[The original, German version of this article appeared on September 19, 2021, as part of the AI column "Aus dem Maschinenraum der KI" in the economics section of the German Sunday newspaper Welt am Sonntag, p. 23.]

[Translated with www.DeepL.com/Translator (free version) - AI technology made in Europe, see https://en.wikipedia.org/wiki/DeepL_Translator, and subsequently polished and modified by the author.]

Messages from the AI engine room

Artificial intelligence gives people their voices back

To talk to machines will soon be commonplace — Germany and Europe should have a say in this conversation.

By Kristian Kersting

Val Kilmer is one of my Hollywood heroes — not only because of his roles in "Top Gun" or "Batman Forever": In 1997, when I was still studying computer science in Freiburg, he gave his voice and body to the master thief Simon Templar in "The Saint". Actually, Simon Templar wants to retire. But then he falls in love with the scientist Emma Russell, from whom he is to steal a formula for "cold fusion". He switches sides and turn on the powerful billionaire Tretiak. In the end, Emma and Simon manage to get the formula made available to the whole world for free. Wow! Passion for science, solving the global energy problem, love, and action — what else is there to want for a young student?

Most impressive, however, is Kilmer's battle against throat cancer, which the now 61-year-old seems to have won in the meantime. Kilmer had a tracheotomy — opening a direct airway via a cut in the front of the neck and the trachea — and, like so many others, had to speak with a respirator that made him sound dramatically different and raspy. Thanks to Artificial Intelligence, or AI for short, he can now express himself with his "own" voice again.

Having machines speak like humans is a long-standing dream of human-computer interaction and AI. Getting machines to produce human-like speech, however, is not easy. The synthesized speech should be as close as possible to the human voice and easy to understand. The human voice, however, is highly complex and ticks quite quickly, with 16,000 or more samples per second and important features across different time scales. So far, only learning artificial neural networks with many layers — hence deep learning — has taught machines a good pronunciation.

To this end, spoken sentences of human speakers such as Val Kilmer are recorded. Indeed, you may not need a professional actor, but a certain quality of pronunciation does not hurt. The neural network then systematically looks at different sampling rates to, for example, capture as many cross-references as possible between thousands of different time points and, thus, the nuances in pronunciation that make a voice human. Then, to generate spoken language, a value is drawn for each time point from the probability distribution estimated by the neural network. This sample is then fed back into as input to the neural network to make a new prediction for the next time step, and so on. This step-by-step construction is computationally demanding but results in highly expressive and human-like voices.

This enables the medically indicated restoration of voices like Val Kilmar's and perhaps someday conversations with AI versions of deceased elderly relatives. The Shoah Foundation of the University of Southern California in the US is already taking this step. It has collected more than 55,000 audio-visual testimonies of survivors and witnesses of the Holocaust and uses them for the “Dimensions In Testimony” project, allowing visitors to ask questions that prompt real-time responses from the survivors in the pre-recorded video interviews.

Unfortunately, the public only knows these methods as "deep fakes" — the contraction of "deep learning" and "fake”. Recently, videos of Tom Cruise circulated on the web, showing the 58-year-old Hollywood star playing golf, as well as telling jokes and performing a sleight of hand. Back in 2018, we were all amazed when US comedian Jordan Peele put the words "President Trump is a total and complete dipshit" in Barak Obama's mouth. All deep fakes.

Yes, dual-use items and technology that can be used for both peaceful and military aims exist not only in chemistry, physics, electrical engineering, and many other fields. We should not be too afraid of deep fakes. One can only agree with Sarah Wachter of Oxford University in the UK: we need a differentiated approach. Individuals and society should be protected from deep fakes, but we should not ban deep fakes completely because of satire or freedom of expression and thus run blindly into a competitive disadvantage. The increasing commercial and cultural uses of deep-fake technology are very promising.

As a first step towards demystifying deep fakes, we should follow the suggestion of Australian AI researcher Toby Walsh from 2015 and label AI systems with a "Red Flag". According to Wikipedia, the Red Flag Acts were a series of Acts of Parliament in the United Kingdom regulating the use of mechanically propelled vehicles on British public highways to prevent accidents during the latter part of the 19th century. Not only did the law require a maximum speed of 4 miles per hour, but it also required a pedestrian carrying a red flag to walk in front of road vehicles hauling multiple wagons. Fortunately, we do not need a person to walk in front of AI systems.

Already today, Alexa, Siri, Duplex and Co. speak so naturally that you can hardly tell them apart from real people. They say "um" and "uh" as if they were thinking, even if they are not really doing so. Either way, talking to AI system will soon be commonplace, and Germany and Europe should have a say in this conversation.

Kristian Kersting is Professor of AI and Machine Learning at TU Darmstadt, co-director of the Hessian Center for AI (hessian.ai), and **winner** of the "German AI Award 2019". His AI column "Aus dem Maschinenraum der KI" appears regularly in the German Sunday newspaper Welt am Sonntag.