.

[Wermter and Sun, 2000] Stefan Wermter, Ron Sun, Hybrid Neural Systems. Springer-Verlag, Heidelberg, Germany. 2000.

[Zadeh, 1975] L.A. Zadeh, The concept of linguistic variable and its application to approximate reasoning, Inform. Sci. I: 8, (1975), 199-249

## Authors' Information

*Andrey Gavrilov – PhD, Associate Professor, Novosibirsk State Technical University, Karl Marx av., 20, Novosibirsk, 630092, Russia; e-mail: andr_gavrilov@yahoo.com*