Albert R. Efimov – Head of Robotics Laboratory, PJSC Sberbank; postgraduate student, Russian Academy of Sciences, makkawity@gmail.com

Efimov, Albert "Technological background of Individuality of Human and its Computer imitation." *Artificial societies.* 14.4 (2019). DOI: 10.18254/S207751800007645-8
https://artsoc.jes.su/s207751800007645-8-1/

TECHNOLOGICAL PREREQUISITES FOR INDISTINGUISHABILITY OF A PERSON AND HIS/HER COMPUTER REPLICA

Introduction

Some people wrongly believe that A. Turing's works that underlie all modern computer science never discussed "physical" robots. This is not so, since Turing did speak about such machines, though making a reservation that this discussion was still premature. In particular, in his 1948 report [8], he suggested that a physical intelligent machine equipped with motors, cameras and loudspeakers, when wandering through the fields of England, would present "the danger to the ordinary citizen would be serious." [8,]. Due to this imperfection of technology in the field of knowledge that we now call robotics, the methodology that he proposed was based on human speech, or rather on text. Other natural human skills were too difficult to implement, while the exchange of cues via written messages was much more accessible for engineering implementation in Turing's time. Nevertheless, since then, the progress of computer technology has taken forms that the founder of artificial intelligence could not have foreseen.

Nowadays, images of people, animals, and fictional creatures on a computer screen or in a movie are nothing new. However, the previous generation of computer graphics technology – from Disney's cartoons to high-end computer graphics of the best-equipped Hollywood studios – only imitated reality, using it as the basis for their designs. Developers of computer graphics have created a new virtual reality that is so similar to the world we observe as the best works of classical painters: it does not merely reflect the world around us but also the artist's individual view of it, and the latter affects the audience's impressions. The new generation of technologies that is emerging right now, before our eyes, is making another approach available. Computer graphics are going to become more than an alternative to the real world, or rather, they will exist as its active extension: a video or photo image of reality will be taken as the basis for new images, video or audio samples. The resulting images may have never existed in reality, yet it will look completely realistic. In this regard, it is perfectly appropriate to recall the ancient Greek story of two painters, Zeuxis and Parrhasius [3].

Zeuxis and Parrhasius were rivals, each of them willing to paint over the wall of a great temple. Quite an audience gathered, and the two rivals came forth, each of them bringing veiled paintings. Zeuxis was the first to pull back his veil, and lo, there was a bunch of grapes, so similar to real ones that birds flocked to peck at them. The audience applauded. "Now you pull the veil away!" Zeuxis said to Parrhasius. "But I can't," Parrhasius replied, "the veil is just what I have painted." And Zeuxis dropped his head. "You win!" he said. "I have deceived the eyes of the birds, but you have deceived the eyes of the painter" [3].

In our opinion, virtual reality, recreated with the maximum likelihood of computer graphics (which is an incredibly intellectual tool of an artist, who is also a specialist in computer graphics), can be correlated with the skill of Zeuxis. In this case the technology of "complementing / augmentation" of reality related to processing images of real objects in artificial neural networks is associated with the skill of Parrhasius. This old tale can also provide the starting point for our further philosophical reasoning.

We do not propose to compare objective reality, computer graphics, virtual reality, and augmented reality. Rather, it is a comparison of the ways in which a machine communicates with a person – how much a person can trust a machine (medium) to deliver a message to another person. Can a machine do this as efficiently as a person? We can extend the well-known aphorism of M. McLuhan that "the medium is the message" [5] and declare that "the machine is the medium." M. McLuhan could not imagine television without humans. But now this is not only conceivable but already real.

E.LENA – digital avatar, television announcer

In early 2019, Sberbank introduced the first digital Russian-language television announcer E.LENA (Electronic ***LENA***), generated as a life-like image of a television announcer. Using artificial neural networking technologies to improve images is not new: every smartphone already has several applications that modify the users' photos satisfying their vanities (removing wrinkles, blurring the background, and correcting colors) in an online mode. We already begin to perceive the surrounding reality as if it had been processed through filters of popular applications, and there already exists even a special designation (hashtag) ***#nofilter*** for the normal, clear view of objective reality. However, the complementing / reconstruction of reality is a fairly new technology: the objects presented to the user in a video sequence seem real documents, as they realistically reflect scenes or objects that are familiar to the user. At the same time, what happens to these objects has never happened in reality and cannot happen. In English, this phenomenon came to be labeled by the catchy word ***deepfake,*** and it is frequently associated with negative phenomena of socio-political life.

It is difficult to say who first proposed the idea of complete digitization of an actor or television presenter. This is going to take special research into history of pop culture and science fiction. However, the idea of a thorough digitization of a professional actor was first explicitly implemented in the little known sci-fi movie *The Congress* (2013).

For the first time, a digital replica of a television announcer was presented by the Chinese company ***Sogou*** that developed a platform solution, commissioned by the Chinese news agency ***Xinhua*** in November 2018. A little later, the Russian Sberbank independently developed and presented a similar technology in the Russian language. The Sberbank digital television announcer can fully automatically read out any text. This enables us to use this solution as a news television announcer on Sberbank corporate television. Currently, dozens of newsreels have been produced using this technology. The audience are hundreds of thousands of employees and customers of Sberbank, who have watched the news delivered by the digital television announcer through various communication channels. Further, we consider this technology in more detail.

"What E.LENA has under her hair?"

The voice of E.LENA is speech synthesized by artificial neural networks of deep learning. In order to create this voice, it was necessary to train neural networks using specially prepared recordings of the original speaker’s voice (she is a professional actress) and to develop software that allows converting arbitrary text into speech.

Designers form E.LENA’s facial expressions using an ensemble of artificial neural networks, pre-trained on specially prepared data: video materials and 3D scans of the prototype actress (currently E.LENA’s voice samples and video images belong to different actresses). As a result of these two-stage transformations that take place without human intervention, we obtain the facial expressions and the speech of a digital television announcer. Then, using automated technological tools and components of computer vision and speech recognition systems, they apply processing, and, as a result, errors are detected and eliminated. Following this, a realistic video is ready for use. The whole complex is a holistic solution based on several independent technologies with AI components.

At the moment, the service for converting text to video operates only as a trial version with the corporate television service of PJSC Sberbank. About 50 different newsreels have been produced using this program. The current implementation of E.LENA has quite a number of drawbacks: poor synchronization of lip movement and spoken text, a limited range of poses, an unnatural voice, etc. However, technology is developing rapidly, and, in the very near future, many companies and research centers will be able to develop high-end products.

The current implementation of E.LENA can deceive many people and make them believe in her reality. Sberbank conducted a survey in its corporate community, which included 1.5 million users of the Odnoklassniki.ru social network, about the “nature of Elena,” presenting two samples to the visitors of this group at the same time: one was the human female announcer, and the other was the digital announcer E.LENA. Even in the imperfect current implementation, over 25% of the 22 thousand survey participants made a mistake (or doubted) about determining the nature (digital or human) of the announcer.

Verbal and non-verbal communication

Human interaction is based on our mutual understanding of the meaning of communication, reflecting not only the intentions in specific communication as they are manifested in our speech and language but also in the context of interaction, which can be geographical, temporal, or semantic. In communication, we also take into account numerous social and cultural characteristics of interlocutors (for example, in the academic environment we tend to use expressions that are different from those suitable while shopping in the market). Generally, in order for a machine (a computer or a robot) to understand a person, it is necessary to ensure understanding of all the three aspects of meaning that we put into speech: language, context, and culture. Therefore, an approach to the study of artificial intelligence that is focused solely on processing of natural languages, seems insufficient to reveal the meaning embedded in communication.

All the three modalities of revealing the meaning (language, culture, and context) are contained not only in the literal meaning of words but in implicit cultural data, which D. Everett calls “dark matter” [7]. The latter does not only consist of semantic combinations of words and phrases but also, for example, of gestures accompanying locutory actions.

In his book, Mladen Dolar gives an example of how K. Stanislavsky instructed his drama students to develop fifty different ways to say the phrase, “Tomorrow night,” implementing various intentions [4]. Interlocutors modify their facial expressions, gestures, and intonation in accordance with the syntactic structure of the sentence and use them as explanations to specify implicit information contained in the speaker's and the listener's cultural or personal backgrounds. D. Everett rightly notes that "language never expresses everything, and culture fills these gaps" [7]. The traditional approach to study of artificial intelligence, based on the text, and in fact on "teletype" messages that came down to us from the era of analog electronics, ignores the hidden "dark matter" of communication, since the interpretation of the message (according to Everett) considers not only verbal reasoning but also the accompanying gestures and facial expressions.

Digital television announcer as a tool for studying person-machine communication

Since the moment A. Turing suggested replacing the question, "Can machines think?" with his *imitation game*, where he proposed to exchange "notes" or teletype messages in communication, AI researchers have, in fact, largely forgotten to pay attention to how messages are transmitted between the judge and the subjects through the so-called "Turing wall," which separates the participants in the imitation game. Hundreds of scientific and popular works on artificial intelligence have completely ignored the issue of "non-verbal" communication with machines.

One of very few exceptions is the 1977 Soviet popular science film *Who Is Behind the Wall?* (1977), where the "Turing Wall" became a video wall. According to the authors of the present article, E.LENA could become a new tool for studying the problems of person-machine interaction and of artificial intelligence by extending A. Turing's method, which later became known as the Turing test. Modern progress in the field of creating "augmented reality," digital avatars, and television announcers such as E.LENA poses another important question: can a machine create the same interpretative basis of speech as a human person, using not only a certain set of words to express thoughts or intentions, but also means of non-verbal communication: facial expressions and gestures? Will a machine possess such an arsenal of tools for a communicative act as a person possesses? Or will we remove the Turing wall only to see microcircuits and batteries of a fraud machine.

The technologies for creating ***augmented digital reality*** give us the opportunity to form a new ***imperfect special Turing test,*** to use the term suggested by A.Yu. Alekseev [1]. An imperfect special Turing test focuses on testing only one component of the original Turing test. In this case, the proposed special Turing test (STT) is aimed at testing the non-verbal communication capabilities of computing sofware. According to A. Alekseev, STT is described by the following components: the subject of testing, an implementation scheme, test questions and answers. In addition, A. Alekseev suggests supplementing the description of the testing itself with discussion (similar to the way A. Turing approached the analysis of objections to his original test) and a description of the sociocultural consequences. When describing the proposed special Turing test "E.LENA," we use the proposed approach.

Subject of testing

In fact, E.LENA is a simulation set of the virtual environment of a television studio. The function of a television studio as a mass medium is to form a specific picture of the world for its audience. The subject of testing proposed in the "E.LENA" STT is the person's ability to perceive information offered by a digital TV presenter, broadcast from a digital television studio. In fact, E.LENA (or rather the software package that creates it) converts text information into an audiovisual format. According to the creators of the system, this should be similar to the format of a television news studio. In the test proposed, we determine how people perceive information delivered by digital television announcers, and whether there is a difference between a person's perception of information when it is transmitted to him by a human announcer and by a digital television announcer.

Like the original Turing test, E.LENA's test leads to a binary result. If the observer's perception of the information is no worse than when viewing the news voiced by a human host, then we consider the test is passed. If the observer comprehends less of the information when the news is voiced by the robot, then the test is not passed.

Further, we will refer to the specific implementation of the E.LENA test as "the experiment."

Implementation environment

Let us clarify the terms used in all design options of the experiment. The object of the study is an ***observer*** in the terminology of A. Turing, or judge (*J*), when we use A.Yu. Alekseev's term. E.LENA is a tool ("medium," in the terminology of M. McLuhan), which is used to study *J's* reactions to information received ("message," in M. McLuhan's terminology).

E.LENA is the digital avatar, or television announcer, i.e., a software product that converts text compiled by the experimenter into a video sequence. The conversion takes place instantly (one second of video is generated in less than one second).

A Judge (*J*) is a person who views each video sample, including E.LENA as a television announcer. The subject of testing (research) is *J's* reaction to E.LENA: an error in the perception of information and rejection (acceptance) of E.LENA as a television announcer (source of information). *J* is the main object of the experiment. The Human (*H*) is a person who acts as a ***television announcer*** in various versions of this experiment. The television announcer (*T*) is the role that *H* or E.LENA can fulfill.

In the course of this specific experiment, the *J* is supposed to watch videos in which people talk about themselves or answer sets of questions. In TV commercials, a *H* or E.LENA (wearing similar clothes and sitting in the same studio) read out their texts, which mostly coincide in theme and style, following which *J* is asked to assess the skills of these announcers. If *J* does not distinguish E.LENA from the other presenters, we consider the ***test passed***.

Testing software

This E.LENA special Turing test exists in two versions. ***Version 1. "Find the robot."*** In this version of the experiment, *J* sequentially watches several video fragments in which *H* or E.LENA can play the role of the announcer. Each video fragment lasts under 25 seconds (in the terminology of news journalism, this fragment is called "a soundbite") (School of Journalism, 2017). In accordance with the rules for formation of newsreels, the number of such sequences should not exceed four. That is why in the proposed test the number of soundbites voiced by various announcers, including E.LENA, is limited to four [2]. The background picture for all video clips remains the same. All *T* texts are on the same topic (culture, sports, weather, etc.). The task of *J* is to determine in which of the fragments E.LENA performed the role of the television speaker. The test is considered passed if the probability of determining E.LENA in all video clips does not exceed 50%.

Version 2. "Happiness comes when people understand you." In this version of the experiment, *J* watches several video clips in which a person or E.LENA can act as presenters. Each clip lasts no more than 25 seconds. A total of four video clips with various presenters are played (among which there are several *human presenters* and E.LENA). The background picture for all the sequences is the same. All texts of television announcers belong to the same topic (sports, culture, weather). The task of *J* is to determine for each of the video clips viewed, the sequence of key facts or opinions that are explicitly expressed. Examination of *J* is carried out right after the viewing. The test is considered passed if *J's* average error when answering questions after watching the video clips with human presenters is the same as when answering questions after watching E.LENA's videos.

Each of the variants is carried out in series, where the roles of *J* are performed by different people. At least 10 episodes should be performed with each *J*.

Discussion and possible objections

The key objection that can be made against the development of the proposed STT is the following: ***this development does not belong in the AI field – it rather represents a new type of computer graphics.*** This objection mixes the technology that we use to obtain a digital avatar (in particular, a television announcer) and its general suitability for TT implementation. Imitation game players could use the image of a digital television announcer in order to more successfully deceive the *J* of a classic TT. On the other hand, following Everett, we admit that non-verbal communication can matter no less than understanding speech proper. And this explains the widespread use of video communication among the younger generation. As the above quotes indicate, A. Turing suggested using teletype only because in his time it was impossible to even imagine something like E.LENA. Now this has become a reality that cannot be ignored in studies of the interaction of persons and machines.

Another objection may be based on the final and successful passage of the E.LENA test. We can say that in this case, the AI conducting a dialogue with *J* "from behind the wall," will be but a reflection, and not so much of the *J*, but of the image that served as the basis for E.LENA. Yet we never claimed this to be a "general" intelligence test. It just "equalizes" the non-verbal capabilities of man and of the machine. While before the arrival of E.LENA and similar systems, machines did not have the possibility of influencing human interpretation channels in full, now such possibilities can be investigated and subsequently implemented.

Conclusion: the socio-cultural implications

Of course, we are still very far from the far-reaching ideas of A. and B. Strugatsky that "any employee with a M.Sc. degree could create models based on his own doubles" [6]. Now the process of creating a full digital replica of a television announcer takes months of work, if we employ a solid multidisciplinary team of video processing engineers and developers. But changes are happening very fast. The experience of operating E.LENA in Sberbank shows that the profession of a television announcer may recede after several years. However, we cannot be quite sure whether a digital avatar will help people to better understand artificial intelligence. We only know for sure that this will help us to feel better about the inevitable future.

REFERENCES

1. Alekseyev A.Yu. *Kompleksnyy test T'yuringa: filosofsko-metodologicheskiye i sotsio-kul'turnyye aspekty* [The comprehensive Turing test: philosophical, methodological and socio-cultural aspects]. Moscow: IInteLL, 2013. (In Russian)

2. Gavrilov K. *Kak delat' syuzhet novostey i stat' mediatvortsom* [How to make a news story and become a media creator]. Saint Petersburg: Amfora, 2007. (In Russian)

3. Davydova L.I., Kon'kova G.I., & Chubova A.P. *Antichnyye mastera. Skul'ptory i zhivopistsy* [Antique masters. Sculptors and painters]. Leningrad: Iskusstvo, 1986. (In Russian)

4. Dolar M. *A voice and nothing more.* Cambridge, MA: MIT Press, 2006.

5. McLuhan M. *Understanding Media: The Extensions of Man.* New York: McGraw-Hill, 1964.

6. Strugatsky A.N. & Strugatsky B.N. *Ponedel'nik nachinayetsya v subbotu: Skazka dlya nauchnykh rabotnikov mladshego vozrasta* [Monday starts on Saturday: A tale for young scientists]. Moscow: Detskaya literatura, 1965. (In Russian)

7. Everett D.L. *How Language Began: The Story of Humanity's Greatest Invention.* New York: W. W. Norton, 2017.

8. Turing A. Intelligent Machinery (1948). In: B.J. Copeland (Ed.) *The Essential Turing* (pp. 395–432). Oxford: Oxford University Press, 2004.

9. Shkola zhurnalistiki [School of Journalism] (In Russian)