PRIA JOURNAL SPECIAL ISSUE XXI NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE WITH INTERNATIONAL PARTICIPATION (CAI-2023)/SECTION 7 "APPLIED BIOMEDICAL SYSTEMS"

Social–Emotional Conversational Agents Based on Cognitive Architectures and Machine Learning

A. A. Dolgikh[a,*], A. V. Samsonovich[a,], and D. V. Tikhomirova[a,***]**

[a] *National Research Nuclear University MEPhI, Moscow, 115409 Russian Federation*

* *e-mail: tolick.dolgih2013@yandex.ru*

** *e-mail: alexei.samsonovich@gmail.com*

*** *e-mail: dvsulim@mail.ru*

Abstract—Large language models can recognize the topic of arbitrary utterances, as well as their emotional coloring, but do not understand the logic of emotions, even though they can often generate adequate responses in a given context. On the other hand, cognitive architectures such as eBICA are capable of modeling the dynamics of emotional states in the general case but require assistance in understanding the meaning of statements and generating responses to them. This paper presents a way to integrate large language models and eBICA, allowing them to complement each other. The "Virtual Receptionist" and "Virtual Psychodiagnostician" paradigms have been studied, and the results are encouraging.

Keywords: large language models, cognitive architectures, emotional intelligence, artificial intelligence

DOI: 10.1134/S1054661824700640

INTRODUCTION

Intelligent conversational agents, capable of maintaining a conversation with a user and possessing elements of emotionality, are playing an increasingly important role in the life of society, and the demand for them is growing rapidly. Examples of their applications include a personal interlocutor agent [1]; a virtual tutor [7]; a rehabilitation conversational agent [4]; a robot storyteller, which stimulates the development of creative abilities in children [5]; a virtual patient for training doctors [3]; and others. At the same time, existing conversational agents (including Siri, Alice, ChatGPT, etc.) lack the socioemotional intelligence that would allow them to be accepted by a person on a social level.

The problem to which this work contributes is the creation of an artificial social agent, which, in addition to performing functions in its subject area, is capable of (a) establishing and maintaining contact with the user at a social level compatible to a human and (b) internally evaluating the user socially to work with him/her more effectively. For certain applications, these abilities are critical.

Today, social conversational agents are implemented either on the basis of statistical machine learning methods, primarily large language models (LLM) based on deep neural networks, or on the basis of cognitive architectures such as ACT-RE, extended Soar, EMA [9], eBICA [13, 14] and others. In both cases, difficulties arise in choosing appropriate behavior in an unforeseen social context. Cognitive architectures have to be created manually, and it is impossible to foresee all possible situations when they will be used. To train statistical models, large amounts of data are needed, which in the field of social decision making are usually difficult to access.

In this work, a combined approach is developed: namely, a way to integrate machine learning and cognitive modeling methods, allowing them to complement each other. The result is demonstrated using the "Virtual Hotel Receptionist" and "Virtual Psychodiagnostician" paradigms.

1. eBICA COGNITIVE ARCHITECTURE PRINCIPLES

The cognitive architecture of eBICA, described repeatedly (for example, [13, 14]), is not a large embodied software environment with its own programming language, like Soar, Act-R, Icarus, etc. We deliberately took a different path, developing eBICA as a theoretical framework that today has dozens of specialized implementations but is constantly changing and expanding, adapting to new tasks and subject areas. Such flexibility would not be possible if implemented as a single large system like Soar. At the same time, in this case, it is the general framework and theoretical understanding of the principles of the solution

Received December 15, 2023; revised January 10, 2024; accepted January 23, 2024

that remain fixed and are of the main value for practice, making it possible to create a technological line for the development of socioemotional agents for a variety of applications.

Key elements of eBICA, in addition to the typical set of memory systems for cognitive architectures, are semantic maps, mental perspectives, and moral schemas [14]. The principle of operation of the moral schema is to determine some "normal" situation in a given social environment and to use the means available to the agent to maintain or achieve such a "normal state." A minimal simplified explanation of the basic principles of the eBICA theory is given in the following paragraphs.

The first step is to introduce appraisals for all significant events and actors operating in the environment. Appraisals are entered as vectors that take values in a semantic space defined by a set of semantic scales relevant to a given paradigm (for example, in the VAD model these are Valence, Arousal, Dominance: that is, "positivity," "excitability," and "dominance"). Let X and Y be vectors of appraisals of two actors, also designated by the letters X and Y. Suppose the actor X performs a discrete action a in relation to the actor Y. Let a and a^+ are appraisals of action a, defined as the expectation of the effect of a on X and Y, respectively. The following rule for updating appraisals of X and Y upon performing an action a is postulated:

$$X := (1 - rw)X + rwa, \qquad (1)$$

$$Y := (1 - rw)Y + rwa^+, \qquad (2)$$

Here r is a constant, model parameter, with $0 < r \ll 1$, a w is a positive value characterizing the significance of the action a. Thus, each action is characterized by two vectors a, a^+ and scalar w, and these values may depend on the context.

The configuration of the moral schema in eBICA is determined by feelings towards the actors—participants in the schema. The feeling concept F_X is defined as the appraisal values of actor's X in normal conditions [14]. Then it can be shown that to maintain a normal state, it is sufficient for the actor to choose actions whose appraisals are as close as possible to the values of the corresponding feelings and are symmetrically distributed relative to them. In a state close to normal, feelings are fixed. They can change in a conflict situation, when the discrepancy between appraisals and feelings is significant and cannot be eliminated by choosing acceptable actions with fixed feelings. The conditions for activation and deactivation of moral schemas are described in [14]. While appraisals and feelings are related to the value system, physiological drives related to the somatic area also play a significant role in the choice of behavior.

Appraisals and feelings may be different in the mental perspectives of different actors. However, it was found experimentally [13, 15] that equations (1), (2) together with the laws of moral schemas [14] make it possible to calculate the dynamics of interpersonal emotional relationships and thus predict the mental states of agents based on their observed behavior.

In this work, eBICA was used to generate datasets for training neural network components, as described below. This allowed the agent to adequately respond to user behavior.

2. EXPERIMENTAL PARADIGMS

Within the framework of the hotel virtual receptionist paradigm, the following model situation is defined: a visitor comes to check into a hotel, and a dialogue takes place between him and the receptionist at the reception desk. The subject of consideration here is the dynamics of emotional interaction between the user and the virtual registrar, and the role of the latter can be played by both a virtual actor (bot) and a hidden living person (confederate).

The research questions are: Can a virtual actor be considered socially acceptable to the user? Trustworthy? Attractive? Capable of evoking emotions? Having individuality? Is it possible to determine the characteristics of a participant's personality by his/her behavior in this paradigm? As a result, can we say that the virtual actor in this paradigm establishes socioemotional contact at the human level?

The motivation for choosing the "Virtual Psychodiagnostician" paradigm can be explained as follows. Currently, there are many questionnaires through which the users can find out their personality characteristics (for example, [16]), but for this they needs to answer a large number of boring questions. The "Virtual Psychodiagnostician" paradigm is a game in which the user is invited to take part in a dialogue taking place in an imaginary real-life-like situation (for example, ordering lunch in a restaurant, choosing furniture in a store, etc.) to find out their personality type. The main advantages from the user's point of view are accessibility, ethics, fun, and ease of taking the test under the condition of anonymity. The research questions in both paradigms are the same.

3. SYSTEMS DEVELOPMENT AND IMPLEMENTATION

3.1. Virtual Receptionist

Structurally, the virtual receptionist can be divided into two components: a graphical interface implemented using virtual reality tools and a server component that manages the dialogue with the user. This structure helps to separate the visualization of the reg-

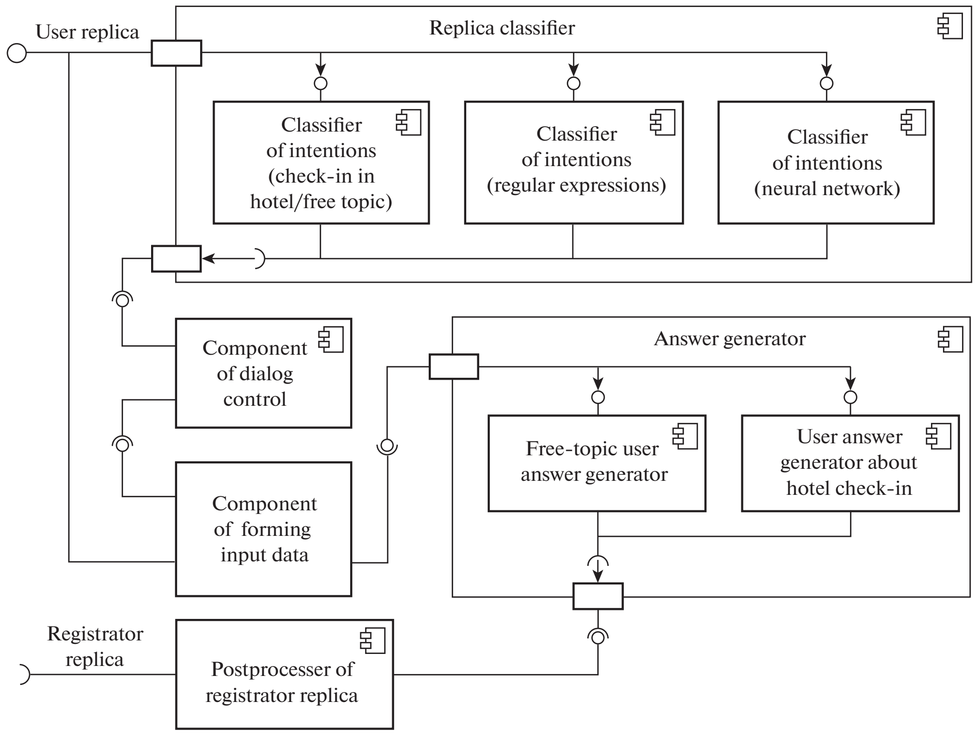

Fig. 1. Structure of the server part of the virtual registrar.

istrar from the construction of the dialogue with the user, which makes it possible to integrate the server not only with the virtual environment but also with other means of interaction with users.

The server part (Fig. 1) is the main part of the virtual recorder. It includes a component for classifying user remarks, a component for generating registrar responses, and a component for processing remarks (the latter performs the function of automatic text editing).

The classification component calls multiple classifiers sequentially. First, it is necessary to determine what type of dialogue this utterance belongs to. Within the framework of the “Virtual Receptionist” paradigm, two types of dialogue are distinguished: dialogue about checking into a hotel and dialogue on a free topic. Since the list of possible phrases is significantly narrowed within the dialogue about checking into a hotel, this classification is performed by searching for keywords in the user’s response. Next, some specific details need to be classified, such as the user’s last name, first name, and patronymic. To solve this problem, a rule-based classifier was used: the presence of a name in a prepared a priori list was checked.

Finally, classification based on user intent was performed. For this purpose, a neural network classifier was used. To detect intent based on the user’s phrase, the *rubert–base–cased–sentence* model was additionally trained, which is included in the library *deeppavlov*. This model is built on the BERT architecture. To train the classifier, a dataset of 231 user utterances was created, divided into 5 categories: greeting, ending a dialogue, booking a room, getting a room, general questions. Cross-entropy was used as the loss function. The calculated metrics after additional training of the neural network for five epochs are as follows: F1(macro) = 0.8902, F1(micro) = 0.8972, F1(weighted) = 0.8922.

The next important component is the receptionist response generator. It was implemented using two neural networks. One for generating dialogue on a free topic, and the other for generating dialogue on the topic of checking into a hotel.

The neural network for generating the receptionist’s response to the topic of checking into a hotel is based on the Transformer architecture. To implement the network, the following parameters were chosen:

• embedding size (regular and for word position): 256;

• number of heads in the attention mechanism: 8;

• number of encoder blocks: 6;

• number of decoder blocks: 6;

• the size of one piece of data arriving at the input of the neural network (batch): 128.

To obtain the resulting sequence of tokens, a greedy decoding algorithm was used. The peculiarity of the algorithm, as can be judged from its name, is that at each generation step, the token whose probability is

Fig. 2. Examples of gestures performed by the F-2 robot.

the highest at this step is selected. This algorithm works well for generating small sequences of words where the sentence is short.

The total number of trained parameters was 5686725. Note that to train the neural network, 10000 dialogues were generated using an algorithm built on the basis of the eBICA cognitive architecture (see Section 1). This cognitive model worked with a prewritten dialogue script that allowed for slight variations. The variations consisted in the choice of options for statements from both the user and the registrar. Each version of each statement was a priori rated by another group of participants on the Valence, Dominance, and Arousal scales in a separate empirical study. As a result, eBICA could determine the choice of one of several options for each utterance for each virtual interlocutor, guided by a model of developing social relations between the interlocutors. Thus, a set of synthetic dialogues was obtained. Each dialogue was divided into pairs <user phrase, registrar phrase>. After this, the neural network was trained on a dataset consisting of such pairs for 20 epochs.

To generate answers to questions on free topics, a pretrained neural network was used, trained on datasets compiled from the Russian dialogue corpus (based on the Russian DialoGPT model). Due to the large number of parameters in this network, it was not fine-tuned.

The functions of the virtual environment visualization component include converting an utterance received from the user into text, and an utterance received from the server into a voice message, as well as exchanging data with the server.

3.2. Virtual Psychodiagnostician

The system that implements the Virtual Psychodiagnostician was described by us earlier in a position paper [2]. Here we present preliminary results of its implementation, which is still at an intermediate stage. In particular, eBICA has not yet been used to control the behavior of a psychodiagnostician, although the possibility and feasibility of its use here are quite clear.

The virtual psychodiagnostician conducts a conversation with the user according to several specified scenarios (see Section 2), addressing the user with prewritten questions. The use of the F-2 robot, developed in the laboratory of Kotov at the Kurchatov Institute, makes it possible to voice text and express specified gestures along with other nonverbal signaling. Robot F-2 [17] can potentially express a large number of emotions and communicative intentions. A software interface based on BML (Behavior Markup Language) makes it possible to work with a database of patterns of emotional dynamics including elements of body language, gesture, gaze, facial expression, and speech, that had to be calibrated on emotional scales. For the purposes of the project, we, with the help of participants, assessed the robot's gestures on emotional scales (the limited space of the article does not allow us to present the results of these assessments). Examples of gestures are shown in Fig. 2. The robot is also available as a virtual avatar.

The main difficulty in this subproject was the analysis of user utterances to determine the user's personality type. The problem was solved using machine learning methods. Today, machine learning algorithms are used to create models that classify personality types according to the Big5 model based on data from questionnaire responses, as well as large-volume free text (20000 words) written by a participant [10], but access to this proprietary system is limited.

Our task here is to train the system to recognize personality types based on dialogues within selected scenarios. Due to the lack of large volumes of data of the required type, training was carried out on the basis of an open access dataset, including the results of the Big5 test of individual participants and the text written by them [8]. The scikit-learn library for Python was used to implement machine learning algorithms. For data preprocessing, the open access NRC Word-Emotion Association Lexicon database was used [11, 12], consisting of 14155 words with ratings for categories of valence (negative tonality, positive tonality) and emotions, such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust. We also used the GloVe embedding model.

4. EXPERIMENT: MATERIALS AND METHODS

Sixteen college students at National Research Nuclear University MEPhI, all native Russian speakers, age 20–22 males and females in equal proportion took part in this study. Testing of the implemented systems confirmed their compliance with the functional requirements and objectives of the study.

4.1. Virtual Receptionist

The experimental procedure in the "Virtual Hotel Receptionist" paradigm included three stages.

At the first stage, the participant in the experiment assessed the appearance and environment of the registrar agent in the virtual space for two minutes, after which he/she completed the Artificial Social Agent Questionnaire (ASA) [6], answering questions about the first impression of the agent.

The ASA questionnaire is a tool for assessing human interaction with a virtual agent, developed by an international working group of researchers in the field of artificial intelligence. We used a short version of the questionnaire, which allows us to create a profile of an artificial social agent on five scales: the attractiveness of the agent, its social acceptability, trust in the agent, the presence of personality in the agent, and the ability to evoke emotions in the user.

The receptionist avatar in this study could be controlled either by a virtual actor or by a human confederate hidden in another room, who entered the text of his questions on the keyboard. Two participants were alternately used in the role of the confederate. The rest of the participants in the experiment did not know that a confederate could be used.

At the second stage, the participant underwent the registration procedure in virtual reality. The registrar's questions were voiced to the participant through speech synthesis, the participant speech was recognized—a dialogue took place, as described in Section 3. The experiment lasted until the registration procedure was completed. If the registration procedure did not complete within 10 min, the experiment was terminated. At the end of the second stage, the participant again completed the short version of the ASA questionnaire in full, rating the receptionist after the interaction.

At the third stage, the participant went through the same registration procedure again in virtual reality. The third stage differed from the second only in that at the second stage the registrar was controlled by a confederate, while at the third stage it was controlled by a virtual actor, or vice versa, if at the second stage the registrar was controlled by a virtual actor, then at the third stage it was controlled by a confederate. Upon completion of the third stage, the participant again completed the questionnaire.

4.2. Virtual Psychodiagnostician

The experiment in the "Virtual Psychodiagnostician" paradigm took place in the form of a conversation between the participant and the robot according to a prewritten script, as described above. After a few general questions had been asked by the robot, the participant and the robot played a game in the form of a dialogue (Section 2). Note that the phrases voiced by the robot were accompanied by gestures, facial expressions, gaze, and body language. For each phrase, a gesture was selected that corresponded to the intention and emotional connotation of the phrase, according to the previously established emotional connotations of gestures (see Subsection 3.2).

5. RESULTS AND ANALYSIS

5.1. Virtual Receptionist

As a result of the study, psychological profiles of a virtual registrar were obtained and compared before and after communication, as well as the profiles of a registrar controlled by a virtual agent and a registrar controlled by a human. It was found that the impression of a virtual receptionist is formed gradually, and before communicating with him, the impression is determined mainly by the characteristics of appearance. After communicating with the registrar, the assessment is influenced by the registrar's behavior.

Figure 3 presents the results of a comparison of the profiles of a registrar controlled by a virtual agent before and after communication with the participant, as well as profiles after communication with a registrar controlled by a human confederate.

The darkened stripes in Fig. 3 display the characteristics of the virtual agent. Significant differences are indicated with asterisks.

Calculations were carried out using the Mann–Whitney U test with account for the Bonferroni correction. The following was found. The virtual actor's ability to evoke emotions increased upon interacting with him ($P < 0.001$). The assessment of the presence of individuality for the virtual actor was higher than for the human confederate ($P < 0.0095$).

5.2. Virtual Psychodiagnostician

In the case of the Virtual Psychodiagnostician, the trained system was used to analyze real dialogues conducted by the Virtual Psychodiagnostician in the guise of the F-2 robot with the participant. For each of the five scales, a binary score was obtained based on dialogue analysis. The results were compared with the results of the standard Big5 test in its Russian adaptation, which participants took independently at a different time, also divided into two categories for each scale. The comparison results are presented in Fig. 4.

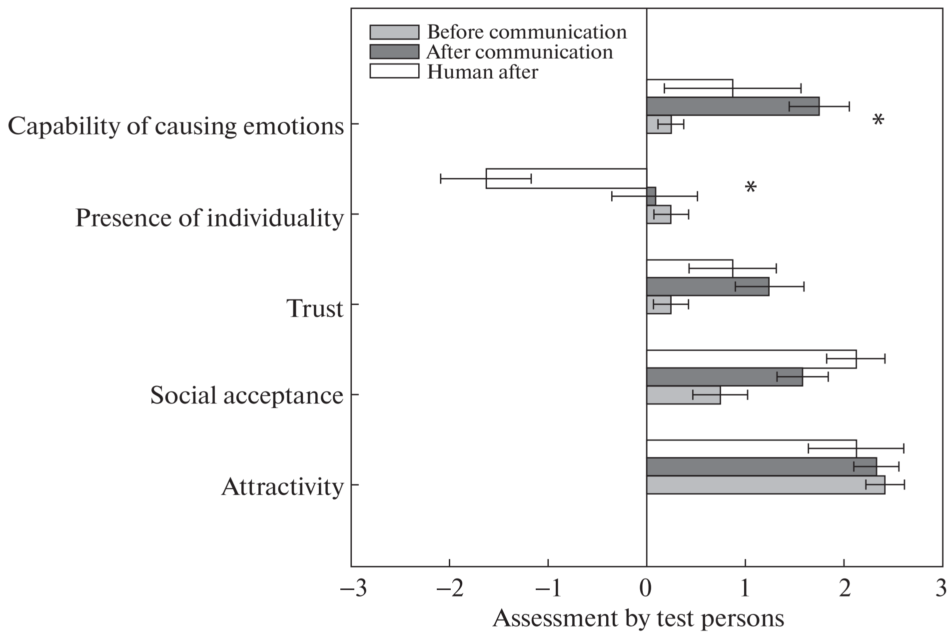

Fig. 3. Comparison of psychological profiles of virtual agent-managed and human-managed registrars before and after the registration procedure.

The maximum accuracy of classification of the user's psychotype (i.e., the Big 5 scores) was achieved for the characteristic "openness to experience" and turned out to be above 90%, and for half of the other characteristics it exceeded 70% (Fig. 4). We expect that connecting eBICA or its neural network equivalent to robot control will allow us to improve the accuracy.

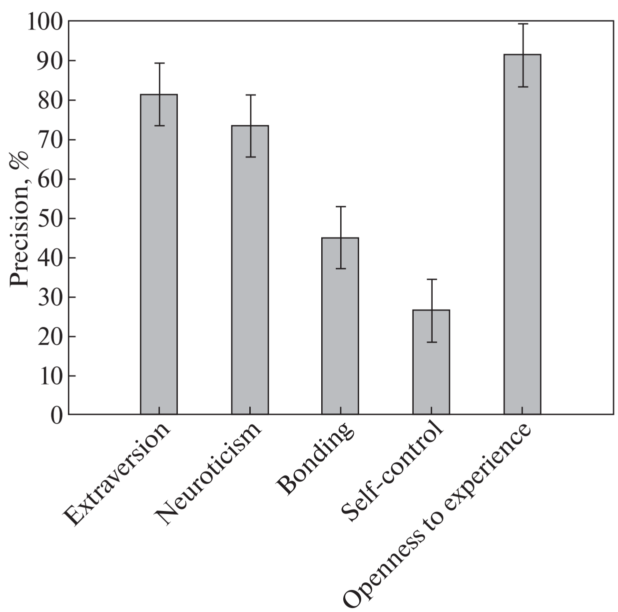

Fig. 4. Results of personality type recognition based on a dialogue with the robot. Whiskers show the standard error estimate.

CONCLUSIONS

In this work, a method of integrating machine learning methods (in particular, LLM) and cognitive modeling (based on eBICA), allowing them to complement each other, is implemented and investigated in experiments with human participants. Encouraging results were obtained in the "Virtual Hotel Receptionist" and "Virtual Psychodiagnostician" paradigms. They indicate that socioemotional conversational agents based on cognitive architectures and machine learning, when implemented correctly, can be socially compatible with humans. In our experiments, the artifact was perceived by participants as having individuality at a level not lower than that of a human. On the other hand, when interacting with an artifact, people revealed their personality characteristics at a level sufficient to reproduce the results of psychological tests. On a number of scales, no significant differences were found between the human and the artifact, which indicates the human-likeness of the artifact. Overall, the results indicate the feasibility of a virtual agent or robot capable of establishing human-level social contact while also assessing the user's psychology.

ACKNOWLEDGMENTS

We would like to thank those who contributed to the work on this project: MEPhI students Alena Anisimova, Aleksei Mikhnev, and Vladimir Tsar'kov.

FUNDING

This work was supported by the Russian Science Foundation grant no. 22-11-00213, https://rscf.ru/en/project/22-11-00213/.

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. M. R. Ali, S. Z. Razavi, R. Langevin, A. Al Mamun, B. Kane, R. Rawassizadeh, L. K. Schubert, and E. Hoque, "A virtual conversational agent for teens with autism spectrum disorder," in *Proc. 20th ACM Int. Conf. on Intelligent Virtual Agents* (Association for Computing Machinery, New York, 2020). https://doi.org/10.1145/3383652.3423900
2. A. S. Anisimova, A. Potemkina, P. Chervakov, V. Komza, D. Maksimov, I. Panin, A. Vaselyuk, and A. V. Samsonovich, "Artificial Psychologist: An intelligent virtual/robotic assistant based on a cognitive modeling framework," Procedia Comput. Sci. **213**, 793–800 (2022). https://doi.org/10.1016/j.procs.2022.11.136
3. P. Blache, M. Abderrahmane, S. Rauzy, and R. Bertrand, "An integrated model for predicting backchannel feedbacks," in *Proc. 20th ACM Int. Conf. on Intelligent Virtual Agents* (Association for Computing Machinery, New York, 2020). https://doi.org/10.1145/3383652.3423948
4. F. D. Díaz et al., "Empathic smart conversational agent for enhanced recovery from abdominal surgery at home," in *HCI International 2022—Late Breaking Posters*, Ed. by C. Stephanidis, M. Antona, S. Ntoa, and G. Salvendy, Communications in Computer and Information Science, Vol. 1654 (Springer, Cham, 2022), pp. 406–413.
5. M. Elgarf, G. Skantze, and Ch. Peters, "Once upon a story: Can a creative storyteller robot stimulate creativity in children?," in *Proc. of the 21th ACM Int. Conf. on Intelligent Virtual Agents* (Association for Computing Machinery, New York, 2021), pp. 60–67. https://doi.org/10.1145/3472306.3478359
6. S. Fitrianie, M. Bruijnes, F. Li, A. Abdulrahman, and W.-P. Brinkman, "The artificial-social-agent questionnaire," in *Proc. 22nd ACM Int. Conf. on Intelligent Virtual Agents, Faro, Portugal, 2022* (Association for Computing Machinery, New York, 2022), pp. 1–8. https://doi.org/10.1145/3514197.3549612
7. A. Hartholt, A. Reilly, E. Fast, and Sh. Mozgai, "Introducing Canvas: Combining nonverbal behavior generation with user-generated content to rapidly create educational videos," in *Proc. 20th ACM Int. Conf. on Intelligent Virtual Agents* (Association for Computing Machinery, New York, 2020). https://doi.org/10.1145/3383652.3423880
8. M. Kosinski, MyPersonality Project (2016). https://web.archive.org/web/20160518020419/http://mypersonality.org/wiki/doku.php?id=download_databases. Cited May 26, 2023.
9. S. C. Marsella and J. Gratch, "EMA: A process model of appraisal dynamics," Cognit. Syst. Res. **10**, 70–90 (2009). https://doi.org/10.1016/j.cogsys.2008.03.005
10. J. A. Martin, "Watson Personality Insights introduction and how to access Watson without SDK," Analytics Vidhya https://medium.com/analytics-vidhya/watson-personality-insights-introduction-and-how-to-access-watson-without-sdk-89eb8992fff2. Cited May 26, 2023.
11. S. M. Mohammad and P. D. Turney, "Crowdsourcing a word—emotion association lexicon," Comput. Intell. **29**, 436–465 (2013). https://doi.org/10.1111/j.1467-8640.2012.00460.x
12. S. M. Mohammad, NRC word-emotion association lexicon (2022). http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. Cited May 26, 2023.
13. A. V. Samsonovich, "Emotional biologically inspired cognitive architecture," Biol. Inspired Cognit. Archit., **6**, 109–125 (2013). https://doi.org/10.1016/j.bica.2013.07.009
14. A. V. Samsonovich, "Socially emotional brain-inspired cognitive architecture framework for artificial intelligence," Cognit. Syst. Res. **60**, 57–76 (2020). https://doi.org/10.1016/j.cogsys.2019.12.002
15. D. V. Tikhomirova, A. A. Chubarov, and A. V. Samsonovich, "Empirical and modeling study of emotional state dynamics in social videogame paradigms," Cognit. Syst. Res. **60**, 44–56 (2020). https://doi.org/10.1016/j.cogsys.2019.12.001
16. Truity. The big five personality test. https://www.truity.com/test/big-five-personality-test. Cited May 26, 2023.
17. A. Kotov, http://www.f2robot.com. Cited May 26, 2023.

A. A. Dolgikh. First year postgraduate student at National Research Nuclear University MEPhI. In 2021, he received a Bachelor's degree in Applied Mathematics and Computer Science, and in 2023, a Master's degree in Software Engineering.

A. V. Samsonovich. Professor in the Department of Cybernetics of the National Research Nuclear University MEPhI, Head of BICA Lab (International Research Laboratory "Cognitive Architectures and Semantic Technologies," Department of Cybernetics, NRNU MEPhI), Scientific Advisor of the Institute of Cyber Intelligence Systems of NRNU MEPhI, Affiliate Faculty in the Department of Bioengineering at George Mason University in Fairfax, Virginia, USA. Holds PhD in Applied Mathematics from University of Arizona (1997).

D. V. Tikhomirova. Senior Lecturer in the Department of Intelligent Cybernetic Systems of the Office of Educational Programs, Junior Researcher of the Institute of Intelligent Cybernetic Systems of National Research Nuclear University MEPhI. Interdisciplinary specialist in information technology, psychology, and management, MBA. Graduated with honors from the Moscow Engineering Physics Institute with a degree in "Applied Mathematics and Computer Science" in 2009 and "Economics and Enterprise Management (Energy)" in 2010. Graduated from the Moscow Institute of Psychoanalysis with a Master's degree in Psychology in 2021.