Processing Like People, Understanding People, Helping People—Toward a Future Where Humans and AI Will Coexist and Co-create

Takeshi Yamada

Abstract

Artificial intelligence (AI) has been making remarkable progress in recent years and has even been approaching the level of human performance for certain functions, but it still has its limitations. In contrast, human beings are highly advanced and complex, which is why they are also imperfect and prone to mistakes as reflected by their vulnerability to bias and illusions. This article introduces NTT initiatives in communication science to bring AI technology closer to a human level and to develop an even deeper understanding of human beings with the aim of closing the gap between AI and humans and achieving AI that can help people.

Keywords: artificial intelligence, communication science, brain science

1. Introduction

Recent developments in artificial intelligence (AI) have been truly remarkable. In the beginning, computers were especially good at performing batch processing of large amounts of data that humans could not process and at performing high-speed processing on behalf of humans for tasks that humans were weak at. However, thanks to recent advances in deep learning, computers are approaching—and surpassing in some cases—human abilities in areas where they have long been behind, for example, speech and image recognition and natural language processing that humans are inherently good at. In the future, we can expect progress in AI to accelerate in this area of media processing.

Nevertheless, neural processes are complex, with many of them still unexplained. It is said that a level of AI performance exceeding the abilities of the complex human brain still lies somewhere in the future. In contrast, humans, though being very advanced and complex creatures, appear at first glance to be imperfect since they can make mistakes under the influence of cognitive bias and be tricked by illusions into thinking that something that does not exist is real.

With the above in mind, the mission of NTT Communication Science Laboratories, which incorporates the words *communication science* in its name, is to connect and close the gap between computers (AI), which will continue to develop rapidly within a limited range, and humans, whose complexity also makes them imperfect (**Fig. 1**). Specifically, we look to build a theoretical foundation and develop innovative technologies toward person-to-person and person-to-computer *heart-touching communication* [1].

As one straightforward example of building a theoretical foundation, we have proposed a highly efficient coding method for sending and receiving messages up to the limit of coding efficiency (Shannon limit). This is described in “Transmission of Messages

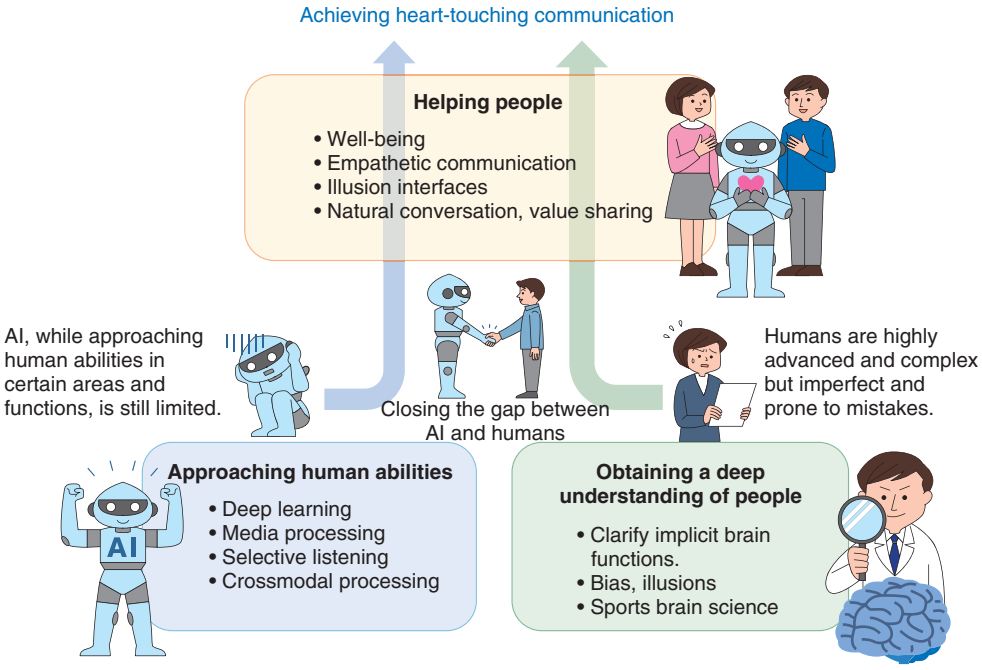

Fig. 1. Mission of communication science.

to the Efficiency Limit—Implementation of Tractable Channel Code Achieving the Shannon Limit" in the Feature Articles in this issue [2].

Furthermore, to truly achieve heart-touching communication, we must, of course, study technology that can approach human abilities with a focus on media processing. It is also important, however, that we explain human functions and characteristics and obtain a deeper understanding of people overall with the aim of developing technology that can truly help people.

2. Technology approaching human abilities

There are still many processes today that are difficult for computers to accomplish but that humans do exceptionally well. Of course, the accuracy of machine translation has been improving by leaps and bounds, and it has even become possible for an AI system to correctly answer to some extent fill-in-the-blank questions in the English portion of a Japanese university entrance exam [3]. Nevertheless, computers have yet to reach the level at which they can deeply understand the meaning of a sentence or exhibit commonsense.

It is also true, however, that computers have approached the level of human abilities in specific areas such as image recognition and speech recognition through the use of deep learning technology. Take, for example, a meeting or party where it is common for more than one person to be talking at the same time or for music to be playing in the background. Despite such noisy conditions, a human is able to zero in on the voice characteristics of the person he or she wants to listen to and to understand what that person is talking about. This is a distinctive feature of human hearing known as selective listening, which is a typical example of the broader concept of selective attention.

Computers have traditionally been weak at selective listening, but at NTT Communication Science Laboratories, we have applied proprietary deep learning techniques to develop technology that enables computers to catch only the words of the target speaker based on the voice characteristics of that person—much like humans do—and have begun to roll out this technology [4].

The key to enabling such media processing technologies to progress even further toward human abilities is crossmodal processing, which refers to processing that can cross the boundary of a single modality such as speech, video, or text. In the past, the conventional approach was to research media such as speech, video, and text separately using

different analysis techniques. Today, however, thanks to the advent of deep learning that has taken up the role of a common language, recognition, generation, and conversion across multiple modalities are becoming possible.

Humans, on the other hand, have always made use of multiple senses (the five senses) in perceiving the outside world. For example, just by hearing a sound, humans are capable of imagining to a certain extent the situation associated with that sound at that place. For people, this type of crossmodal processing is commonplace in everyday life. In addition, the phenomenon of sensory substitution is well known as a means of replacing a sensory function that has been lost due to an injury or other reason with another functioning sense, as is done by visually impaired people who use their fingertips to read Braille printing.

For humans, it is natural on seeing a photo of a person's face to imagine to some extent the voice associated with that face. Could computers be made to do the same? At NTT Communication Science Laboratories, we have taken up the challenge of enabling computers to actually perform such crossmodal processing. For example, the aim of cross-media scene analysis technology is to use sound to recognize all active events in a scene even at a location situated in a camera's blind spot. The latest crossmodal processing technologies now being pursued at NTT Communication Science Laboratories are described in "See, Hear, and Learn to Describe—Crossmodal Information Processing Opens the Way to Smarter AI" [5].

3. Technology for obtaining a deep understanding of people

In the above way, computers are approaching human abilities in specific areas if not surpassing them, but it appears that more progress is needed if AI performance is to exceed the complexity of the human brain. Humans, on the other hand, are sometimes swayed by cognitive bias or fooled by illusions that lead to completely unexpected mistakes, as reflected by the ease at which some people are taken in by bank transfer scams. The Illusion Forum website managed by NTT Communication Science Laboratories provides information on a variety of illusions that can make a person doubt one's own eyes or ears [6].

In a famous experiment conducted by Christopher Chabris and Daniel Simons [7], subjects are shown a video of six players in white and black shirts passing around basketballs to each other and instructed to count the number of times that the players in white pass one of the balls to each other. Here, at nine seconds into the video, a gorilla walks into the scene, faces the camera, pounds his chest majestically, and finally exits. Nevertheless, about half of the subjects are so engrossed in counting that they never notice the gorilla. In this way, a human turning his/her attention to something fails to notice other things that are happening in the same scene. In other words, the flip side of the remarkable human characteristic of selective attention is selective inattention. In addition, a person focusing on something does not even notice that this is happening. Thus, it is not only the elderly who get taken in by bank transfer scams.

In this way, the complex nature of humans also makes them imperfect as reflected by their tendency to be fooled by bias or illusions. In contrast, AI, while limited at present, is steadily advancing. To therefore close the gap between humans and AI and achieve coexistence and co-creation between them, it is essential that we obtain a deeper understanding of human beings in all their complexity before believing—without careful consideration—the idea that AI will one day surpass the human brain.

To this end, NTT Communication Science Laboratories is expending effort to clarify and understand implicit brain functions related to the basic human senses of seeing, hearing, and sense of movement. Here as well, illusions can provide important clues to understanding such implicit brain functions.

Understanding implicit brain functions is a challenging task. The brain activity patterns, for example, may vary greatly across individuals. Focusing on top-ranking athletes as subjects, we are working to explain the outstanding abilities of these individuals from the viewpoint of brain science and to find out how mind, technique, and body in humans are interrelated as part of our efforts in sports brain science.

For example, we have taken up the challenge of clarifying the mechanism of how a top hitter in baseball can judge whether the incoming ball is slow or fast and adjust the timing of his swing accordingly all within a very short period of time of about 0.1 second [8]. Sports brain science can be regarded as a new technology and an ambitious undertaking that departs from conventional sports science and sports analysis techniques that are mainly concerned with body training.

Incidentally, crossmodal processing as mentioned above takes place on a variety of levels within the

brain. For example, when people view an ordinary video, the brain initially processes the information on color, form, and movement separately and integrates that information later. As a result, any inconsistencies that might exist among those different modalities of information will be corrected in the integration process.

This brain processing mechanism was used to devise Hengento at NTT Communication Science Laboratories [9, 10]. When experiencing Hengento, the user obtains color and form from a static object while obtaining movement from monochrome video projected on that object. Since color and form are static here, a spatial inconsistency occurs with movement. However, the brain, which attempts to see an object in a consistent manner, will correct for this inconsistency when integrating movement, color, and form. Consequently, in the Hengento experience, the user notices no inconsistencies among movement, form, and color and falls under the illusion that the color and form of the static object are actually moving. The name Hengento is derived from a Japanese word meaning illusory transforming lamps.

4. Technology for helping people

The results obtained in sports brain science research are not limited to sports. They can also be used to bring implicit mental and physical abilities into full play in the everyday life of human beings. That is to say, they can be used as knowledge for improving well-being in people. Improving human well-being is a qualitatively elusive problem, so we are tackling it in a quantitative manner from the viewpoint of human science and establishing design guidelines to enhance the sense of well-being.

One example of this approach is our work in measuring the effects of empathetic communication that occurs when a number of people come to share the same space [11]. Additionally, given the eye-straining effects of display devices such as televisions and smartphones that surround us in our modern society, we are proposing a method for self-checking the state of one's eyes on a routine basis using a general-purpose tablet device in a game format. This method is described in "Measuring Visual Abilities in an Engaging Manner" [12] in this issue.

At the same time, while illusions play an important role as clues to explaining implicit brain functions, they also hold the key to filling in the gap between humans and AI and to facilitating interfaces and feedback designed to help people in their daily lives. At NTT Communication Science Laboratories, we have developed a device called Buru-Navi that generates the illusion of being pulled by some force as an interface that exploits human illusions. We are also working on a means of making a sitting person feel as if he/she is actually walking. This development is described in "Creating a Walking Sensation for the Seated—A Sensation of Pseudo-walking Expands Peripersonal Space" [13].

In fact, we have announced a series of interesting interfaces in this area, including the aforementioned Hengento that makes a printed picture or photograph appear to move simply by projecting light on it, Hidden Stereo that enables a viewer to enjoy three-dimensional (3D) video while wearing 3D glasses and vivid 2D video when not wearing them, Ukuzo, an optical projection technique that gives 2D objects such as those on printed matter a 3D floating effect by projecting shadow-like patterns onto them [14], and Danswing papers, which gives an impression of motion to static paper objects and was selected as a top 10 finalist for the 2018 Best Illusion of the Year contest [15].

In future research, we plan to work on new types of interfaces that use illusions while simultaneously investigating the possibility of novel forms of perceptual expression that exploit illusions to create experiences that cannot be achieved by physical means.

To achieve natural dialog between humans and robots or AI, dialog-processing technologies such as speech recognition and natural language processing will be important, and it would seem at first that human biases and illusions are unrelated. However, AI has yet to reach the point of being able to understand the full meaning of a sentence or exhibit commonsense the way humans do, so dialog between humans and AI is mostly limited to a one-question/one-answer format at present. As a result, if an inconsistency arises in what is being said with what was said shortly before, a glitch in the process will quickly be exposed, and the dialog will be short-lived. It is therefore necessary to learn how to make effective use of such limited abilities and to exploit human biases and illusions so that AI appears *smart* to humans.

At NTT Communication Science Laboratories, we have achieved dialog processing that enables natural ongoing conversation even in a one-question/one-answer format by skillfully dividing up tasks between two robots. Furthermore, with the aim of breaking away from the one-question/one-answer format, we focused on the fact that much of what users talk about

concerns event-related information and therefore proposed a technique for structuring and understanding user utterances in units of events. This technique is described in "Chat Dialogue System with Context Understanding" [16]. In this way, context understanding can be improved, and simulated experiences of systems that match events can be shared, which should result in dialog that can truly help people such as by inducing empathy between people and robots.

5. Conclusion

As described above, human beings are advanced and complex creatures, while AI, though approaching the level of human performance in certain areas and functions, is still limited. Achieving intelligence that surpasses that of humans is not that simple. Humans, on the other hand, are complex and imperfect; they can be taken in by bank transfer scams, be mistaken about cause-and-effect relationships, be vulnerable to bias, and be prone to mistakes. It is also known from optical illusions that humans may not always be observing physical quantities for what they really are. It is therefore important to close the gap between humans and AI and achieve AI that can help people by refining AI technology to approach the level of human abilities while simultaneously deepening our knowledge of human characteristics. This is the mission of NTT Communication Science Laboratories as we work toward achieving heart-touching communication.

References

[1] T. Yamada, "Shift to New Dimensions—Further Initiatives to Deepen Communication Science," NTT Technical Review, Vol. 16, No. 11, pp. 14–18, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa1.html

[2] J. Muramatsu, "Transmission of Messages to the Efficiency Limit—Implementation of Tractable Channel Code Achieving the Shannon Limit," NTT Technical Review, Vol. 17, No. 11, pp. 34–39, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa6.html

[3] R. Higashinaka, H. Sugiyama, H. Isozaki, G. Kikui, K. Dohsaka, H. Taira, and Y. Minami, "Taking the English Exam for the 'Can a Robot Get into the University of Tokyo?' Project," NTT Technical Review, Vol. 13, No. 7, 2015.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201507ra2.html

[4] M. Delcroix, K. Zmolikova, K. Kinoshita, S. Araki, A. Ogawa, and T. Nakatani, "SpeakerBeam: A New Deep Learning Technology for Extracting Speech of a Target Speaker Based on the Speaker's Voice Characteristics," NTT Technical Review, Vol. 16, No. 11, pp. 19–24, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html

[5] K. Kashino, "See, Hear, and Learn to Describe—Crossmodal Information Processing Opens the Way to Smarter AI," NTT Technical Review, Vol. 17, No. 11, pp. 12–16, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa2.html

[6] Illusion Forum (in Japanese), http://www.kecl.ntt.co.jp/IllusionForum/

[7] C. Chabris and D. Simons, "The Invisible Gorilla," http://www.theinvisiblegorilla.com/videos.html

[8] D. Nasu, "Timing Adjustment of Baseball Batters Determined from Motion Analysis of Batting," NTT Technical Review, Vol. 16, No. 3, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201803fa3.html

[9] T. Kawabe, T. Fukiage, M. Sawayama, and S. Nishida, "Deformation Lamps: A Projection Technique to Make Static Objects Perceptually Dynamic," ACM Transactions on Applied Perception, Vol. 13, No. 2, Article 10, 2016.
https://dl.acm.org/citation.cfm?id=2874358&dl=ACM&coll=DL

[10] Press Release issued by NTT on February 17, 2015, https://www.ntt.co.jp/news2015/1502e/150217a.html

[11] J. Watanabe, Y. Ooishi, S. Kumano, M. Perusquía-Hernández, T. G. Sato, A. Murata, and R. Mugitani, "Measuring, Understanding, and Cultivating Wellbeing in the Age of Technology," NTT Technical Review, Vol. 16, No. 11, pp. 41–44, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa6.html

[12] K. Maruya, K. Hosokawa, and S. Nishida, "Measuring Visual Abilities in an Engaging Manner," NTT Technical Review, Vol. 17, No. 11, pp. 17–22, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa3.html

[13] T. Amemiya, "Creating a Walking Sensation for the Seated—A Sensation of Pseudo-walking Expands Peripersonal Space," NTT Technical Review, Vol. 17, No. 11, pp. 23–27, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa4.html

[14] T. Kawabe, "Ukuzo—A Projection Mapping Technique to Give Illusory Depth Impressions to Two-dimensional Real Objects," NTT Technical Review, Vol. 16, No. 11, pp. 30–34, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa4.html

[15] T. Kawabe, "Danswing Papers," http://illusionoftheyear.com/2018/10/danswing-papers/

[16] H. Narimatsu, H. Sugiyama, M. Mizukami, T. Arimoto, and N. Miyazaki, "Chat Dialogue System with Context Understanding," NTT Technical Review, Vol. 17, No. 11, pp. 28–33, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa5.html

Takeshi Yamada
Vice President and Head of NTT Communication Science Laboratories.

He received a B.S. in mathematics from the University of Tokyo in 1988 and a Ph.D. in informatics from Kyoto University in 2003. He joined NTT Electrical Communication Laboratories in 1988. He was a visiting researcher at the School of Mathematical and Information Sciences, Coventry University, UK, from 1996 to 1997. He was a group leader of the Emergent Learning and Systems Research Group from 2006 to 2009 and an executive manager of the Innovative Communication Laboratory from 2012 to 2013 at NTT Communication Science Laboratories. His research interests include data mining, statistical machine learning, graph visualization, metaheuristics, and combinatorial optimization. He is a Fellow of the Institute of Electronics, Information and Communication Engineers and a senior member of the Institute of Electrical and Electronics Engineers, and a member of the Association for Computing Machinery and the Information Processing Society of Japan.