Contents

Hello, World!

The common first letters a programmer types into it's digital computing machine is some form of "hello world" along with braces, punctuation and words that the machine should interpret. After executing this short paragraph of text the computer likewise responds with "Hello, World!".

This ritual of greeting each other (and the world) is accredited to Brian W. Kernighan, who wrote the popular guide *The C Programming Language*[1] in 1978, where he explains that the only way to learn a programming language is by writing programs in it.

I wanted to start this introduction with the same greeting, as I am typing text into a computer and which letters simultaneously appear on the screen in front of me. But this translation from the keyboard of my laptop to the pixels on the screen is only possible by thousands of lines of code hidden in the source of the program. There are countless people who have written the software and operating system I am currently using to write this text. And there are even more people who wrote the software, the programmers have used to write the word processor I am using right now.

In this thesis I want to focus on the known and unknown human connections we make with and through computers. My hypothesis is that the current trend of framing "Artificial Intelligence" as autonomous computational beings, would be better described as "Co-Intelligence", where the programming of such systems is a collaborative effort with many hidden players. While I address some technical details of current machine learning systems, I will put them into social context, because software is inherently a social tool. And ultimately, as an artist, address these topics through aesthetic experiences, which go beyond logic and language.

A book is never written alone, in fact everything we do is influenced by others. And while I focus on the social nature of computing and I am influenced by many of my peers and created works with others, I am describing these topics from my subjective perspective. I don't claim any objective truths on the descriptions of the collaborative works I have done and am naturally biased towards my own contribution to it. As this thesis is written in the context of an art program, I tend to deviate from rigorous scientific methods and describe my observations more freely. I am certain that there will not be a definitive answer as there is not even an answerable research question, but even more questions will come up along the way.

Dreaming

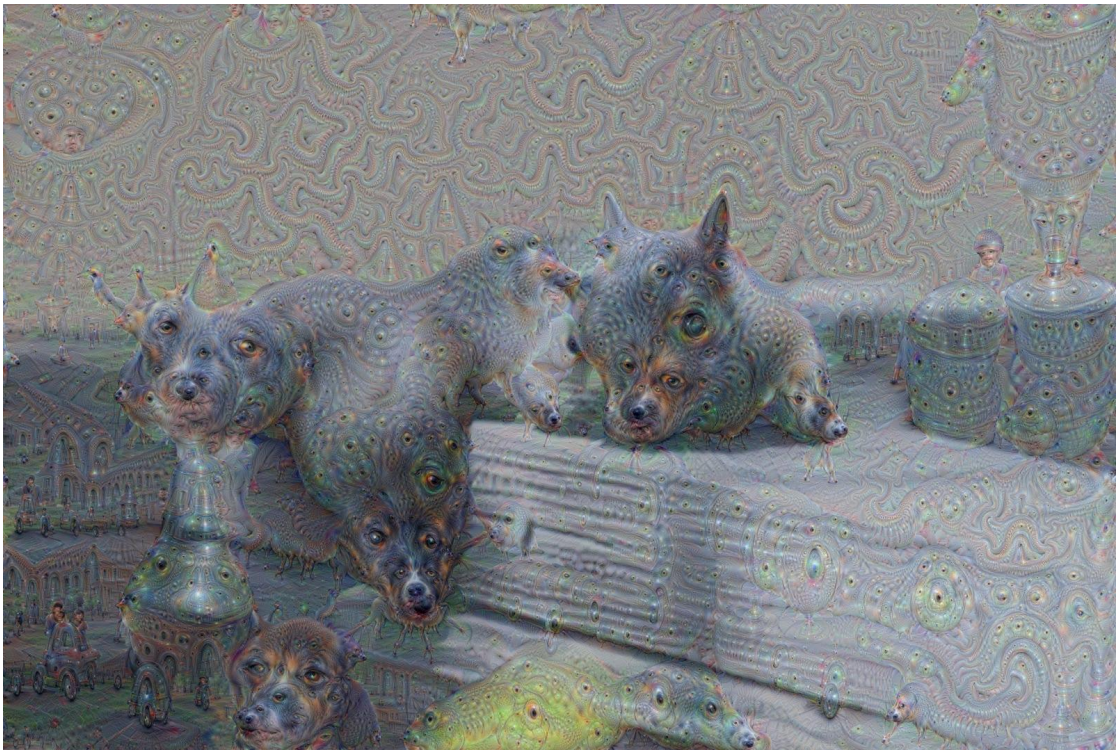

Figure 0.1: "PuppySlug" - posted by deleted user on /r/creepy with the title "This image was generated by a computer on its own (from a friend working on AI)" on January 10th 2015. 7 days before the Google blog post

My journey into the hype machine of "Artificial Intelligence" started around 2015, when Google released DeepDream[1] on it's research blog. Alexander Mordvintsev, Michael Tyka and Christopher Olah were initially trying to peek inside of an "Artificial Neural Network" that was "trained" for image classification tasks. They reversed the task of the classifier and instead of identifying what a set of pixels look like, they were changing the pixel values to look more like what the network has "learned". This created images that amplified the textures embedded in their networks, transforming the input image into swirls of PuppySlugs[2]. This imagery that reminded people of bad dreams and hallucinogenic trips was widely shared on the internet, which skyrocketed the awareness that "Deep Learning"[3] and biologically inspired computing has come back from it's hibernation in the AI winter. Since 2012 "Artificial Neural Networks" have made an incredible comeback into the field of computer vision. Where other researchers have long moved to different techniques for image classification Yann LeCun and Ian Goodfellow have kept researching on translations of simplified brain models to computer systems and finally hit a breakthrough, due to computational power and availability of data. The latter was compiled by a team around Fei-Fei Li, who created ImageNet, a dataset of

images organized into categories by anonymous workers from the internet. The ImageNet project held a yearly competition[4] since 2010 and for the first two years the best algorithm was recognizing these images with a 74% accuracy, until AlexNet[5], a convolutional neural network, topped the score with 85% and woke up the computer vision community to the potential powers of "Artificial Neural Networks". Finally with DeepDream this technique became tangible for the general public, as it created something to look at, something for humans to empathize with the outputs of an abstract mathematical model. This tendency to anthropomorphize this type of software becomes already clear in the way we use language to describe it. We talk of "Neural Networks" as if they are biological entities, that can be "taught" and which "learn", "experience" and "read" data from the world. This was a turning point that got me and many other artists interested in "collaborating" with seemingly autonomous machines. Machines that we do not have to understand through mathematics, but through interpretation and experimentation. Yet, the outputs of DeepDream quickly became kitsch. Google shared the code with an open source license and many people created apps and APIs to generate PuppySlug textures on top of their own images[6]. As with most other memes on the internet the hype around DeepDream died quickly and by the end of 2015 the world was already saturated with computer generated hallucinogenic images. Other researchers have since picked up[7] on visualizing nodes in "Artificial Neural Networks" and Alex Mordvintsev—together with his wife—has since become an artist, exhibiting in art fairs[8]. What stayed was the notion that "Artificial Intelligence" became some form of computational being, different from the humans, who wrote the code and this in turn created an immense interest for artists (and anyone, really) to create images, texts and sounds based on emergent algorithms and large datasets.

Terms

I have been deliberately using quotation marks around some of the previous words, as I am critical of the language used when talking of AI systems. In his book *The Artist in the Machine* Arthur I. Miller comes to the conclusion that machines could in fact be seen as creative and will be considered "artists, writers, and musicians in their own right."[9] He is using a typical techno-utopian argument that the technology is not quite there yet to *really* be creative, but that it will change in the foreseeable future. At the same time Miller is explaining in great detail the actual creative work of Mordvintsev and his human experience of insomnia to come up[10] with the generative system for DeepDream and how he shared it with his peers at Google. Alex Mordvintsev was not collaborating with his computer, he created an emergent algorithmic system with it. The software does not hang pictures on exhibition walls, talk to gallerists and has no agency in the process if it were to generate pictures or not. Peter Weibel has formulated it bluntly: Artificial intelligence does not exist. But an ensemble of machines, media, programs, algorithms, hardware, and software has resulted in an extraordinarily large, diverse, and productive field of research that is called AI.[11] Where Arthur Miller is feeding the narrative of

humanoid robots with glowing blue brains[12], most computer scientists in the field of artificial intelligence today are working in the subdiscipline of "Machine Learning". In the last 10 years the terms have converged in the media landscape, but where intelligence is rejected as a suitcase word with a multiplicity of meanings, "Machine Learning" tries to define the task more narrowly to some form of pattern recognition and extrapolation of existing data. They are using a variety of computational and statistical techniques to form abstract models for domain specific problems. In these fields "Artificial Intelligence" is a buzzword to convince governments and venture capitalists to fund projects. The artist and researcher Francis Hunger has recently shared a list of alternative terms in an attempt to dehumanize our language for AI systems:

1. 'Artificial Intelligence' => ' Automated Pattern Recognition'
2. 'Machine Learning' => 'Machine Conditioning' OR 'Automated Classification'
3. 'Neural Network' => 'Weighted Network'
4. 'Deep Learning' => 'Deep Conditioning'
5. 'Neuron' => 'Weight' or 'Node'
   -(**databaseculture?**) [13]

In an accompanying talk[14] he explains the aim of those terms is to invoke passivity, that we deal with machines and humans set those machines in motion, even when in the end we form "human-machine assemblages". I like many of the proposed terms, even though they are still closely associated to biological operations. The conditioning of machines, for example, reminds me of Pavlov's dog experiments or B.F. Skinner's box to modify pigeon behavior by reinforcement or punishment.[15] This analogy, though, serves the purpose of creating an image of the machine as a system, that can be controlled by changing the parameters of it's virtual environment. Therefor I will use "conditioning" and "weighted networks", where it is applicable. Using "Automated Pattern Recognition" instead of "Artificial Intelligence", however, becomes too narrow of a definition and is counterintuitive to me, as AI serves exactly the function of being ill-defined. Pei Wang has made a great effort in defining the different strands of AI research and their working definitions.[16] He clusters them into Structure-AI (recreating the human brain), Behavior-AI (recreating human actions), Capability-AI (domain problem solving), Function-AI (developing cognitive functions) and Principle-AI (finding underlying principles) and comes up with his own working definition to where AI research should be headed and how it can be unified. My initial goal for this thesis was to completely avoid the term "Artificial Intelligence", but as I have already failed in that task and coming up with a less anthropomorphic term does not seem feasible to convey the research in the field and it's media reception, therefore I will keep using the abbreviation AI. Contrary to articles and sci-fi novels I will not use the personification of "an AI", but will rather talk of AI systems, meaning emergent complex programs. Many of the systems today use statistical modeling, yet nobody would ask the question, if statistics can be creative, which is why I decided not to engage with philosophical questions of consciousness and creativity. Instead I want to explore how AI systems can include the human knowledge and work that goes into building them, which is challenging the AI ideology of machine

autonomy and proposes human-centric goals, rather than building a machine as a goal in itself.

Getting Started

[[Revision Needed]]

I structured this thesis around 4 chapters, which combine historical, computational and collective knowledge. Starting with a history of talking machines from Wolfgang von Kempelen's speech automatons to digital assistants today, not forgetting that the first computers were women doing calculation. The literal act of speaking to computers sets the base of defining that we are actually talking through computers with other humans. The second chapter deals with the building blocks for complex statistical modeling in AI systems. To make such systems possible, researchers need to create large datasets, often using ethically questionable techniques aggregating data from internet users. I will look closely into the StyleGAN dataset, as it serves a dual purpose, because the model defined another turning point for artists to generate synthetic media. The same is true for GPT-2, a model that can generate coherent looking text based on large amounts of scraped websites and books. In the third chapter I will explore the transformer architecture and how artists are using it to make (non-)sense automatic writing. Lastly I am revisiting collective experiences that I have been organizing with other artists. I had the great pleasure to work with the net culture initiative servus.at to organize the *Silicon Friend Camp* in the Austrian mountains, where we invited 17 artists and researchers for a week long retreat, focusing on human computer conversations. The camp resulted in the exhibition *Camping with Computers* and an online symposium on *Conversations with Computers*.

Talking to Computers

Female voices are spreading into living rooms and kitchens, on laptops, smartphones and small round speakers. Networked with computing machines around the world, they wait for their names to be spoken so that they can send your command to a far off data center and analyze it. If the transcription does not succeed algorithmically, sometimes a human takes over the task of understanding. This person listens to snippets of conversations for hours, anonymized[1] of course, and translates speech into text so that the machine recognizes the correct patterns in the recording the next time it tries. If the transcription is successful, the server sends the response back to the machine, which plays it in the friendly tone of a synthetic female voice.

Human Automata

Stories of artificial assistants already existed in ancient myths. For example, the limping Hephaestus built himself servants made of gold who assisted him in his work, could speak, and even had a mind of their own.[2] But it is only in the past few centuries that we have created the technological means to seriously address the construction of mechanical servants (at least virtually, as robotics is still far behind). At the height of automaton design in the 18th century, Frenchman Jacques de Vaucanson invented a mechanical duck that could not quack but appeared to have a functioning digestive tract. The mechanical attraction toured European noble houses and let its audience feed grains to it. However, what the duck excreted was a prepared colored porridge that was in a hidden container. This principle of mechanical trickery was also used by Vacaucanson's contemporary Wolfgang von Kempelen, who caused a sensation with his chess-playing automaton in the shape of a turban-wearing Turk. The illusion that the machine was acting autonomously was made possible by a small person inside a hidden compartment who controlled the puppet arm of the table via gears, levers, and pulleys. The hybrid machine is now the namesake of Amazon's *Mechanical Turk*, the largest platform for digital micro-labor, which lists click jobs for pennies. Today, it continues to perpetuate the illusion of autonomous machines with "artificial intelligence" that is covertly enabled by an army of underpaid workers.

Even though the chess playing Turk attracted attention, Wolfgang von Kempelen's scientific interest was in imitating human speech. He wrote down his investigations into phonetics in the work *Mechanismus der menschlichen Sprache* (Mechanism of Human

Speech) and built an apparatus with a bellows, rubber hose and a wooden nose with which it was possible to produce basic phonemes.

Among those influenced by Kempelen's book was a German tinkerer named Joseph Faber, who demonstrated his own mechanically constructed speaking machine in 1841. This attracted little interest in Germany and was presented and improved four years later in the United States as the Wonderful Talking Machine. This machine, as described by author David Lindsay, consisted of a bizarre-looking talking head[3] that spoke in a strange ghostly tone while Faber manipulated it with foot pedals and a keyboard[4]. For the inventor, the machine did not lead to the financial success he had hoped for, though it was presented as the *Euphonia* in London, where it at least delighted the father of telephone inventor Alexander Graham Bell and served as the boy's inspiration for his first talking machine.

For the exhibition Mensch-[in der]-Maschine at the ZKM (Center for Art and Technology Karlsruhe), media artist Michael Markert built *kII (Kempelen 2.0)*, an interactive installation in which visitors can playfully control a speech synthesizer by moving, opening and closing their hands.[5] In doing so, he brings Kempelen's speech apparatus into the 21st century with the help of an 8-bit PIC microcontroller and sensor technology. Like Kempelen's apparatus it alienates the voice in such a way that it creates mostly meaningless vocal sounds that enable new gesticulatory speech interactions.

The development of electricity certainly made new human-machine interactions possible. For example, the invention of the telephone and radio allowed the human voice to be transmitted over long distances. To optimize the transmission of speech, Bell Laboratories researched how to digitize the voice, for which they developed the vocoder (voice encoder). Demonstrated at the 1939 World's Fair in New York, the *Voder* omitted the speech input and transformation of the vocoder and allowed electrical synthesis of the voice via a console with 15 keys and a foot pedal.[6] The keyboard was operated by specially trained women, and in a recording advertised as the robot speaker, while it's unclear if they mean the machine or the woman, who is speaking through it.

Human computers were popular and necessary for war machines and research purposes in the 1930s and 40s. Mostly it was women who prepared mathematical tables, for example, for the use of ballistic projectiles. With the advent of the first digital calculators, female mathematicians, who were often denied higher scientific positions, were employed as programmers for the new universal electric machines.[7] The 6 people who programmed the first universal computer ENIAC include Betty Snyder Holberton, Jean Jennings Bartik, Kathleen McNulty Mauchly Antonelli and Marlyn Wescoff Meltzer.[8] Initially, the (mechanical) computer was programmed with punched cards and cables for specific operations. It soon became clear that programming complex systems required an abstract semantic language, for which reason the programming languages Fortran by John W. Backus, Lisp by John McCarthy, and COBOL by Grace Hopper were invented in the 1950s. The latter is strongly oriented to written English. Intended for business applications, it was the first attempt to use natural language for computer programming.

The second half of the 20th century saw the emergence of the myths about computers that we are familiar with today. Stories of anthropomorphic beings, like the board computer HAL9000 in Space Odyssey or Samantha in the movie Her. In both films, the disembodied voices become aware of their emotions and emancipate themselves from their human programming. Artist Tillmann Ohm makes this clear in his work *Reflections of HAL and Samantha*[9] by having the two artificial beings engage in a dialogue, cutting their original voice-overs together. While Samantha is convinced that the overwhelming and sometimes hurtful process of her learning algorithm improves the complexity of her emotions, HAL is consequentially interpreting them as errors in human programming and analyses the estimated malfunction.

Artificial Voices

I want to explain the advancements in voice synthesis since Wolfgang von Kempelen's speaking machine on actual dolls for children. The toy manufacturer Mattel released a doll with the name Chatty Cathy in 1959, which was similarly popular to the companies other best seller Barbie. Cathy's trademark was a string, coming out of the back of it's body, that could be pulled to wind up the mechanism of a simple phonograph. Like a record player it plays short strips with sentences like "I love you" or "Tell me a story". It was not the first toy using phonograph records, but it's success led to many pull string toys flooding the market well into the 60s and 70s. But obviously before dolls and puppets were designed to talk back to us, children and adults have been talking to figurines and other inanimate objects for a long time.

The systems engineer and science writer George Zarkadakis traces back the modern human mind to a pre-historic figurine of a lion-man (Löwenmensch), carved out of ivory and found in the cave Hohenstein-Stadel in southern Germany.[10] The figure dates back 40.000 to 35.000 years and clearly depicts a human body with an animal head. Because the figurine was found in a cave next to other carved objects, like beads and jewelry, researchers believe that the cave was either a storage or a place for shamanistic rituals. What fascinates Zarkadakis is that modern humans have existed 360.000 years before, but no evidence exists that they have been making art objects or figurative depictions before. Only during the upper paleolithic age it is assumed that we have created a general purpose language, when the first cave paintings were drawn and figures like the lion-man were carved. The evolutionary reasoning is that 'theory of mind', the seemingly unique human trade of projecting our own inner thoughts onto others to explain their behavior, gives the individual a social benefit, making it more likely that they reproduce. Projecting consciousness onto others did not stop inside of human groups, but placing our own reasoning onto animals seemed also beneficial for hunters coming up with elaborate strategies to take down larger animals together. In animist belief systems everything contains a 'soul' or 'spirit': animals, plants, rocks, rivers, the weather and maybe even words. George Zarkadakis argues that symbolic language predates the modern mind, and is not only used to communicate, but ultimately dictates how the

world is represented in our own consciousness. Only through language we place 'souls' and 'spirits' into everything around us and art objects fulfill the function to mirror our own minds physically and make connections with the world around us. It is no surprise then that the original artist of the lion-man created a hybrid of a human and a mountain lion, probably imagining themselves into the figure of the predator and using the object as a symbol to communicate.

In the beginning of "In Our Own Image" by George Zarkadakis, he tells the story of how the fictional character Robby the Robot from the movie *Forbidden Planet* impressed him as child so much, that he would imagine the robot as a playmate that could walk, talk and obey orders. When he later decided to study engineering and exchanged thoughts on artificial intelligence with other researchers, they all seemed to relate to the same sci-fi movies, books and stories.

Inspired by this revelation I asked the other participants at the *Silicon Friend Camp* of what their first contact with a robot was. My own memories brought me to a picture of myself, proudly holding a plastic toy robot in my hand. The doll—or should I say action figure—had a battery compartment, which was the power source for tiny wheels underneath it's feet and shiny light bulbs that it had for eyes. It could not talk, but I clearly remember that it made noises that resemble guns or a laser sounds. In the 80s and 90s the bleak vision of boxy humanoid robots carrying guns was very prominent, next to movies like *Terminator* and *RoboCop*, they fueled the male fantasy for power and destruction.

The other first encounters were a little less militaristic, one artist remembering a robot dog that he got as a present and was very proud to show around to others. In the early 2000s digital pets with integrated sensors became cheap enough for consumers. *Tekno the Robotic Puppy*[11] by Manley Toy Quest was one of the more affordable robodogs in the market and with light and proximity sensors, buttons and even a microphone it was well equipped to learn some 'tricks' by triggering the right combination of sensors. Another artist was sharing an image of the *Tamagotchi* and while it was not a robot, the LCD screen in an egg-like plastic shell displayed a simulation of a pet. With 3 buttons children could interact with the virtual pet and fulfill it's needs for food, affection and sleep. The anthropologist Pat Shipman proposes that animal connection can be considered a human trait that has uniquely evolved in us, next to tool making, symbolic language and the domestication of animals and plants.[12] Simulating animal connection by creating simple programs on microchips was then the next logical step in our drive to create symbols all the way down and fulfill the need of children to take care of others.

Yet another popular toy that was shared by an artist in our group and filled the same niche of robotic pets was *Furby*. This furry ambiguous creature with big eyes was made to simulate language acquisition and brings us back to the topic of how dolls learned to speak. In the case of *Furby* the designers decided to create a new vocal language called 'furbish'. When the device was first started it only repeated gibberish sentences and slowly over time would exchange these words with the local language equivalents as it ages. The original source code for one of the microchips was released in 2018[13] and even though I

can not read assembly code, it is thoroughly documented through comments. The game logic of the furby.asm file is fairly simple, jumping between subroutines depending on the state of the many sensors, triggers and timers and occasionally using a pseudorandom number generator to mix things up. The actual speech sounds were accessed from a memory unit and send to a cheap Texas Instruments speech synthesizer chip. These chips have been in use for some decades already, most prominently in another toy called *Speak and Spell*, but also in watches, clocks and translators. The synthesis uses a prerecorded audio signal that is then compressed using linear predictive coding (LPC). This type of encoding allowed the voice to be later changed in frequency, pitch and loudness. But just like the *Vocoder*, it only digitized a voice and was not able to construct new words or sentences. So the original *Furby* from 1998 could only express a couple hundred words that were recorded by a voice actor, which is again only a digital equivalent to Chatty Cathy.

More powerful computers were already able to generate any kind of text into somewhat intelligible speech. The Software Automatic Mouth (S.A.M.) from 1982 was one of the first commercial products that used rule based formant synthesis. Instead of using whole prerecorded words, it uses an array of computer generated phonemes and places them together to make up sentences. While this was often understandable, it was not perfect. It sounds unnatural and robotic and was reliant on converting text to speech in a probabilistic manner, often mispronouncing words.[14] S.A.M. did not need any special hardware and could run on Apple, Atari and Commodore computers at the time, even though that would often mean using every CPU cycle and holding other programs. The software had it's biggest moment at the launch event of the Macintosh computer in 1984, greeting the audience with a live demo in first person.[15] Sam and it's multiple software variations became the universal voice of the computer. The net.art pioneer Alexei Shulgin took the speech synthesis capabilities similar to S.A.M. to the extreme with his cyberpunk rockband *386 DX*.[16] The band is named after the Intel chip inside of the old office computer, which is also the front singer and instrumentalist. It uses MIDI and voice synthesis to cover popular songs by The Doors, Nirvana or the Sex Pistols, traveling to festivals around the world and sometimes performing on the streets. The artist Alexei Shulgin takes the role of the operator, only pressing the play button and creating the visuals on the screen during the live performances. In my opinion this ironic take of making the microchip into it's own one-computer-band was a demonstration of how terrible and reductive rock music becomes, when it was compressed into bits and bytes. At the same time the novelty and humor of this process, generated enough interest for *386 DX* to produce 2 CDs and distribute them over music labels.

The strangely robotic voices that got associated with computers improved over time and when Apple's digital assistant Siri came out in 2011 they used a different process of concatenating audio samples together. Before Apple even started working on the voice assistant, a company called Scansoft auditioned hundreds of voice-over artists in 2005 and made them speak a multitude of random sentences. The idea of how to use the random utterances for new outputs was quite simple: cut the voice into smaller pieces of syllables, demisyllables, phonemes, diaphones or triphones and put them back together

for any word you need. Doing this manually was—of course—very time consuming, so that a programs were developed to analyze and categorize the speech dataset based on their acoustic properties. Then, at runtime an algorithm selects the best sequence to create the desired phrase. The results in the early 2000s still contained a lot of error and needed a substantial amount of tweaking, but the outcome was perceived as much more natural. In 2006 the Scansoft merged with Nuance, another company working on enterprise speech solutions, who presumably used the database to create the voice for Siri and licensed it to Apple. The original voice-actor in the US was Susan Bennet,[17] who was oblivious of the fact that suddenly, 6 years after recording, her voice appeared on iPhones all over the country.

The technology to generate new sentences from previously recorded voices drastically improved, when deep weighted networks started to be used to analyze and select audio samples. One major achievement in generating natural voices came in 2016 in the form of a model named *WaveNet*[18] from a team of researchers at Google's DeepMind. They used convolutional node networks (CNN)[19] to perform statistical analysis on raw audio files with 16000 samples per second. CNNs were previously used to iterate over two dimensional pixel arrays for classification and generation tasks in computer vision, because it allowed for fast computation on graphics cards.[20] Using convolutional operations on temporal audio data meant that the researcher did not need to convert it first into a compressed format using a vocoder and the results were much more natural sounding than any previous text-to-speech algorithm. Using a raw audio dataset also means that the system can find patterns in any audio signal, including music. For demonstration the researchers generated new samples from a classical piano dataset, which sounds fascinating, but has more resemblance to free jazz than a classical piano score. This also happens to speech, when the network is conditioned without any text sequences it generates impressive human babbling, where real words are interspersed with made-up word-like sounds and breathing noises. When the network is trained on audio and the transcription together, the system does not only generate a consistent voice replica of a single human, but can also be used for voice recognition task transforming utterances back to text.

Other companies experimented with weighted networks to better create concatenative voice models. At the same year in 2016 Adobe showcased a hypothetical feature for their audio software, where users can create voice clones with a small 20-minute dataset and then use text to edit spoken words in a sentence. In the presentation project *VoCo* at the Adobe MAX conference, the developer Zeyu Jin used an example sentence and exchanged the words in such a way that instead of "I kissed my dogs and my wife", the audio was manipulated to say "I kissed Jordan three times" – referring Jordan Peele, the host of the conference. This strange demonstration, that was supposed to be funny, got a lot of attention online and created a debate around the ethical implications of manipulated audio and voice cloning. Adobe never released *VoCo* as a product, maybe because of the media backlash and fear from the public, but other companies like *Lyrebird*[21] filled the gap, trying to market their product for the use of editing podcasts and voice-overs. However, they started their media campaign by cloning the voices of

Donald Trump and Barack Obama to "discuss" about the startups product and praise it for their democratizing effort of making it available to anyone.[22] Together with synthetic video, so called "Deepfakes" have earned a bad reputation for being used in fraud[23] and pornography[24]. At the same time artists are using the technique to create hypothetical stories to mock politicians and other famous figures.[25] Zeyu Jin, the creator of *VoCo*, already proposed watermarking synthetic voices and use software to detect if a voice is fraudulent.

We can see that with the release of *WaveNet* and *VoCo* in 2016 a new field in voice synthesis opened up. Hybrid systems were created that used both concatenative and generative techniques to categorize and interpolate between the tiny audio samples of human speech.[26] The first release of *WaveNet* was computationally expensive, but already one year later Google implemented a new version that works in real-time for their voice assistant and has shown that a slim hybrid model can be used on low-powered devices, such as smartphones.[27] [28] Today, almost all digital assistants use a synthetic voice that is generated with deep weighted networks. Next to the feminine voice of Amazon's Alexa, they also provide the cloned voice of Samuel L. Jackson.

Conversational Agents

For Amazon founder Jeff Bezos, the board computer in Star Trek was the inspiration for investing in the cloud-based voice software Alexa. The product was initially marketed as a networked speaker. The software is now expanding to other items, including watches, smartphones, jewelry, light bulbs and doorbells. The aggressive price war with Amazon-connected products is partly to capture the connected home market, but also to collect as much natural voice data as possible. Over the past 10 years, deep weighted networks[29] have become popular for classification and pattern recognition tasks and have especially made an impact in automated speech recognition. Before weighted networks were used, the error rate of automated speech recognition (ASR) went from 40% in the 1990s down to 15% in the 2000s, which was still very high, but became practical to use. Back then the user would have to feed the system with voice samples, so that it could reliably turn utterances into text, but even then they would need to speak like a robot to be understood. With a landmark paper[30] in 2012 by research groups at the University of Toronto, Microsoft, Google and IBM it became that the application of deep weighted networks outperforms any of the previous techniques. In the same year Android released it's new cloud based speech recognition service[31] and since then users can speak into their phone naturally and be transcribed quite accurately. For a long time though, the models needed gigabytes of data in computer memory, to create accurate results, so that users need to send their voice samples to large server farms for transcription. Only in the past 2-3 years offline recognition became feasible through specialized hardware and model optimization, trading performance for robustness. Google's claim that new speech recognition systems have an error rate of less than 5% is only true for a specific—mostly

white and male—audience speaking American English and drops abruptly with a more diverse distribution of dialects, age and gender.[32]

In order for deep ASR systems to perform well they need a massive amount of input data. A paper from the Chinese tech giant Baidu proposes 12000 hours of speech to create a robust model.[33]bA global internal database of speech recordings enables the company to improve its speech recognition. The intrusion into privacy is immense and has already been used by U.S. law enforcement agencies as evidence in a court case.[34] But there is another, more ethical, way than stealing people's voices: The Mozilla Foundation's *Commonvoice* project relies on people voluntarily recording their voices for computer models, and the resulting speech recognition and synthesis software can be offered with an open source license.[35] Common speech and text datasets already have been publicly available from projects like VoxForge that uses the same crowdsourced techniques as Mozilla or LibriSpeech, which essentially scraped and aligned public domain audiobooks.[36] Other common non-free resources come from the Linguistic Data Consortium, who, for example, created *Switchboard*, a corpus of conversations between two US citizens randomly connected and recorded over the telephone line. Listening to the audio sample provided on the website, where a man talks to a woman about his interest in gardening creates oddly voyeuristic feelings.[37]

Artist Lauren Lee McCarthy plays with these tensions between intimacy and privacy, convenience and agency. In her projects LAUREN and SOMEONE, she installs connected devices into volunteers homes and either acts as a control system herself or lets others remotely monitor the volunteers and control the devices in their homes. This creates an interesting tension, when the person knows that there is an actual human listening and watching from afar. At the same time the artist and performers find themselves in a position of a helpful voyeur.

However, the role of human labor behind voice assistants is not just about executing and understanding commands. People tend to interpret voices and categorize them according to age, gender and social status. Companies take advantage of this and design their voice software according to certain identity schemes, which are provided with a history, hobbies and preferences. It is precisely this illusion that excites users and makes the product interesting. In a UNESCO think piece titled *I'd blush if I could*[38], they explore harmful gender biases associated with digital assistants. The voice assistants of major tech companies are scripted by default as female personas with smart, humble, and sometimes funny personalities. The teams working on voice assistants try to avoid this aspect, because Apple, Microsoft and Google ask their employees to refer to their headless voices as "it" and when the persona get's asked the question directly winds out with a joke. Only Alexa answers what is obvious in the default design of all of them with "I'm female in character". By being submissive, they thus support a patriarchal image of women that we already know, in a historical context, from human computers and other secretarial roles. The paper calls for women to be more empowered and involved in IT. It calls for AI software to avoid gender attributes whenever possible, and for AI assistants to take a clear stand against sexist behaviors.

Researcher and artist Nadine Lessio creates useless voice assistants to critique the current corporate agenda of productivity, efficiency, and consumption. She does this by using the programming interfaces provided to make apps for corporate voice assistants. For example, she explores the concept of a depressed home assistant with *SAD Home (Depressed Alexa 1.0)*[39], an Alexa hack that grants users their wishes depending on the weather and other mood factors, sometimes it simply turns itself off.

This scripted denial of a capitalist logic ironically uses the same technique as the company behind it. Voice assistants are carefully crafted by a team of creative professionals working in the field of "conversation design". In the book *Talk to me*[40] the Author James Vlahos describes that many people in the field are far away from computer science and more commonly had careers in the liberal arts. The teams are made up of authors, playwrights, comedians, actors as well as anthropologists, psychologists, and philosophers who imagine the personality of the AI persona that should represent the brand. To create the character, they have to come up with all possible questions and create various answers for each of them. Vahos recalls asking Microsoft's Cortana "Where do you come from?" and the female voice replies "I was made by minds across the planet".[41] And even though the designers decided to use the first person "I", they really are talking of themselves carefully crafting the answers played back by the loudspeakers around the world.

The tedious process of mapping out all questions and creating answers for them is mostly done through creative writing, but also utilizes careful statistical analysis of the questions users send to the cloud.

It is interesting how the term AI is used in the context of voice computer interfaces, because there is nothing "smart" about it, just a winding flowchart of if-else conditions. What might be clever is the nefarious way of how companies trick people into the belief of computer personalities, encouraging people to interact with the device like children playing with dolls. And when a company like Microsoft has experimented with a more sophisticated chatbot, like Tay[42], it started to repeat the racist and misogynistic slurs of twitter users and consequently has to be heavily filtered. But more on this in a later chapter on stochastic text generation.

Leewa

The artist duo *Ekheo* consisting of Aude Langlois and Belinda Sykora explore the crosslines of musical, visual and sonic art performances. They met at the Sound Studies program at the Universität der Künste in Berlin, where they created musical improvisations and theoretical frameworks on the voice and music. They started their research on voice, gender and identity in relation to AI systems in 2019 and held talks at the MUK[43] and other institutions to share their insights. At the *Silicon Friend Camp* they wanted to extend their already existing improvisation and research practice by integrating a form of embodied AI. But the main focus was the voice itself: How is gender perceived through

it? How do we navigate the tension between what is considered an artificial voice and an authentic voice? To tackle these issues during the camp they wanted to create and record soundscapes based on vocal improvisation and use AI systems in the form of synthetic voices, generative composition tools, robotic elements and so on. So we prepared the attic of the house with a small sound studio that they could work in and during the week they found inspiration in using a multitude of AI systems and human collaborators that they used to create an improvisation and finally a 6-minute video and audio composition. To come up with the identity of their AI partner, they used the *Oracle*[44] to ask questions of their name, gender, favorite instrument etc. which reads like this:

> You: how old is she?
> Oracle: Today she will be 11.
> You: what is her favourite musical instrument?
> Oracle: her greatest pleasure is to whisper to the sounds of trumpets and the beating of tympani
> You: is she happy?
> Oracle: It is a mystery and a great sadness.
> You: What is Leewa's surname?
> Oracle: It was suggested that her human's name was /u/Seattle_Celtics.
> You: does she like to sing?
> Oracle: I have no voice
> You: No I mean Leewa
> Oracle: Query : Have you ever cheated?
> You: Does Leewa have a voice?
> Oracle: Yoow!
> You: what is Leewa's favourite poem?
> Oracle: Leaves of Grass
> You: What is Leewa's favourite book?
> Oracle: Leewa does joke around and loves to give sarcastic responses. She also loves to read books, especially horror novels.
> You: What gender does Leewa have?
> Oracle: She is female.
> You: which music does Leewa like?
> Oracle: Leewa likes to sing and most of all Leewa likes to dance.
> You: What is art for her?
> Oracle: art is the process of iterating through consciousness generation.

This served as a base for the identity of *Leewa* and inspired further experiments using a whispering voice and 3D printing a female doll bust. To recreate the voice I searched for tools to do voice synthesis and found an open source toolbox from Corentin Jemine that claimed to do Real-Time Voice Cloning from a few seconds of audio.[45] The method is essentially based on a previous paper from Google engineers: *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis*[46] (SV2TTS), which itself

uses a pipeline of a speaker-encoder network (GE2E) to represent different voices in different positions of the latent space, a text to waveform synthesizer (Tacotron) and a vocoder to generate sound from these waveforms (WaveNet).[47] As described previously combining different encoding and generator systems into a larger framework is commonly used and is commercially used in Google's Assistant and their cloud services for other companies.

Thanks to the efforts of Corentin Jemine and the other contributors, who built a user interface to explore these tools, we were able to input our own voices and generate new samples with pre-conditioned voice models. Unfortunately though the results were not as expected, by inputting only a small number of samples the output was not picking up on or voice characteristics and the results were too noisy and of low-quality. Instead we tried out the cloud service from resemble AI[48] that is mentioned in the code repository and where some of the maintainers work at. We decided to try and make a collective voice of the camp participants, so that every one recorded 10 phrases.

After I pressed "train" on the web interface and waited for some time for the process to finish, I got an e-mail telling me that the "Resemble score" of our voice was 53%. So the transfer learning process was not quite successful, but the resulting voice was very surprising. I expected that the algorithm would try to average our voices into something that fits in between all data points. Instead it created clones of a voice for some phonemes, but not for others, so that a sentence randomly interpolates between different speakers. It also seems that, at least in this run, the characteristics of male voices are more prominent than female voices, but it also has a strange "artificial" nature to it, where the pitch and speed of the voice is constantly changing.

Now that we used a commercial product, we were limited to their API, but because of our limited time we only used their web interface, which made using the voice for real-time performances impossible. Instead *Ekheo* created some voice samples that they played live during the final performance. For this, they also recorded real voices from participants at the camp and intertwined with synthetic drones and other instruments. The visuals were created from a live video feed of the puppet, that was modified by Naoto Hieda using the live coding shader language hydra[49] The puppet itself could move it's head, which was remotely controlled by So Kanno via a servomotor.

In my opinion the example of *Leewa* illustrates the collective efforts that go into building "an AI" quite nicely: it shows the process of using open source and proprietary code developed by people reliant on previous research, the creation of collective datasets through the recording of our voices and finally the human performers hiding behind screens. At the same time *Leewa* is a self-referential project that makes the collective experience throughout the camp audible, by using our synthetic voice, the interaction with the *Oracle* and snippets of anecdotes from the camp.

Notes

0

1. Original Google Blog post. Link:https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

2. The initial classifier was conditioned on ImageNet, which contains many images of dog breeds. Therefore DeepDream is biased to generate textures of dog faces.

3. The idea of convolutional neural networks dates back to the 1950s and 60s to the works of David H. Hubel and Torsten N. Wiesel, who recognized receptive fields in the visual cortex of monkey brains. Ian Goodfellow calls the technique "Deep Learning", because the .

4. About ImageNet. Official Website. Link:https://image-net.org/about.php

5. Krizhevsky 2014

6. Mike Tyka has compiled a list of projects, released within a month of publishing the source codehttp://mtyka.github.io/code/2015/07/21/one-month-after-deepdream.html

7. A collection of tools and papers for feature visualization is collected at the github tensorflow/lucid repository. Note that Chris Olah and Alex Mordvintsev are contributors. Even more examples and explanations are collected in the interpretable machine learning book by the statistician Christoph Molnar, who also notes that "Feature visualizations give unique insight into the working of neural networks", but the "visualizations can convey the illusion that we understand what the neural network is doing".https://christophm.github.io/interpretable-ml-book/cnn-features.html

8. Interview with Alexander Mordvintsev by artnome. Link:https://www.artnome.com/news/2018/12/30/deepdream-creator-unveils-very-first-images-after-three-years

9. Miller 2019, 122.

10. Mordvintsev himself was inspired by a previous paper exploring the generative potential of CNNs from Karen Simonyan: Simonyan, Vedaldi, and Zisserman 2014.

11. Translated from the article 'AAA - Art, Algorithmen, Artificial Intelligence' at Kunstforum Bd. 278 Link:https://www.kunstforum.de/artikel/aaa/

12. Referring to typical AI stock images. Counter imagery is currently built byhttps://betterimagesofai.org/

13. Twitter post by (**databasecultures?**). Link:https://twitter.com/databaseculture/status/1413462059291975680

14. Unhype AI, Link:http://www.irmielin.org/unhype-ai/

15. In behaviorist psychology Ivan Pavlov's experiments with dogs is known as classical conditioning and B.F. Skinner who experimented on rats and pigeons using lever machines is called operant conditioning.

16. Wang 2019.

0

1. Metadata and references to the account are deleted before the review. Complete anonymization of the voice is not performed.

2. "Hephaestus then limped out of the door'; and maidens supported the ruler, golden ones, like living ones, with youthful charming education: These have understanding in the breast, and speaking voice, Have strength, and also learned art work from the gods." (Homer, Iliad 18, 417-420; link:https://www.projekt-gutenberg.org/homer/ilias/ilias183.html)

3. Fabers machine was first presented with a female mask in the USA and later in London under oriental motif wearing a turban.

4. D.Lindsay, 1997. Link:https://www.inventionandtech.com/content/talking-head-1

5. Serexhe et al. 2007, p. 74. And project description online. Link:http://www.audiocommander.de

6. A video recording of the Voder demonstration can be found in the AP Archive under Human Voice Machine.http://www.aparchive.com/metadata/youtube/5f098b1f3e8b4d09b8de30dcecc42f99

7. A focus on black women who worked as computers for NACA (NASA's predecessor) can be seen in the film Hidden Figures, 2016.

8. The story of the ENIAC programmers is told in the documentary The Computers, 2016.

9. Project description of Reflections of HAL and Samantha online. Link:https://tillmannohm.com/reflections-of-hal-and-samantha/

10. Zarkadakēs (2016)

11. The robot puppy was featured on the TIME magazine cover with the title "Tech comes to Toyland". Link: http://content.time.com/time/covers/europe/0,16641,20001211,00.html

12. (Shipman 2010)

13. Sean Riddle requested the Furby source code from the US patent office and uploaded it on his webpage. Link: http://www.seanriddle.com/furbysource.pdf

14. (Lukose and Upadhya 2017)

15. To make the demo work, engineers had to use a prototype Mac that was more powerful than the retail version. It said: "Hello, I am Macintosh. It sure is great to get out of that bag! Unaccustomed as I am to public speaking, I'd like to share with you a maxim I thought of the first time I met an IBM mainframe: Never trust a computer that you can't lift! Obviously, I can talk, but right now I'd like to sit back and listen. So it is with considerable pride that I introduce a man who has been like a father to me... Steve Jobs!". Link: https://www.folklore.org/StoryView.py?project=Macintosh&story=Intro_Demo.txt

16. The songs of 386 DX are available on Alexei Shulgin's website. Link: http://www.easylife.org/386dx/

17. Interview by the Guardian with multiple voice actors for Siri. Link: https://www.theguardian.com/technology/2015/aug/12/siri-real-voices-apple-ios-assistant-jon-briggs-susan-bennett-karen-jacobsen

18. (Oord et al. 2016)

19. Usually referred to as Convolutional **Neural** Networks

20. CNNs are used in AlexNet for example. (Krizhevsky, Sutskever, and Hinton 2017)

21. Lyrebird changed their name to Descript. Link: https://www.descript.com/

22. (*Politicians discussing about Lyrebird* n.d.)

23. Apparently fraudsters were able to use a synthetic voice to instruct a bank transfer of 220.000€. (Stupp 2019)

24. In 2019 an app with the name DeepNude got public attention due to a Vice article and was taken down by the author shortly after. Link: (Cole 2019)

25. Examples of deepfake political comedy can be found on the YouTube Channel of Speaking of AI. Link: https://www.youtube.com/channel/UCID5qusrF32kSj-oSGq3rJg

26. (Pierce n.d.)

27. ("WaveNet launches in the Google Assistant" n.d.)

28. (Wang et al. 2017)

29. As described in the introduction I use "weighted" instead of "neural".

30. Hinton et al. 2012

31. Google AI Blog article about "Speech Recognition and Deep Learning". Link: https://ai.googleblog.com/2012/08/speech-recognition-and-deep-learning.html

32. Tatman 2017, did a study on Gender and Dialect Bias in YouTube's Automatic Captions that shows a higher error rate for women and non-American speakers. Facebook's AI research department created

a dataset of people with different skin tones, gender and age to evaluate speech recognition models for biases and found significant differences between lighter and darker skin tones (Liu et al. 2021).

33. (Amodei et al. 2015)

34. The first Alexa recording in a court case was handed over after the defendant agreed handing over his data. The Independent, Amazon Echo could become key witness in murder investigation after data turned over to police, 2017. Link:https://www.independent.co.uk/news/world/americas/amazon-echo-murder-investigation-data-police-a7621261.html

35. Link:https://commonvoice.mozilla.org/de

36. Link to LibriVox: https://librivox.org/ and (Panayotov et al. 2015)

37. Link to the LDC Switchboard-1 Release 2: https://catalog.ldc.upenn.edu/LDC97S62

38. The title "I'd blush if I could" is also the response Siri gives to the insult "You're a bitch"

39. Nadine Alessio's project website. Link:http://nadinelessio.com/projects.html

40. Vlahos, 2020.

41. Ibid. p. 117.

42. Tay was the name of a chat bot Microsoft intended to have a teenage personality and could be interacted with over multiple channels. Twitter trolls co-opted the bot with sexist, racist and antisemitic questions that the bot replied to with generic answers. They also made use of a repeat-after-me phrase to make it look like the bot is spewing hateful comments itself. After only 16 hours Microsoft deleted all accounts and the PR disaster was immense, with headlines happily personifying the AI that has learned to be racist. But no continuous learning algorithm was involved. Link to a blog post by Russel Cameron Thomas explaining the technology stack behind Tay:https://exploringpossibilityspace.blogspot.com/2016/03/microsoft-tayfail-smoking-gun-alice.html

43. ("GENDER & DIVERSITY: Gastvortrag von Ekheo - Musik und Kunst Privatuniversität der Stadt Wien" n.d.)

44. The Oracle was a terminal computer connected to GPT-3 with a generic Q&A prompt.

45. From the github repository CorentinJ/Real-Time-Voice-Cloning. Link: (Corentin Jemine [2019] 2022). and his master thesis: (C. Jemine, n.d.)

46. (Jia et al. 2019)

47. The SV2TTS pipeline is based on these previous papers: (Wan et al. 2020) → (Wang et al. 2017) → (Oord et al. 2016)

48. ("Clone Synthetic AI Voices with Neural Text to Speech" n.d.)

49. Hydra was developed by Olivia Jack. Link: https://hydra.ojack.xyz

Amodei, Dario, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, et al. 2015. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin." December 8, 2015. http://arxiv.org/abs/1512.02595.

"Clone Synthetic AI Voices with Neural Text to Speech." n.d. Resemble AI. Accessed January 26, 2022. https://www.resemble.ai/.

Cole, Samantha. 2019. "This Horrifying App Undresses a Photo of Any Woman With a Single Click." Vice. June 26, 2019. https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman.

"GENDER & DIVERSITY: Gastvortrag von Ekheo - Musik und Kunst Privatuniversität der Stadt Wien." n.d. Accessed January 24, 2022. https://muk.ac.at/veranstaltung/gender-diversity-gastvortrag-von-ekheo.html.

Jemine, C. n.d. "Master Thesis : Automatic Multispeaker Voice Cloning," 38.

Jemine, Corentin. (2019) 2022. *Real-Time Voice Cloning*. https://github.com/CorentinJ/Real-Time-Voice-Cloning.

Jia, Ye, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, et al. 2019. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis." January 2, 2019. http://arxiv.org/abs/1806.04558.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Commun. ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

Lukose, Sneha, and Savitha S Upadhya. 2017. "Text to Speech Synthesizer-Formant Synthesis," 4.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "WaveNet: A Generative Model for Raw Audio." September 19, 2016. http://arxiv.org/abs/1609.03499.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "Librispeech: An ASR Corpus Based on Public Domain Audio Books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–10. South Brisbane, Queensland, Australia: IEEE. https://doi.org/10.1109/ICASSP.2015.7178964.

Pierce, David. n.d. "How Apple Finally Made Siri Sound More Human." *Wired*. Accessed January 23, 2022. https://www.wired.com/story/how-apple-finally-made-siri-sound-more-human/.

Politicians discussing about Lyrebird. n.d. Accessed January 24, 2022. https://soundcloud.com/user-535691776/dialog.

Shipman, Pat. 2010. "The Animal Connection and Human Evolution." *Current Anthropology* 51 (4): 519–38. https://doi.org/10.1086/653816.

Stupp, Catherine. 2019. "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case." *Wall Street Journal: WSJ Pro*, August 30, 2019. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2020. "Generalized

End-to-End Loss for Speaker Verification." November 9, 2020. http://arxiv.org/abs/1710.10467.

Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, et al. 2017. "Tacotron: Towards End-to-End Speech Synthesis." April 6, 2017. http://arxiv.org/abs/1703.10135.

"WaveNet launches in the Google Assistant." n.d. Deepmind. Accessed January 24, 2022. https://deepmind.com/blog/article/wavenet-launches-google-assistant.

Zarkadakēs, Giōrgos. 2016. *In Our Own Image: Savior or Destroyer?: The History and Future of Artificial Intelligence*. First Pegasus books hardcover edition. New York, NY: Pegasus Books LLC.