FORSCHUNG

DEEP LEARNING IN MULTIMEDIA

DR.-ING. MARINA GEORGIA ARVANITIDOU, SEBASTIAN PROKESCH

The advances in *Artificial Intelligence* (AI) and more specifically in Deep Learning, have hardly left any technical domain unaffected. The idea of Artificial Intelligence is not new: Although the first functional Neural Networks have been introduced in the late 50s, only recently did the field gain so much attention in the scientific and business world. What happened that suddenly Artificial Intelligence has gained so much attention? What is the impact on the multimedia landscape and what are the emerging challenges for broadcasters?

Die Entwicklungen hinsichtlich Künstlicher Intelligenz (KI) und insbesondere im Bereich von Deep Learning haben kaum einen technischen Bereich unberührt gelassen. Die Idee der Künstlichen Intelligenz ist nicht neu: Obwohl die ersten funktionalen Neuronalen Netze bereits Ende der 50er Jahre Gegenstand der Forschung waren, hat das Gebiet erst vor kurzem in der Wissenschafts- und Geschäftswelt so viel Aufmerksamkeit erhalten. Was ist passiert, dass Künstliche Intelligenz plötzlich so viel Aufmerksamkeit erlangt? Welche Auswirkungen hat dies auf die Multimedia-Landschaft und vor welchen Herausforderungen stehen die Rundfunkanstalten?

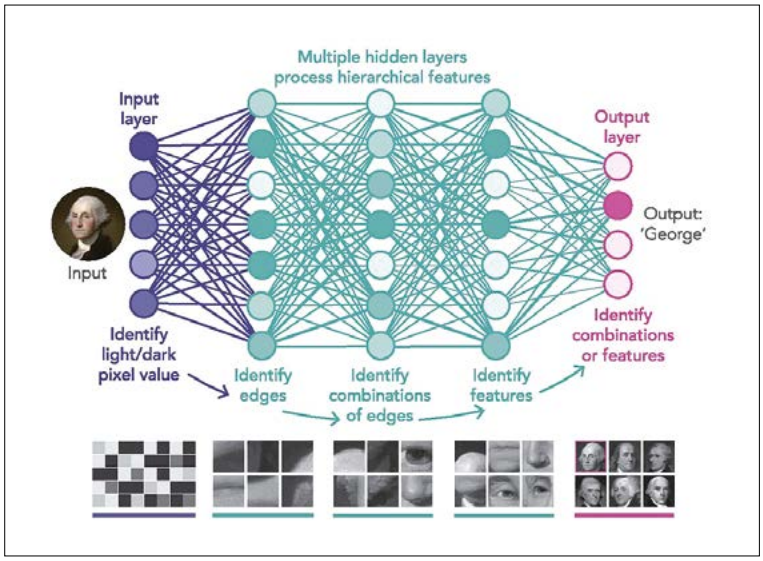


Figure 1: Conceptual design of a neural network for the task of image classification. The computer initially "understands" raw sensory input data just as a collection of pixel values. The role of the hidden layers is to extract increasingly abstract features from their input. The final layer is then able to recognize parts of objects in the image.

Source: M.M. Waldrop, "News Feature: What are the limits of deep learning?", National Academy of Sciences, 2019

Shortly before the beginning of the past decade, Fei-Fei-Li, professor at Stanford, launched [1], a publicly available database of more than 14 million labeled images in more than 20.000 categories. While, at that time, most of the AI research focused on improving or creating new models, professor Li focused instead on data and managed to expand available datasets used for training AI algorithms. This was a revolutionary contribution to the scientific world: despite the fact that internet is flooded with all kinds of images, labeling these appropriately is an essential requirement for training neural networks. At the same time, the significant increase of contemporary hardware capabilities (such as Graphics Processing Units and Tensor Processing Units) enabled the training of very complex and resource demanding models, namely the Deep Learning models that excelled in efficiency over other traditional computer vision approaches.

Deep learning is a subfield of machine learning, a technological field, which belongs to the more general field of Artificial Intelligence and includes technologies that enable computer systems to improve with experience and data. Deep learning methods are based on neural network architectures that contain multiple *hidden* layers (Figure 1). It is based on the philosophy of *connectionism:* while an individual biological neuron or an individual feature in a machine learning model is not "intelligent" on its own, a large population of these neurons or features acting together can exhibit intelligent behavior [2]. The fact that the number of neurons must be large, is essential and it is one of the key factors in obtaining good results. The success of neural networks nowadays is due to the dramatic increase in the size of the network that we can use today.

The above-mentioned introduction of ImageNet has been a milestone in the Deep Learning revolution. In the following years, it resulted in the conception of dramatically improved network architectures and has expanded in many application fields. This, on one hand, has given rise to new challenges and, on the other hand, promises significant improvements of journalistic workflows. In the following of this article, we discuss topics that emerged with deep learning and are shaping the media landscape: Deepfakes and the challenge to detect them, Natural Language Processing that enables powerful capabilities related to text, how these technological capabilities revolutionize Open Source INTelligence communities' workflows, as well as how deep learning can support effective data mining of the broadcasters' archives.

Deepfakes

The first thing that probably comes to mind when thinking of *Deepfakes*, is in the use of videos, such as the ones that first appeared a couple of years ago depicting celebrities, for example Barack Obama mouthing words that another person recorded. Deepfakes are a synthetic type of media, manipulated by Deep Learning algorithms, where the person saying or doing something is realistically replaced by another person. Neural network architectures such as *Generative Adver-*

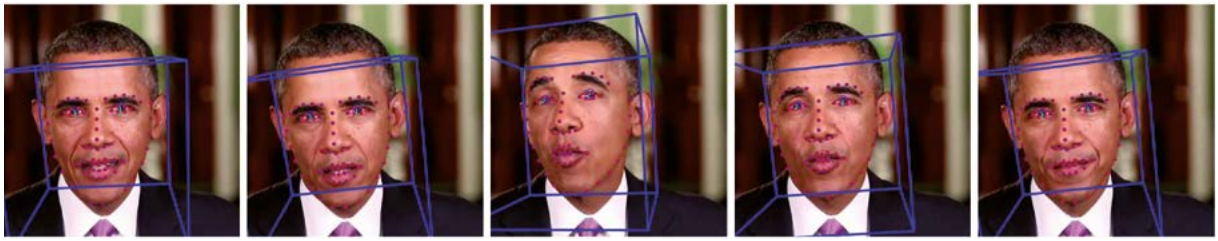


Figure 2: Facial and head movements in a video are tracked in order to model specific actions. Source: "Protecting World Leaders Against Deep Fakes", Agarwal and Farid, CVPR 2019 Source: Agarwal et al., "Protecting World Leaders Against Deep Fakes", IEEE Conference on Computer Vision and Pattern Recognition 2019

sarial Networks (GANs) are commonly used for the creation of Deepfakes. In order to make the DeepFake sound or look like a target person, one needs to train such a GAN with the target person's speaking voice, video or even just photos. And public personalities, who appear on numerous videos online, are an easy "target" to begin with.

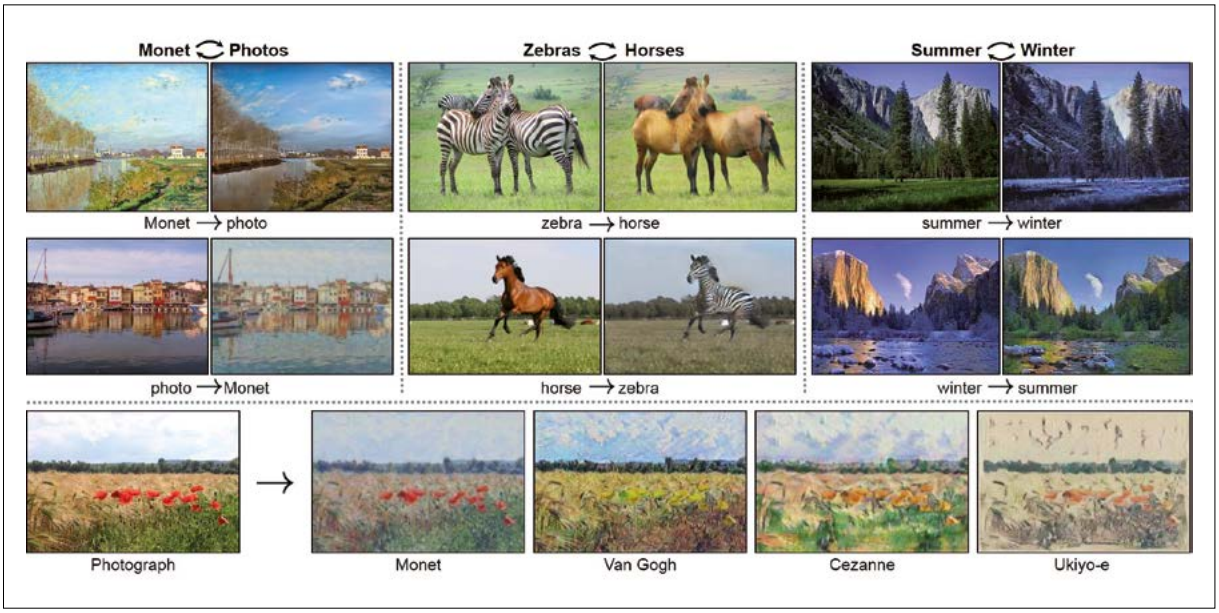
Deepfakes have been used to misrepresent politicians, mislead audiences, create harmful online content and have lately triggered growing concerns for political events such as elections, with social media and influencing platforms recently announcing considerable investments for the introduction of counter measures against Deepfakes.

The rapid advancements in technology to produce Deepfakes is rapidly evolving and the game between Deepfake creators and the experts who try to catch them resembles one of cat-and-mouse game. As Deepfake creators can access publicly available content and create any kind of new content, they tend to be one step ahead of the experts that try to detect them. As soon as a novel feature in the detection of Deepfakes is identified, for example that Deepfakes can be identified because people's mouth doesn't move naturally, a malicious programmer will quickly develop an algorithm to remedy that problem [3]. Key role in the successful detection plays again data: The lack of realistic datasets that can be used to test out new Deepfake detection technologies, is currently one of the biggest limitations.

In order to address this, Facebook has recently commissioned a first-of-its-kind dataset, using paid actors, for the AI community to use and launched the *Deepfake Detection Challenge* [4]. The challenge is in association with several universities and the goal is to find the best solution against Deepfakes which seem to be a real a problem in social media.

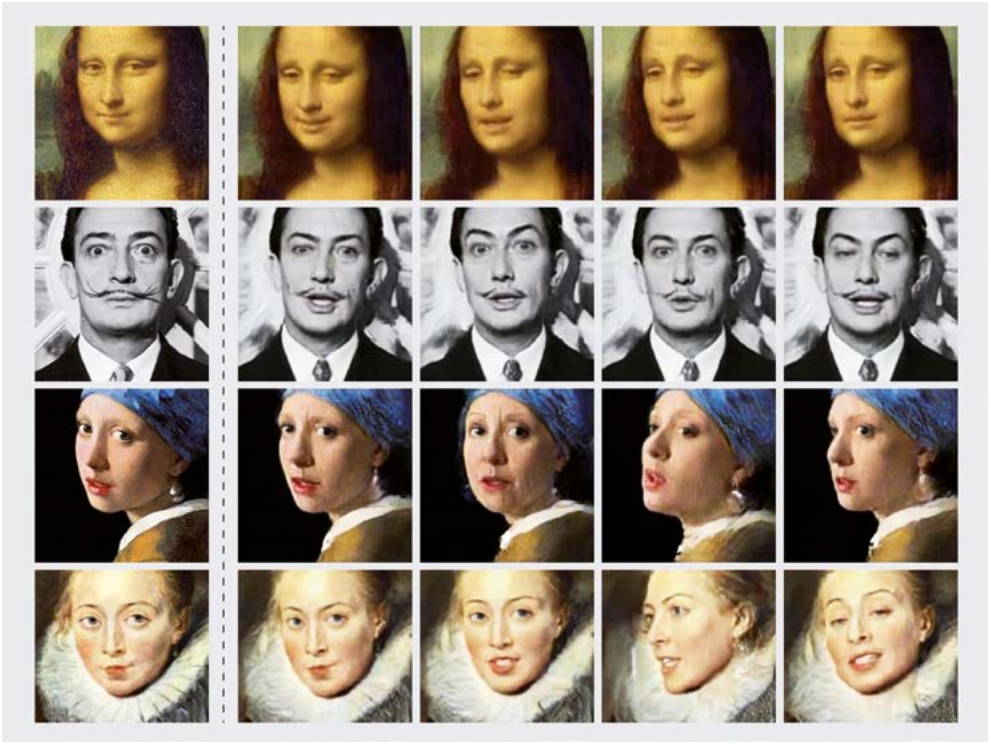
Of course, Deepfakes constitute only one facet of the general misinformation spread. Other mechanisms are for instance *fake news,* that are human-generated and *fake text,* that is generated by AI. In 2019, OpenAI, a non-profit research company, published a powerful model named GPT-2 *(Generative Pre-Trained model-2)* [5] which can generate text articles, namely fake text that are sufficiently convincing to be human ones. This release has given rise to concerns about potential misuse of the model and resulted in them releasing, at first instance, only a cut-down version of this model. These technological advances in such capabilities alongside the rising public concern about the impact of fake content are likely to have a major impact on the media industry in the coming years, as suggested by recent studies [6].

However, it has to be noted that the technology behind Deepfakes can also be used for legitimate commercial purposes, such as dubbing foreign-language films. For example, *Deep Video Portraits* [7], a video editing technique that uses machine learning techniques to transfer head pose, facial expression and eye movement of the dubbing actor on to the target actor to accurately sync the lips and facial movements to the dubbing audio, could save time and reduce costs for the film industry.



Source: JY Zhuet al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", IEEE International Conference on Computer Vision, 2017.

Figure 3: Example work of GANs present impressive results for style transfer, where the algorithm learns to automatically "translate" an image from one into the other and vice versa. Moreover (bottom) a natural photograph can even be rendered into the styles of famous artists.



Source: https://www.wired.com/story/deepfakes-getting-better-theyre-easy-spot/

Figure 4: "Mona Lisa smiled!" Deepfakes generated from a single image (painting).

Natural Language Processing

Media is all about communication and data. Language is a vital part of communication. *Natural Language Processing* (NLP) is the technological field that deals with machine processing (text reading, text generation etc.) and understanding of human (natural) language. Initially, NLP was handled by rule-based systems which used writing rules for grammar, sentence structure etc. Even though rule-based NLP has been an active research field for decades, it has gained remarkable attention lately thanks to the advances in Deep Learning.

In recent years, researchers have been showing that techniques with confirmed value in computer vision, may be useful also in the field of NLP. For example, *Transfer Learning,* i. e. pre-training a neural network model on a known task and then performing fine-tuning using the trained neural network as the basis of a new purpose-specific model, has introduced major improvements in the performance of NLP. The introduction of GPT-2 by open.ai and BERT by Google [8], triggered even more research interest in the last two years and enabled powerful performance of NLP related tasks. BERT *(Bidirectional Encoder Representations from Transformers),* a state-of-the-art language model for NLP, makes use of an attention mechanism *(Transformer)* that learns contextual relations between words in a text. One of the main breakthroughs of BERT is its *bidirectionality,* meaning that the model is able to learn the context of a word based on all of its surroundings (both left and right of the word).

One of the most popular and widely used application of NLP is language translation (e.g. English, French). Google translation service improved dramatically when, in late 2016, they switched from old phrase-based statistical machine learning translation algorithms to deep learning-based ones. In parallel, the introduction of the *Tensor Processing Unit* (TPU) helped towards the superior, almost human-quality, translation system dream becoming true. In addition to translation tasks, NLP tasks include language identification, text summarization, natural language generation (convert information into readable language), natural language understanding (machine reading comprehension) and speech recognition among many others.

Each NLP task may find applications in multiple fields in the media sector. Automatic subtitle creation (speech to text) used for accessibility purposes as well as archiving, automatically producing content, which is not based on a specific template, hate speech detection in user generated content, virtual assistants for customer service are just a few examples. Special language cases, such as spoken local dialects, are also important cases that can be supported by automatic speech recognition tools.

As expected, the rapid advances in NLP currently concern mainly the English language. The German speaking journalistic community faces currently the challenge to transpose and implement these advances to German. Nowadays, many newly established companies in Germany are working towards this goal, specializing in a broad band of application fields (such as banking, insurance and media). One of the main challenges towards this is the appropriate dataset creation. As the technology of machine learning is mature enough to perform adequately in specific tasks, the lack of appropriate annotated datasets that will enable learning of further capabilities, is vital. For this reason, recently many new startups have been established to focus on this specific goal: data annotation.

Open source Intelligence

User Generated Content (UGC) is continuously produced and it is accompanied by the challenge of moderating or exploiting it. On one side, there is growing awareness amongst the public, business and policy makers to the potential damage of harmful online material [6]. Online platforms are already taking measures to protect public from harmful content such as nudity, fake news, hate speech using algorithms that automatically detect them in newly appeared UGC. On the other side, the digital data wealth available is an extraordinary opportunity to gain insights and extract valuable information about behavior, trends and correlations that are "hidden" in large data volumes, from heterogenous data sources. This is what the teams of *data journalism* are pursuing: by combining the journalistic know-how on reporting, storytelling and finding stories, with the capabilities of automated approach-

Figure 5:
Natural Language Processing aims to improve communication between humans and computers.

Source: https://aliz.ai/

es, driven by statistics and machine learning techniques to process large data volumes in a more effective way. The in-depth examination of different kinds and sources of data, such as the routes of ships and planes, publicly available user data and public authorities' documents can lead to more concrete results and observations regarding specific topics of interest, or even reveal hidden issues that were not deemed important from the beginning.

Furthermore, investigative journalists, who rely on Open Source INTelligence (OSINT) also can benefit from the capabilities that the above-mentioned technologies offer. Open Source Intelligence is the collection and analysis of information that is gathered from publicly available, open sources. It has been initiated by efforts of investigators, such as Eliot Higgins of Bellincat [9], an independent international collective of researchers, investigators and citizen journalists using open source and social media investigation to probe a variety of subjects and has led the way in reinventing investigative journalism in the digital era. Groups like Bellingcat, have pioneered creative new ways of getting to the truth by exploiting digital resources. According to the *Harvard's Nieman Foundation of Journalism* [10]: *"An OSINT investigation is not one single method to get at truth, but rather a combination of creative and critical thinking to navigate digital sources on the web"*. The Foundation predicts that in 2020 OSINT will increasingly attract attention amongst journalistic communities. In line with this is also BBC's recent decision to make training journalists *"in the art of open source media"* a top priority. Therefore, an increasing number of journalists seem to become familiar with this method.

OSINT communities are researching and verifying available online content with the goal of trustworthy reporting, which is becoming even more critical today. To achieve that, research workflows that are automated or improved in accuracy are needed. Open source intelligence used for journalism builds on a wide range of digital data such as satellite imagery, social media, databases of wind, weather or any other form of data in order to better understand what happened at a specific place and point in time. Sophisticated machine learning algorithms can support the extraction of the desired information from these data sources in order to improve the collection of intelligence both qualitatively and quantitatively.

Emerging challenges for Broadcasters

The Institute für Rundfunktechnik (IRT) is focusing on identifying and understanding how AI-driven automation can improve the efficiency of algorithms (i. e. quality of result, execution time or both) as well as opening innovative ways for tackling new challenges. Deep learning, offers new opportunities in automatic recognition of features in data, allowing the analysis of complex data inputs such as human speech, images, videos and text. Some technologies are more mature than others, such as recommendation systems and speech-to-text conversion, others are relatively new, such as **fake news detection,** automated **summary generation** or automated **article generation**. Moreover, there is also often the need for tailored solutions, for example the transcription of spoken dialects that require the creation of annotated datasets before training the neural network. In this way, automation can serve in freeing up journalists' time so that they can focus more on creative work. In the following, we identify and discuss how the latest technological advances in the field of deep learning can support journalistic workflows.

Broadcasters archive all of their productions and thus have built up a considerable collection of digitalized material. Archivists have always evaluated and annotated the content with meaningful descriptions and keywords. However, despite all their efforts, editors often have difficulties to find images, audio clips or videos that they can reuse in current productions. Either the search engines used are too simple and do not match content that is similar, or the right metadata is missing in the archive because the needed annotation was not considered as important at the time and therefore omitted.

Data mining with machine learning and deep learning tools offer the chance to re-index the archive material and annotate it with even more metadata in a structured way. New intelligent

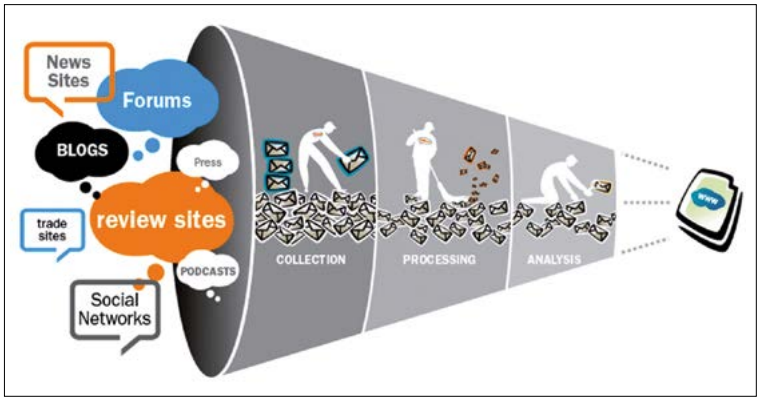


Figure 6: The concept of open source intelligence.
Source: http://www.osintux.org/

search and **recommender engines** can then retrieve more reusable information. This can offer editors relevant material to emphasize an argument in their story or give more information from past events and can save productions from retaking shots and thus reduce production costs.

There are different kinds of information that can be extracted from the content, such as audio and visual information. This information can be used for further analysis as well as for the **creation of metadata**. Information about the person appearing in the clip can be extracted from its visual information. Moreover, locations where the clip was shot or is supposed to play can be determined. Detecting objects that appear and analysing shot boundaries as well as detecting scene changes are also other features that are hidden in the visual part. With scene interpretation, the content of a scene can be described or for sport content, the detection of important events like scoring a goal can be marked by the exact time code. For audio, similar information can be extracted. For example, it is possible to **identify the person speaking** as well as when it does it using voice recognition. **Sentiment analysis** on the voice can also give information about the mood of the speaking person. Detecting objects that are either part of the scene or in the background may also add valuable metadata. One of the most valuable capabilities that deep learning offers is arguably **speech-to-text conversion**, since the generated transcripts can further be used with different text analytic tools.

All of these tools can be valuable but each of them presents limitations. The speech-to-text tools, for example, work especially well for English content. For German, although there are models that recognize standard German well, a lot of the content produced by broadcasters is in various **dialects** where these models cannot not perform sufficiently well. Because of the limited size of the market, vendors typically don't prioritize the training of dialectic content and therefore in-house models need to be developed. For that, existing models can be trained with a dialect training set that requires extremely large amount of annotated training data. An alternative approach for this is transfer learning, where a pre-trained neural network model on standard German can be used, and then perform fine-tuning, using a smaller training dataset based on the dialect. For object detection for example the recognized objects should match the ontology model used by broadcasters. To detect regional or even local landscapes a specific model needs to be used that learned the landscapes that need to be identified. To identify persons, the model needs to be provided with the respective facial- or speech feature vectors. To fulfil these needs, broadcasters need to build up own training sets and train own models.

New user interfaces are also essential, in order to support the usage of these metadata objects. User interfaces that enable fuzzy searching for given terms and present related results. Recommender engines can be used to retrieve relevant assets and to sort them. With the generated transcript, cropping an audio clip at the important part can be done by marking the sentence, which is easier than marking the right part in an audio wave form. A visual search interface could allow the user to search for similar pictures or, when applying **face detection**, to search for the person of interest. Towards this, we have built a browser extension to showcase how an augmentation of a video stream with additional information could look like. The idea was to give end-users more information about people appearing in a video stream. The same technology could be easily adopted in systems that journalists use for their research. The demo can be found at the IRT-Lab (https://lab.irt.de/in-browser-face-recognition-to-amend-information-on-appearing-persons/).

Ethical and societal aspects

The new technological possibilities that Artificial Intelligence brings are driving major social and economic changes. As with any new technology, the application of AI is associated with both opportunities and risks. There are concerns about rights violation and safety risks, related to the data needed for the algorithmic decision-making. In this context, the European Commission is currently setting out its plans for a common AI data strategy, for the coming years, in order to address (among others) specific issues of interest for the audiovisual media sector. Through the European Broadcasting Union, members have the opportunity to communicate their opinion in order to be taken into account in a consolidated manner in the ongoing works of EC on the topic. According to the EC's recent white paper on AI [11]: *"A solid European regulatory framework for trustworthy AI will protect all European citizens and help create a frictionless internal market for the further development and uptake of AI as well as strengthening Europe's industrial basis in AI."* , so that every developed solution will ensure that its outcome is not only reliable and beneficial, but also that at the same time it respects the ethical guidelines and law regulations.

References
[1] http://www.image-net.org/
[2] I. Goodfellow et al., Deep Learning, MIT Press, 2016
[3] https://actu.epfl.ch/news/epfl-develops-solution-for-detecting-deepfakes/
[4] https://deepfakedetectionchallenge.ai/
[5] https://openai.com/blog/better-language-models/
[6] Use of AI in online content moderation, Office of Communications UK, 2019
[7] https://www.bath.ac.uk/announcements/ai-could-make-dodgy-lip-sync-dubbing-a-thing-of-the-past/
[8] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
[9] https://www.bellingcat.com/
[10] https://www.niemanlab.org/2020/01/osint-journalism-goes-mainstream/
[11] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Source: IRT

DR.-ING. MARINA GEORGIA ARVANITIDOU

is lead R&D engineer artificial intelligence at the Institut für Rundfunktechnik (IRT).

www.irt.de

Source: IRT

SEBASTIAN PROKESCH

is research engineer at the Institut für Rundfunktechnik (IRT).

www.irt.de