THE MULTIMEDIA POSSIBILITIES OF ARTIFICIAL INTELLIGENCE AND MEDIA CONVERSION

László T. Nagy, Ph.D; t.nagy.laszlo@drhe.hu; (Debrecen Reformed Theological University)
István Károly Boda, Ph.D; boda.istvan@drhe.hu; (Debrecen Reformed Theological University)
Daniel Benediktsson; danielb@unideb.hu; (University of Debrecen)
Erzsébet Tóth, Ph.D; toth.erzsebet@inf.unideb.hu; (University of Debrecen)

Purpose – The aim of the study is to summarise the most common AI-based media generation tools, mostly available for free, and to examine their inputs and outputs from a media and data conversion perspective. The research will also investigate how these AI-based tools offer new options to address different media conversion requirements, in particular the specificities of the multimedia content they generate.

Methods – The study used an empirical approach, analysing media generation processes using specific AI tools. The content produced by the tools was evaluated on the basis of heuristic criteria, focusing on the types of data conversion and the structural and linguistic characteristics of the content.

Results – AI-generated verbal and multimedia content, although based on different editorial principles and inputs, is increasingly indistinguishable from human-generated content. The tools investigated can effectively convert data in different formats, opening up new avenues for the development, creation and adaptation of digital learning materials.

Value – The research highlights the potential of AI-generated content and the underlying technologies (media conversions) to support learning and teaching processes. However, their application requires further critical examination, particularly with regard to the authenticity, quality and pedagogical value of the content.

http://doi.org/10.52036/1335793X.2025.1.33-39

INTRODUCTION

The applicability of multimedia tools as possibilities in support of learning and teaching has often been investigated. We have also tackled this issue. Later on, with the appearance and spreading of social networks, new tools have entered the scene whose freshness and main strength is the knowledge potential hidden in the community — and which is based—organically—on the tools of multimedia and the internet network (T. Nagy, Boda 2014). After the appearance of social media and artificial intelligence the focus today is on synthesized content. One of the best-known applications is the free version of the ChatGTP, a linguistic model that concentrates on textual communication. Other AI-based solutions are capable — beyond multimedia and character-based communication — of creating and converting voice/sound, pictorial, and video material or even changing their content.

Results of AI research have appeared almost surreptitiously in our everyday lives in recent decades or more closely in the past 5–10 years. Most people have started to use AI-based products or services without even suspecting that the results are produced fully or partially by artificial intelligence in the background.

The appearance of ChatGPT in 2022 has changed everything. The free accessibility of the linguistic model followed by the strong media response has helped access AI and its results even for those who had little or no knowledge about it or were not at all interested.

THE RESEARCH

The use of AI for a specific purpose can only be conceived when we cannot or will not be prepared for all the conceivable relevant in or output (due to their expected

quantity) or an algorithm is expected to perform where the traditional programming solution would require too much human/machine etc. efforts. The analysis of natural languages (e.g. shaping texts) or pictorial analysis are such tasks. The output generated by AI will substitute social knowledge appearing in social networks. The word output is used intentionally instead of the word knowledge because AI — at least on its current level — does not understand the question and neither does the reply, at least not in the sense we humans do.

In our research, the goal has been set to investigate with empirical method, i.e. with the heuristic evaluation of the responses that the tools and solutions which use AI available and mostly free of charge on the internet what multimedia or media conversion capabilities apply. Beyond that, the practical applications of the capabilities mentioned will also be reviewed, with the expectation that they could support certain educational and learning processes. The method of the investigation is justified because the currently available AI applications from the point of view of research theory can only be considered as black boxes where functioning due to limited information cannot be properly known and only the methodology of the Turing test can be applied (Turing 1950).

MEDIA CONVERSIONS

Looking at AI as an entity it is functionally a neural network applying deep learning that, based on the input task synthesizes the expected (i.e. textual) reply in the output. It is important to note that in the case of current solutions what happens is not the interpretation of the task (input) but that the system based on the structure of the problem tries to predict the output reply (prediction), again based on the connected relevant information. Therefore, due to the character of the functioning, the AI (e.g. ChatGPT) can provide "with certainty", a reply to the question that is not correct. In such a case the incorrect result is not intentional but using a special terminology, the AI "hallucinates". As it cannot interpret the question — meaning that it does not understand what is asked about — cannot pass a judgment on the correctness of the reply either.

The utilization of the possible data sources and the selection or composition of the replies (outputs) take place based on some kind of ranking order, e.g. frequency of citations, popularity, and evaluation of the scores or other relevancy factors, etc. In the case of a linguistic model similar to ChatGPT, the output based on the information found in the database is only the composed (possible) word order or set of sentences. Considering all this (or rather despite this) the outputs of the AI mostly provide replies that make us believe the task has been solved by an interpreting (i.e. human) mind (Laky 2023).

It is an important capability of deep-learning neural networks with effective pattern recognition in certain areas (e.g. picture and speech/text analysis) produce better results than earlier machine-learning solutions (Tóth 2020). The pattern recognition capability makes possible data transformation and directed content generation (data synthetisation).

The two concepts cannot be sharply separated. It can be said that one of the two is more emphatic. If e.g. we give a picture of a dog as an AI input, then a conditioned deep neural network can and does provide a textual output like "dog". In this case, the data set composed of picture points is interpreted as pictorial information and transformed into the word "dog" in the language which is then presented as a character sequence and will be an output result. In this particular case, it is data transformation that takes place. (the picture depicting the dog (a pixel set) is transformed into the letter set of the dog.

Figure 1 Transformation data conversion

If it is reversed and it is requested from the AI in a textual form: "Draw a dog." then the conversion act is considered as content generation.

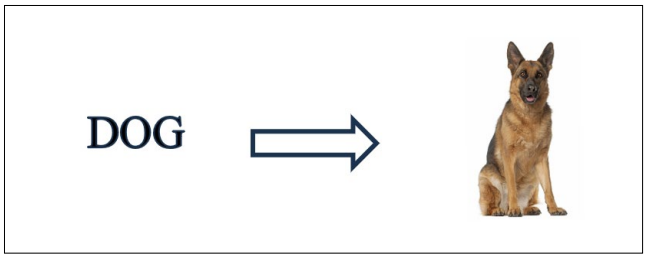

Figure 2 Data conversion with content generation

When the dimensional characteristics are reduced it is transformation, when increased the output is interpre-

ted as a generation. It must also be noted that while in the case of conversions media change (picture → text: jpg → txt) takes place it does not happen in every case. In the following chapters, only those conversion directions that use AI and are available on the internet will be discussed.

The possible application areas of AI can be divided, beyond data analysis and interpretation, in the output side two conversion and functional forms, i.e. transformation and content generation.

TRANSFORMATIONS

input		output
character	→	character (translation)
picture	→	character (character recognition)
picture	→	text (object recognition)
speech	→	written text (speech recognition)
sound	→	text (sound recognition)
music	→	text (music recognition)
text	→	speech (machine reading)
character	→	character (translation)

One of the best-known and most used forms of AI application is translation between natural languages. What happens here is that there is a source language text consisting of characters that need to be transformed into a text that content wise corresponds to the target language characters. This is data conversion but there is no media conversion. Though there have been translation programs for a long time quality translation between languages is still a desideratum. Despite the rapid development in recent years the mechanization of this process still faces challenges depending on content, language, topic, etc.

The best-known application is Google Translator, which is available online in more than 120 languages. The Deepl knows 33 languages and is considered quite strong in the exactness of the translation, in the interpretation of uncertain meanings and it understands Hungarian too (Deepl 2024).

picture → character (character recognition)
The function of character recognition can be found in several document-producing applications as well as in the immediate translation option of Google Lens. Let us think about optical character recognition (OCR) where the relevant (textual) information appears in varied graphic (pictorial) forms in the output, which means e.g. a part of a word or text. In contrast, the output presents the character necessary to recognize the letters (Google Lens 2024). In the case of media conversion, the pictures with textual content can be edited, copied, or sent to the input of other textual applications after character transformation.

This media conversion can in several ways support learning and has become a critical area of AI application. The recognition of the text depends on several factors: e.g. contrast, lightning, looking angle, size, character set, etc. The pattern recognition capability of AI deep learning increases the recognition rate even in handwriting. It is neither necessary nor possible to be prepared for the appearance of any text. The AI, after conditioning, can guess, i.e. interpret the text (the characters).

Besides, after applying OCR (which is advantageous in itself) we get editable text instead of pictorial text, and further data conversion provides us with other functions. The application of Google Lens calls on virtual reality to help immediate translation.

With Google Lens, the phone camera can "see" the foreign language text as in a live video. The OCR supported by AI converts the text into a character sequence (picture → to text) and then the Google translating program translates to the requested language. This data conversion does not result in media conversion (character → character). Then the text in the requested language appears in pictorial form (live video) in place of the original text. In this case: (original) picture → (original language) text in characters → target language text → (virtual) picture and this picture shows the translated text provided we consider the direction of the data flow and their conversions (Google Translate 2024), (Google Lens 2024).

picture → text (object recognition)
Recognition of visual information containing no text (e.g. humans, traffic signs, etc.) is a basic feature of self-driving vehicles. The direction of media conversion is as follows: from a pictorial pixel setting becomes a text (character set) corresponding to the pictorial content in the output. Effective and exact object recognition represents a mainline in AI research. Best known are military, navigational traffic, and security applications.

Beyond this, it can help identify names of equipment, and objects with foreign language replies. It is a means that supports learning both in the case of beginners and small children as well (e.g. connecting object — word form — meaning or learning words etc.). In Google

Lens, object recognition is also an integrated solution (Google Lens 2024).

speech → written text (speech recognition)
In this case, the production of the character (textual) version of the voice speech is expected from AI. This is a technically difficult conversion and one that cannot be easily standardized because the speakers of a language of voice speech speak with different accents, pronunciations, tempo, voice power, speech faults, and background noises. Similar to character and picture recognition, getting prepared for all programmable possibilities is impossible. What can be done is to call on the pattern recognition capability of AI. This function was used at the beginning in industry for voice control.

Today even in the case of complete texts dictating functions can be applied to (e.g. notes, sketches, and similar documents supporting learning) or textual form of an existing voice file or even the subtitling of a video data stream. Recently there have been language-learning applications that use the speech recognition capability of AI to practice pronunciation (e.g. Duolingo, Xeropan) i.e. the text pronounced by the student is compared (pattern fitted) to a standard pronunciation. When the text spoken by the student is similar to the reference (within certain limit values) the system accepts it as correct (Duolingo 2024), (Xeropan 2024).

sound → text (sound recognition)
This is a media conversion when the sound recognition capability of AI can help. The textual description of the non-speech voices, sounds, and noises can be expected in the output e.g. applause, engine noise, thunderbolt, car horn, etc. Concerning hearing-impaired persons, it is possible to prepare a textual description (subtitles) to a video file or other textual output usable for general commands as well.

music → text (music recognition)
In the music recognition function in the case of a played music piece, the converted output can be metadata, e.g. composer/performer, title: Queen – Bohemian Rhapsody or in another case, the result is the textual description of a sung voice which is similar to what is mentioned at the speech → text. In this media conversion case, the description of songs can be expected which can then be used to support text interpretation or even to learn a song text.

text → speech (machine reading)
The AI reads the written text by imitating human speech. The pronunciation of letters, numbers, or words can be implemented by simple, analog attachment (e.g. certain voices or numbers get prefixed, and later will be connected according to intention). Later there appeared the "reading" of the text with the aid of a speech synthesizer where the machine produces the pronunciation of every letter of word tag with the help of a (digital) voice synthesizer. Today, however, these reading forms look quite out of date.

A more developed solution is the application of the deep learning capabilities of AI. Technically it is realized in the following way: using the voice file of the reading by living persons (e.g. 10–100 pages like a teaching database), the AI can "read" any text in the natural voice of the person in the database, material that it has never read in reality. In this case, too, the advanced pattern recognition and predictive capability of the deep learning network are used to achieve the desired synthesized output.

This kind of data conversion is useful because even while traveling is possible to learn or repeat e.g. written textual contents whereas machine reading can be a very important possibility for visually impaired people as well. E.g. Natural Readers, Voicemaker (Natural Readers 2024), (Voicemaker 2024).

CONTENT GENERATION

input		output
text	→	picture (picture generation)
picture	→	text (picture description)
text	→	video (video generation)
picture	→	picture, and picture → video (picture and video transformation)
text	→	sound (music composition)
question	→	reply (chat bot) (text → text)

text → picture (picture generation)
In the picture generation function, the AI produces pictures, and figures from the text based on the instructions of the textual input, e.g. "Draw a running horse on the river bank in sunset, in romantic style, photorealistically." The virtual artist, i.e. the AI produces one or more pictures from the database, according to the instructions. The first trials can then be further refined, according to our needs.

Recently there have appeared several publicly available applications that are capable of producing a close-to-reality pictorial world after description even in photorealistic mode (e.g. Mid-Journey, DALL-E, Stable Diffusion,

Nightcaffe, Replicate, etc.)(MidJourney 2024), (DALL-E 2024), (Stable Diffusion 2024), (Nightcaffe 2024), (Replicate 2024). The applications are appropriate for the preparation of figures or drawings in varied forms or as well as photo like pictures. They can be used as illustrations, examples, or word forms. It is also quite amusing to draw with (mostly English language) instructions, as it is the case when learning words, where not only the picture of the words but e.g. the relation and position of the picture parts are to be defined.

picture → text (picture description)
Compared with the former this is a data conversion of a different course. This happens when the AI prepares a textual description from the input picture, which is a picture description. This possibility is similar to object recognition. At the same time, it is more than that because the relation, position, ratio, meaning, etc. can be symbolically interpreted. Both conversions elicit a change of media (text → picture, picture → text).

text → video (video generation)
Based on the instructions of the textual input AI, generates video media. This media conversion form — because of its complexity — is today not yet widespread, but it has several applications e.g. replicate.com (replicate 2024).

picture → picture, and picture → video (picture and video transformation)
AI produces (or transforms) after (e.g. textual) instructions from a picture another picture or a video. E.g. adds or subtracts or modifies a picture or pictorial objects. It can colour black and white pictures, ameliorate resolutions, correct damaged or missing parts, can make static pictures mobile, etc. As mentioned in the paragraph on from text to picture of the currently available solutions (see above) more is capable of the use or transformation of input pictures.

text → sound (music composition)
Based on the instructions of the textual input, AI generates sounds or music as well. "Musical composition" from text as media conversion is an interesting experimental area within AI research. There are manifold experiments in progress: AI composes music according to instructions — e.g. in a given style — or creates by imitating a given performer.

question → reply (chatbot) (text → text)
In the case of textual communication in chat applications, there is no media shift which means that a textual question elicits a textual reply (e.g. ChatGPT 3.5 language model). Here the expected output is to receive a sensible reply in a human language. As mentioned earlier AI generates the expected textual reply based on the question. But the fact is that the AI available today does not understand the question, just try to predict the possible output reply based on the question itself (Laky 2023). If we take a language model similar to ChatGPT then the reply is no more than a possible word order/sentence set.

"In spite" of all this with the appearance of ChatGPT a complex predictive language model is capable of answering general, complex questions and it is important to note that the model is being developed, its database is being increased, and due to user interactions it is constantly learning and developing. The current version (e.g. 3.5) can now correctly answer questions to which the earlier solutions (or results) before its appearance were not relevant. The latest versions of ChatGPT provide a means that cannot be disregarded in the areas of education or learning support in general.

A detailed analysis of the language model is far beyond the scope of the current study. That, however, can be declared that the ChatGPT can be used as a means that supports meaning and is capable of answering concrete questions in most areas of science and can also prepare analyses, and studies as well as to support language learning.

From version 4 the model "sees", i.e. is capable of interpreting pictures added to the input which makes sure the possibility of complex answers. From September 2023 on the ChatGPT Plus can "speak" and can interpret speech which means it "hears". This can mean that the problem to be solved is in a picture and the solution will be in a written or spoken answer. When the problem is approached from the angle of learning support or teaching the AI can be considered metaphorically as a tutor or consultant with whom everything can be discussed, even in a foreign language (ChatGPT 2024), (ChatGPT 4 2024).

CONCLUSION

Above the currently available AI applications and their media conversion capabilities have been considered. The conversions can be divided into two parts. According to the first route compared with the input data the output can be taken as a transformation. In the second case, it is content generation (in other words synthetisation). When the so-called dimensional characteristics decrease or do not change significantly,

then data transformation (e.g. picture → text), when increases, i.e. abstraction decreases then content generation (text → picture) is what happens. It has also been found that not every conversion entails media changes (see e.g. in the input and output cases of translation programs or textual chat applications often there is no media change, i.e. text → text). In the majority of cases, however, with data conversion media change does happen (e.g. text → picture, picture → video, speech → text, etc.)

Based on the investigation it can also be stated that many applications have multimedia capabilities, which means that they can handle and even transform several media (text, picture, sound, video). It is also a tendency for more and more applications to become multimedia-friendly. Let us have the ChatGPT evolution in mind where compared with the input and output of the first versions (textual question → textual answer) the fee-based version today (Autumn 2024) is already capable of "seeing", "hearing", "speaking". We mention also the Google Lens application which with multiple media conversion and with the aid of virtual reality translates the "seen" text into the requested language. The intense development of the outputs of single entities, looked at from subjective or objective points of view is also a tendency. It means that we get better and better and more humane textual replies through chat applications and that AI produces more beautiful pictures, and drawings, and in the area of sounds, music, and videos we can witness spectacular development.

It is known that AI-generated verbal and multi-media content is produced after entirely different editorial principles and input data than those contents produced by humans. Still, AI-generated inputs can be less and less differentiated from contents produced by humans. Often it is felt that the result would pass the Turing test which simply means that the inquirer would have difficulty deciding whether a human or a machine has replied (Turing 1950).

The investigation has proved that AI as a result of applications often appears in our daily lives and in a way that not even noticed that partial or target tasks are solved by AI (e.g. immediate evaluation of tasks (Ady–Terpecz 2018), (Kohnke et al. 2023), (Tölgyes 2023), controlling via sound commands, improving picture quality, etc.). The current basic research has not aimed at the detailed investigation of the possibilities from the educational or learning side but it is assumed that most applications can be fitted in and can be widely applied in support of learning and sometimes in teaching (Dietz 2024), (Shemesh 2023), (Simon 2023), (Solis 2023), (Xiao–Zhi 2023).

REFERENCES

Ady L., Terpecz G. 2018. Mesterséges intelligencia alkalmazása az oktatásban. Repüléstudományi Közlemények 30. évf. 1. szám. 111–126.

https://www.researchgate.net/publication/331929130_MESTERSEGES_INTELLIGENCIA_ALKALMAZASA_AZ_OKTATASBAN_-_APPLICATION_OF_AI_IN_EDUCATION. (viewed 2024. 12. 05.)

Dietz F. 2020. A mesterséges intelligencia az oktatásban: kihívások és lehetőségek. Scientia et Securitas 1. évf. 1. szám 54–63. https://doi.org/10.1556/112.2020.00009. (viewed 2024. 12. 05.)

Kohnke, L., Moorhouse, B. L., Zou, D. 2023. ChatGPT for Language Teaching and Learning. RELC Journal Vol. 54. No. 2. 1–14. https://doi.org/10.1177/00336882231162868. (viewed 2024. 12. 05.)

Laky Z. 2023. „Az embereket nem nehéz átverni" – a Google magyar tudósa a mesterséges intelligencia öntudatáról és veszélyeiről. [Interjú Szegedy Krisztiánnal.] Válasz Online 2023. 03. 10 https://www.valaszonline.hu/2023/03/10/szegedy-krisztian-matematikus-google-mesterseges-intelligencia-interju/. (viewed 2024. 12. 05.)

Shemesh, H. 2023. The Ultimate Guide for Using ChatGPT for English Learning. https://hadarshemesh.com/magazine/chatgpt-for-learning-english/. (viewed 2024. 12. 05.)

Simon, E. 2023. 6 Ways to Use ChatGPT to Learn a Foreign Language. https://www.icls.edu/6-ways-to-use-chatgpt-to-learn-a-foreign-language/. (viewed 2024. 12. 05.)

Solis, T. 2023. 9 Ways to Use ChatGPT for Language Learning. https://www.scribbr.com/ai-tools/chatgpt-language-learning/. (viewed 2024. 12. 02.)

T. Nagy L., Boda I. 2014. Nyelvtanulás és nyelvtanulási lehetőségek a web 2.0 közegében. In: Bátyi, Szilvia; Navracsics, Judit; Vígh-Szabó, Melinda (szerk.) Nyelvelsajátítási-, nyelvtanulási- és beszédkutatások. Papers in language acquisition, language learning and speech research. Budapest–Veszprém: Gondolat Kiadó – Pannon Egyetem Modern Filológiai és Társadalomtudományi Kar. 199–211.

Tölgyes L. A. 2023. A mesterséges intelligencia integrálása az oktatásba. https://ictglobal.hu/iparagi-megoldasok/a-

mesterseges-intelligencia-integralasa-az-oktatasba/. (viewed 2024. 12. 05.)

A. M. TURING 1950. Computing Machinery and intelligence, *Mind*, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433. (viewed 2024. 12. 09.)

Xiao, Y., Zhi, Y. 2023. An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks. Experience and Perceptions. Languages Vol. 8. No. 3. 212. https://doi.org/10.3390/languages8030212. (viewed 2024. 12. 05.)

SOURCES

ChatGPT. https://openai.com/blog/chatgpt. (viewed 2024. 11. 18.)

ChatGPT4. https://openai.com/research/gpt-4. (viewed 2024. 11. 18.)

DALL-E. https://openai.com/dall-e-2. (viewed 2024. 10. 18.)

DeepL. https://www.deepl.com/en/translator. (viewed 2024. 11. 25.)

Duolingo. https://hu.duolingo.com/. (viewed 2024. 12. 01.)

Google Translate. https://translate.google.com/?sl=auto&tl=hu&op=translate. (viewed 2024. 11. 25.)

Google Lens. https://lens.google/intl/hu/#translate. (viewed 2024. 11. 28.)

MidJourney. https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F. (viewed 2024. 11. 18.)

Natural Reader. https://www.naturalreaders.com/. (viewed 2024. 10. 18.)

Nightcaffe. https://creator.nightcafe.studio/text-to-image-art. (viewed 2024. 12. 08.)

Replicate. https://replicate.com/. (viewed 2024. 11. 18.)

Stable Diffusion. https://stablediffusionweb.com/. (viewed 2024. 11. 18.)

Voicemaker. https://voicemaker.in/. (viewed 2024. 11. 18.)

Xeropan. https://hu.xeropan.com/. (viewed 2024. 12. 01.)

■ **Článok bol recenzovaný. / Reviewed article.**