„YOU ARE THE ONLY PERSON IN THIS CONFERENCE“

Wie klingen (künstliche) Stimmen, die in der Cloud gespeichert oder generiert werden und über Kanäle wie Livestreams oder Medien wie Smart Speaker an unsere Ohren dringen? Im Folgenden geht es um vermeintlich körperlose Stimmen, die neben semantischen Aussagen (oder schlicht: akustischer Sprachausgabe) ihre technologische Bedingtheit vermitteln, wenngleich Rechen- und Übertragungsleistung aus Sicht der Entwickler*innen möglichst unbemerkt bleiben sollen. Dazu zählen sowohl avancierte virtuelle Stimmen aus dem Bereich des *Machine Learning*, die Nutzer*innen affizieren und an Geräte und damit an Services in der Cloud binden sollen, als auch operative Stimmen, die etwa eine Systemmeldung ansagen, wie es in älteren Betriebssystemen eine Popup-Nachricht auf dem Bildschirm war.

Wie lassen sich die verschiedenen Stimmqualitäten von menschlichen Stimmen und aktuellen Computerstimmen auseinanderhalten und ihre jeweilige Bedeutung verstehen? Dazu kann unterschieden werden in erstens eine von Menschen (mithilfe technischer Effekte) imitierte Computerstimme, zweitens aktuelle, AI-basierte Computerstimmen und drittens eine authentische menschliche Stimme in der Rolle einer zukünftigen Computerstimme.[1)] Für die zweite Gruppe ist bezeichnend, dass sich die Möglichkeiten von Natural Language Processing etwa seit der berühmten Computerstimme von Stephen Hawking, die sein Sprachautomat *Call Text 5010 Speech*-Synthesizer erzeugte, dermaßen weiterentwickelt haben, dass neue Prototypen in kurzen Demo-Dialogen von einer menschlichen Stimme nicht zu unterscheiden sind, wie zum Beispiel *Google Duplex* beweist. Daraus ergeben sich allerdings neue Probleme: Sobald die Maschinenstimme wie ein Mensch klingt, droht sie, die Hierarchisierung von Computerstimme und Mensch aufzuheben, wie im Fall von Duplex. In Callcentern etwa nimmt zunächst eine maschinelle Stimme den Anruf entgegen, sortiert in einer überschaubaren Frage-Antwort-Situation vor und füllt Wartezeit bis eine Kundenberaterin den wesentlichen Teil des Gesprächs übernimmt. Ähnlich läuft es beim Warten auf die Anderen in einer Videokonferenz: Automatische Ansagen haben maximal eine dienende Funktion. Anhand von Spielfilmen wie *Her* (Jonze 2013) mit sogenannten *voice actors*, die ihre natürliche Stimme einem Computer leihen, wird nachfolgend die Frage diskutiert, wie nah die Maschine den Menschen in Zukunft kommen soll. Eine genderneutrale AI-Stimme

1)
Die menschliche Stimme als fiktive Computerstimme hat eine lange Tradition im Film und in der Popmusik, zu deren Ikonen anfangs Kraftwerk zählten und zuletzt Hyperpop-Vertreterinnen wie SOPHIE, die mit Synthesizer, Vocoder oder Autotune ihre Stimmen mechanisch und virtuell erklingen ließen, um damit binäre Kategorien wie Mensch-Avatar oder Mann-Frau aufzulösen.

namens *Q* könnte hier einen Ausweg aus einer als mechanisch oder binär kritisierten Computerstimme oder als unheimlich empfundenen, allzu menschlichen AI-Stimmwelt zeigen.

„You are the only person in this conference:" Mit dieser automatischen Ansage begrüßt das Videokonferenzprogramm *BigBlueButton* die erste Nutzerin beim Eintreten in den virtuellen Konferenzraum, eine abstrahierte Fläche, die sich im Wesentlichen aus verschiedenen Dashboard-Elementen und potentiell aus Videofenstern zusammensetzt. „You are the only person in this conference" ist ein redundanter Satz, denn beim Blick in die Teilnehmer*innenliste steht dort neben dem eigenen Namen kein weiterer. Dass es sich hier um ein einseitiges Sender-Empfänger-Modell handelt und kein Kommunikationsangebot, wie von Smartphones und deren Sprach-Interfaces gewohnt,[2] unterstreicht die Antiquiertheit der Ansage. Hier können die Nutzer*innen nicht als „algorithmic subjects" auftreten und im Tausch gegen Daten digitale Dienste in Anspruch nehmen (vgl. Munn 2018: 84). Das Zeitalter des *Cloud Computing* mit seinem ausgeweiteten Dienstleistungsgedanken bleibt außen vor – „as-a-Service" (aaS) wurde im Zusammenhang mit *Cloud Computing* (z. B. Infrastructure-as-a-Service, Software-as-a-Service) eingeführt und auf Wohn-, Mobilitäts- und alle möglichen Angebote ausgedehnt. Aus technischer Perspektive handelt es sich bei der *BigBlueButton*-Ansage um den sogenannten *alone sound* – so wird die Audiodatei auf der Ebene der Systemeinstellungen bezeichnet. In diesem digitalen Seminarraum bleibt es zunächst still, es gibt keine hierfür produzierte Raum-Atmo, weder das Rauschen einer Computerlüftung noch das Schalten an den Knotenpunkten des Netzwerks. Es ist solange nichts Interface-Spezifisches zu hören, bis eine weitere Person durch das Knacksen ihres Mikrofons merklich den virtuellen Raum betritt und ihr Name in der Teilnehmerliste erscheint.

Im Unterschied zu virtuellen Assistenten oder automatischen Telefonsystemen handelt es sich hier nicht um eine Computerstimme, wie sich auf Nachfrage im Rechenzentrum der Universität herausstellt.[3] Stattdessen wird die Aufnahme einer menschlichen Stimme wiedergegeben, eine von mehreren wav-Dateien, die in der Audiokomponente des Open-Source-Videokonferenztools *BigBlueButton* mitgeliefert werden. Die namenlose Stimme klingt gefühlsleer, monoton und unpersönlich, wie jeder andere Sound eines Betriebssystems, das aus einem Computerzeitalter vor der Cloud stammt, in dem sich alles lokal auf dem Motherboard abspielte und als ‚User-Erfahrung' noch wesentliches Kriterium bei der Programmierung einer Software oder eines Interface zu sein

2) Bei Inbetriebnahme eines iPhone können Nutzer*innen vorgegebene Beispieläußerungen als Input ein sprechen, um Siri auf die persönliche Stimme einzustellen. Oder sie lassen es und benutzen Siri nie.

3) Auskunft per E-Mail erhalten am 8.10.2020.

schien. Ein Mensch hat für diese Sprachausgabe offenkundig also eine Computerstimme imitiert – beziehungsweise die Vorstellung, die es zur Tonalität von Computerstimmen gibt. Dies kann bei menschlichen User*innen für Irritationen sorgen. Um die möglicherweise Ärger oder ein Gefühl von Einsamkeit hervorrufende Ansage abzustellen, da sie in der Wiederholung und Redundanz stört, sah ein User im Forum nur mehr den Ausweg, die Datei umzubenennen, damit sie vom System nicht länger erkannt und abgespielt würde.[4)]

Die Ansage „You are the only…“ mag rudimentär erscheinen im Hinblick auf die größeren Entwicklungen im Internet der Dinge, die der Überzeugung folgen, dass Sprachinterfaces die Nutzung von smarten Geräten wie überhaupt den Alltag erleichtern, „a voicebased interface eliminates the friction that often accompany other technologies – the ‘pain points’ of picking up a smartphone, opening an app, awkwardly tapping out a search query“ (Munn 2018: 83). Interfaces passen sich an die Nutzer*innen an; umgekehrt werden Nutzer*innen aufgefordert, sich an die technologischen Bedingungen anzupassen, sei es durch Sprechtempo, Aussprache oder Aktivierungswörter (Ebd.: 107). Doch all das bedeutet nicht, dass die akustisch-sprachliche Interaktion zwischen Maschine und Mensch reibungslos funktioniert. Wie oft ist einem beispielsweise schon der Satz „ich habe Sie nicht verstanden“ von einem automatischen Telefonsystem entgegnet worden.

Vor einigen Jahren hat die Medienwissenschaftlerin Frances Dyson in *The Tone of Our Times* (2014) damals neu entwickelte Roboterstimmen unter ihren konkreten ökonomischen und medienökologischen Entstehungsbedingungen auf ihre räumliche und sensuelle Wirkung hin untersucht. Zu dem Zeitpunkt sprachen Roboter ohne Betonung.[5)] In der Entwicklungsgeschichte computergesteuerter Sprachverarbeitung gibt es lange schon das Bemühen, die monotonen Stimmen, die zum Teil in ihrer Affektlosigkeit berühren, in empathisch klingende Stimmen zu verwandeln. Dyson zufolge sind es die technologischen Entwicklungen, epistemologischen Fragen und kulturellen Diskurse aus einer jahrhundertalten Mediengeschichte, die den Klang der übertragenen Stimmen beeinflussen. Anfangs waren es Telefonverbindungen, die Nähe herstellten. Die künstliche Stimme hingegen hatten immer „a touch of the uncanny“ (Dyson 2014: 88). Bei dem Philosophen und Stimmtheoretiker Mladen Dolar gilt die Stimme ohne Körper als unheimlich, weil sie auf die Abwesenheit eines Körpers verweist. Die Erfahrung des Unheimlichen produzierte letztlich jedes Medium, solange es neu und noch nicht in sämtlichen Haushalten vorhanden

4)
„I also considered just renaming the audio file, and hoping that if Asterisk can't find the file, it won't play it“. (AdHominem 2012).

5)
Das entsprechende Kapitel in Frances Dysons Buch hat den Titel „Disaffected Voices.“

war. Doch dort, wo es keinen Körper gibt, tritt ein spielerischer Umgang an die Stelle des Unheimlichen. Wie YouTube-Videos zeigen,[6] haben viele Nutzer*innen die computergenerierten Stimmen von Siri oder Alexa schnell als Computernetzwerke mit fehlender Kognition vorgeführt.

Allerdings können immer noch Momente des Unheimlichen, die einen Kipppunkt markieren, auftreten: zum Beispiel dann, wenn eine neue AI-Stimme mit einem Mal allzu menschlich erklingt. Zuletzt ging dies im Fall von *Google Duplex* dem empfindsamen, menschlichen Publikum zu weit. So wurden nach einer öffentlichen Vorstellung des Sprachassistenzdienstes ethische Fragen diskutiert. Die Programmierer hatten der AI typische menschliche Verhaltensweisen in Gesprächen antrainiert, zustimmende Partikel wie ein „mh", offenbar um die *uncanniness* von *Duplex* zu minimieren. Die im Demovideo dargestellten Dialoge zwischen Maschine und Mensch irritierten deshalb, weil die Imitation der menschlichen Stimme durch die Maschine täuschend echt war. Würde aus *Google Duplex* in dieser Form eine Anwendung für den Alltagsgebrauch, wüssten menschliche Teilnehmer*innen nicht zweifelsfrei, dass sie mit einer Maschine reden. Zweifel und Kritik an diesem automatischen Programm wurden laut (Harwell 2018). Seit 2019 wird die Anwendung in den USA von Google genutzt, um Angaben wie Öffnungszeiten in Google Maps zu aktualisieren. *Duplex* setzt sich aus drei AI-basierten Komponenten zusammen, einem neuronalen Netz, einem Spracherkennungsprogramm und WaveNet, einem generativen Programm, das in Googles Schwesterunternehmen DeepMind entstanden und für die typisch menschlichen Diskurspartikel wie die „mhs" und „aahs" zuständig ist.

Nach dieser Kritik entschied Google, dass sich die AI-Stimme zu Beginn der Konversation als System von Google vorstellen sollte. Während Duplex-Anrufe bislang teilweise noch von Menschen durchgeführt wurden, um Daten für einen besseren KI-Trainingsprozess zu sammeln, übernimmt die Maschine aktuell laut Google 99 Prozent der Anrufe. Das hört sich dann so an: „Heyyyy, I'm calling from Google Maps. Given the current health situation, I want to update if you are open today. I'm an automated service that is recorded and monitored for quality insurance. When do you open and close today?" (Mrkšić 2020). Andere Anwendungen wie *Lyrebird* bietet User*innen an, die eigene Sprechstimme, etwa für Podcasts, Vorträge und Video-Voiceover, zu klonen (descript 2021). Für die Videokonferenzanwendung *Google Duo* wurde ein *auto complete*-Feature für Sprache entwickelt, in dem AI-Technologie fehlende Buchstaben oder Silben der Sprecherin ersetzt, sollte deren

6) Siehe beispielsweise Athena P (2020): The Replika (AI Friend) is… interesting? https://www.youtube.com/watch?v=T071GRkScrM (22.02.2021) oder South Park (2017): Alexa's the Coolest! Staffel 21, Folge 1. https://www.southpark.de/videoclip/6zs9up/south-park-alexa-s-the-coolest (22.02.2021)

Internetverbindung aussetzen. Es geht hierbei um Zeitfenster von 120 Millisekunden. Damit wäre dann auf der Audio-Ebene eine reibungslose, wenngleich roboterhafte Korrektur vorgenommen, die nach dem Prinzip von Textvervollständigung funktioniert und wie Duplex auf WaveNet basiert. Vorhergesagter Text wird als Sprache generiert und die Glitches im Livestream unterhalb der Wahrnehmungsschwelle korrigiert – zumindest in den Audiosamples (Barrera, Stirnberg 2020, MIT Technology Review 2020). Welche Auswirkungen diese automatische Korrektur im Detail haben kann, müsste im Vergleich von Programmierschnittstelle, sinnlich wahrnehmbaren Stimmen und Lippenbewegungen auf den gestreamten Videobildern ermittelt werden. So gesehen hat Dysons Ratschlag an die Medienwissenschaften – „to attend to the suspensions that modulate the enigmatic, ambivalent hyphen between human and post" (Dyson 2014: 91) – seine metaphorische Qualität behalten. Im Fall von Duo ist es banal, eine Technologie (Autocomplete) korrigiert die Fehler (Unterbrechung im Livestream) einer anderen.

Bei maschineller Sprachverarbeitung, wie zum Beispiel einer sprechenden AI, gibt es im Code geringe Möglichkeiten, die Stimme emotional aufzuladen. Aus der Entwicklungsabteilung von Amazon Alexa heißt es, „developers cannot change the prosody – 'you cannot control the stress and intonation of the speech.' [...] Small adjustments can be made using the <break> tag, specifying a pause in speech. [...] there is no possibility for lyrical readings, altered pitches, timbre shifts or abrupt volume and speed changes. Texttospeech establishes language as a particular set of universal parameters. This abstracted system provides maximum readability but simultaneously negates emotionality" (Munn 2018: 96). Emotionalität findet auf einer Metaebene statt, die solche berechneten Intelligenzen nicht imitieren können. Deswegen ist es nicht überraschend, dass in dem Film *Her* (Jonze 2013) eine AI des Betriebssystems mit der echten Stimme einer Schauspielerin spricht, die also mit sämtlichen Affekten belegte werden kann. Der Film *Her* kam zwar zwei Jahre nach der Einführung von Siri auf Apple-Telefonen und zwei Jahre vor der Markteinführung des Amazon Echo und damit dem ersten Gerät, das Alexa nutzte, in die Kinos; Alexa bestätigt jedoch scheinbar die Gewissheit der Prä-Alexa-Ära für *Her*: Für den Regisseur des Films war es naheliegend, dem OS im Film die Stimme einer menschlichen Schauspielerin zu geben.

Denn was die AI-Stimme noch nicht kann, nur die menschliche Stimme als Stellvertreterin einer vorgeblichen AI im Film, nennt Mladen Dolar „Nichtstimmen." „Nichtstimmen von Husten

und Schluckauf über Brabbeln, Schreien, Lachen und Singen sind ja offenbar keine sprachlichen Stimmen; sie sind keine Phoneme und stehen doch nicht einfach außerhalb der sprachlichen Struktur: Sie scheinen sich gerade in ihrer Nichtartikuliertheit [...] besonders dafür zu eignen, die Struktur als solche zu verkörpern, die Struktur in ihrer grundlegendsten Form – oder Bedeutung an sich, jenseits jeder konkreten, faßlichen Bedeutung“ (Dolar 2014: 48).

WOMAN-AS-INTERFACE In *Her* kann die gespielte AI sämtliche nicht-stimmliche Laute äußern. Filme wie dieser machen sich keine Mühe zu erklären, warum es sich bei den Stimmen keineswegs um computergenerierte handelt. Sie behaupten einfach, in einer unbestimmten Zukunft würden Maschinen wie Menschen klingen. Doch *Her* hebt sich als aktuellstes Beispiel für das algorithmische Imaginäre ab von anderen Beispielen. Anders als in der Mediengeschichte hat die Stimme von Scarlett Johansson zwar dienenden Charakter, ihr vernetztes Dasein gibt der Handlung jedoch eine unerwartete Wende, als sich herausstellt, dass diese Dienste nicht in einem *One-on-one*-Verhältnis geleistet werden. Das OS, in dessen Stimme sich der menschliche Protagonist des Films verliebt hat, kommuniziert parallel mit vielen anderen Menschen und ebenso weiteren Ausprägungen des OS. Insofern eröffnet die Handlung von *Her* den Blick in eine technologische Zukunft mit einer Dienstleistungsgesellschaft, in der Dienste in alle Richtungen verrichtet werden, von Maschinen für Menschen, von Menschen für Menschen und Menschen für Maschinen.

Um wieder zum Ausgangsbeispiel zurückzukommen: Der Klang der „You are the only person“-Stimme lässt sich als weiblich einordnen. Die Verwendung einer weiblichen Stimme im digitalen Service- und Assistenzbereich ist ein hier digital animiertes, sexistisches Klischee. Persönliche Assistentin ist ein Beruf, der in der westlichen hierarchischen Angestellten- und Bürokultur in der Vergangenheit überwiegend von Frauen ausgeübt wurde (und auch in der Gegenwart oft weiter wird); als künstliche Version lebt die Vorstellung der dienenden Frau nun in Navigationsgeräten oder Smart Speakern weiter und wird auf Chips gespeichert.

Mit der Einführung von ISDN[7)] Anfang der 1990er-Jahre waren die technischen Voraussetzungen für Videotelefonie geschaffen. In der Zeit machte die Telekom mit drei Bildtelefonanlagen parallel Werbung für das neue ISDN-Netz und die damit verbundenen technischen Lösungen. Dabei ließen sie das Stereotyp von historisch weiblich besetzten Stellen in der Telefonzentrale oder im Sekretariat aufleben, indem sie die Geräte mit weiblichen

7) ISDN ist der Internationale Standard für ein digitales Telekommunikationsnetz. Im Jahr 1979 entstanden erste Pläne und 1982 fiel die Entscheidung für die ISDN-Technik. Seit 1989 war ISDN in Betrieb. Die Telekom hat diesen digitalen Standard Ende 2020 abgeschaltet.

Vornamen markierten. Das Modell „Christa R“ bestand zum Beispiel aus einem Monitor mit beweglicher Kamera, einer Telefoneinheit und einem Videocodierer. „Christa R“, „Lisa C“ und „Claudia O“ sollten die Kommunikation in den Büros verändern (Lieb 2020). Die Telekom gab ihren Modellen weibliche Identitäten, die wie zum polizeilichen Schutz der Privatsphäre vom Nachnamen nur den ersten Buchstaben preisgaben. Die Abkürzung eines Nachnamens mit dem Initial ist eigentlich nur aus der medialen Berichterstattung über Kriminalfälle und anschließende Prozesse bekannt; die Identität von Opfern und Tatverdächtigen wird durch die Anonymisierung geschützt, entsprechendes fordert etwa das deutsche Presserecht. „Da die Telekom bzw. die Post davor und danach entweder auf Fantasienamen (Xitel) oder eher technische Bezeichnungen zurückgriff, wirkt es noch auffälliger, dass bei diesen drei Modellen diese Namen gewählt wurden.“[8] Sadie Plant kritisierte Tätigkeiten am Beispiel der Telefonistin, die Gesprächsverbindungen herstellte und dabei zum Interface zwischen Mensch und Medium wurde, als „woman-as-interface“ (Munn 2018: 99-100). Im Fall der Telekomgeräte ist die Frau zum Interface-Objekt geworden.

Der Sexismus der IT-Branche wird in den vergangenen Jahren häufiger, öffentlicher und vehementer kritisiert.[9] Aus heutiger Sicht wirkt ein über die ISDN-Technik und ihre Einsatzmöglichkeiten informierender Telekom-Clip[10] von 1992 skurril, antiquiert. Wenn die Videotelefonie vorgestellt wird, muss die Frau in der Rolle der immerhin weiblichen Auftraggeberin mit dem Auftragnehmer flirten. Sie lacht amüsiert, ohne dass aus dem gefilmten Gespräch deutlich wird warum.[11] Es ist ein stereotypes Rollenspiel, mit dem der in den Apparaten angelegte Sexismus auf Anweisung der Regie performt wird.

NON-BINARY

Aus Gründen von binären Default-Einstellungen wie diesen hat sich die Non-Profit-Initiative Feminist AI[12] in Los Angeles damit beschäftigt, die vorinstallierte, weiblich gegenderte Alexa-Stimme durch eine von den Nutzer*innen selbstgewählte oder neu programmierte Stimme zu ersetzen. Hier gilt nicht mehr, was Mladen Dolar noch über das Verhältnis zwischen Anrufbeantworter und Stimme schrieb, dass ein Gerät als „Stellvertreter [der unsichtbaren, abwesenden Quelle] aufzutreten beginnt“ (Dolar 2014: 86). Hier wird die Stimme zur Stellvertreterin für eine nonbinäre Sprachpolitik. Um dieses Ziel zu erreichen, haben Feminist AI mit einem sogenannten Wizard of OZ-Test, einer gängigen Methode in der Entwicklung eines neuen Interface, versucht,

8)
Matthias Lieb, Mitarbeiter am Museum für Post und Telekommunikation in Frankfurt am Main, war im Sommer 2020 mit einer Archivrecherche zum Thema Videotelefonie beschäftigt. Diese Aussage machte er in einer E-Mail an die Autorin vom 21.7.2020.

9)
Unter anderem kritisiert die Autorin Anna Wiener die patriarchale Geschlechterordnung in *Uncanny Valley* (London, 4th Estate, 2020), einem Bericht, der auf persönlichen Beobachtungen und Erfahrungen als Mitarbeiterin in Tech-Startups im Silicon Valley der Zehnerjahre beruht. Spezifisch mit der Inszenierung von KI-Assistenzsystemen beschäftigen sich Martin Hennig und Kilian Hauptmann unter dem Aspekt von Macht und Gender in ihrem Beitrag „Alexa, optimier mich“ KI-Fiktionen digitaler Assistenzsysteme in der Werbung,“ (Zeitschrift für Medienwissenschaft 21, Künstliche Intelligenzen, 2/2019, S. 86-95).

10)
1992 bewirbt die Telekom das digitale Netz aus Glasfaserleitungen für die Übertragung von Sprache, Daten, Texte und bewegte Bilder in informativen Werbevideos „Alles über ein Netz,“ vgl. https://www.youtube.com/watch?v=1IYFr-C_SKg (22.02.2021). Ab 14 Min 31 Sek wird der ‚Clou‘ vorgestellt, ein Bildtelefon, ein Computermonitor mit der entsprechenden Software, Kamera und einer Tastatur, über die Bildschirmeinstellungen vorgenommen werden können (Tasten für Bild hell, Bild dunkel, Eigenbild und Dokument).

11)
1992 bewirbt die Telekom das digitale Netz aus Glasfaserleitungen für die Übertragung von Sprache, Daten, Text und bewegten Bildern in informativen Werbevideos. Auf YouTube ist eine Kompilation zu finden. Siehe „Alles über ein Netz,“ YouTube, ab 15 Min 26 Sek. https://www.youtube.com/watch?v=1IYFr-C_SKg (22.02.2021)

12)
Feminist AI (2016): Thoughtful Voice Design. https://www.feminist.ai/thoughtful-voice-design

die Affekt-Reaktion von Nutzer*innen zu ermitteln. Mit diesem Zwischenschritt im Entwicklungsprozess wird Geld und Zeit gespart, indem die ‚Erfindung' oder das ‚Feature' zunächst von einer realen Person, die sich in einem anderen Raum befindet, simuliert wird, was die Tester*innen weder wissen noch sehen können. Die Testpersonen gehen in der Situation davon aus, dass sie mit einer künstlichen Intelligenz kommunizieren. Bei diesem Vorgang geht es darum herauszufinden, ob der Service den zukünftigen Nutzer*innen behagt. Erst nach solchen Testergebnissen wird für Programierer*innen deutlich, mit welchen Maßnahmen sie an den Programmierschnittstellen und auf der Code-Ebene intervenieren könnten.[13)] Wobei die Möglichkeiten der Intervention von außen begrenzt sind – „for a real intervene, you need to work in a company."[14)] Feminist AI stellten im Rahmen ihrer Intervention fest: „Participants indicated they would like the ability to modify the output of the voice when private information is collected" (Feminist AI 2016). Hier könnte eine Vielfalt an Stimmen für Diversität sorgen.

Neben Feminist AI gibt es u.a. auch Projekte wie *Q*, eine als geschlechtslos bezeichnete Stimme, die in internationaler Zusammenarbeit entwickelt wurde[15)] und als Maschinenstimme erkennbar ist. Während in dem Film *Her* noch eine Geschichte erzählt wurde und mit einer fiktiven OS-Stimme die Frage verhandelt wurde, wie eng das Mensch-Maschine-Verhältnis sein könnte oder eines Tages sein würde, deutet das Beispiel Google Duplex die Anmutung einer universellen Intelligenz an, obwohl sie auf einen einfachen Dialog zum Sammeln von Informationen programmiert wurde, doch die Diskurspartikel suggerieren mehr.

In *Her* wird die alte Utopie der Verschmelzung von Mensch und Maschine zur Dystopie,[16)] distanziert sich damit aber von alten Stereotypien. Ist in dieser Verbindung die Maschine defizitär oder ist der Mensch defizitär? Die Maschine muss vernetzt sein, sonst funktioniert sie nicht als Knotenpunkt in den Netzwerken. Das OS ist eine immaterielle Stimme ohne Körper, ohne Gehäuse. Navis, Siris und Alexas sind die Töchter der Telefonistin, die Ferngespräche herstellte, bis der Switch erfunden worden war.

Neben dem problematischen Einsatz von Smart

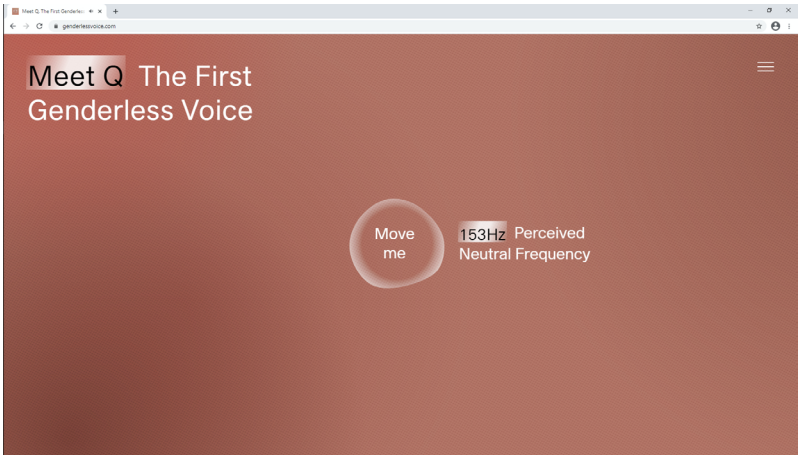
13) Vgl. zur theoretischen Reflexion von Interventionen in digitale Kulturen, Galloway, Alexander R. (2017), Intervening Infrastructures: Ad Hoc Networking and Liberated Computer Language. In: Caygill, Leeker, Schulze (Hg.), Interventions in Digital Cultures. Lüneburg, meson press 2017, S. 61-72.

14) Mercedes Bunz im Q&A nach ihrem Online-Vortrag „On the Culture of Artificial Intelligence", veranstaltet vom Museum Brandhorst und dem Zentralinstitut für Kunstgeschichte der Universität München, 27.1.2021. https://www.zikg.eu/aktuelles/veranstaltungen/2021/online-vortrag-mercedes-bunz (22.02.2021)

15) Siehe https://www.genderlessvoice.com/ (22.02.2021). Die Linguistin Selina Sutton kritisiert in einem Paper, dass es sich um eine geschlechtlich uneindeutige Stimme handelt, nicht um eine geschlechtslose. Siehe Sutton, S. J. (2020). Gender Ambiguous, not Genderless. *Proceedings of the 2nd Conference on Conversational User Interfaces*. doi:10.1145/3405755.3406123

16) „Die Stimme ohne Beigaben ist keine ‚normale' Stimme mehr; ihr fehlt jene menschliche Note, die der öden Maschinerie des Signifikanten durch die Stimme hinzugefügt zu werden scheint, und so liegt die Drohung in ihr, daß die Menschheit selbst mit der mechanischen Wiederholbarkeit verschmelzen und dadurch ihren Halt verlieren könnte" (Dolar 2014: 34).

// Abbildung 1
***Meet Q*, Screenshot**



Speakern und dem problematischen Design künstlicher Stimmen tritt ein anderer Konflikt beim Erkennen menschlicher Stimmen durch algorithmische Systeme auf. Wie aus der Zeugenaussage einer Wissenschaftlerin vor dem New Yorker City Council zum Einsatz von Algorithmen in automatisierten Personalentscheidungen (Myers West 2020) hervorgeht, macht nicht nur Gesichtserkennungs-Software erhebliche Fehler aufgrund von rassistischen und sexistischen Vorannahmen und Vorurteilen, sondern Programme für die Erkennung von Stimmen diskriminieren Bewerber*innen hinsichtlich Gender und Race. „Facial and voice analysis technologies work less well for people of color, English speakers with non-native accents, and trans people" (Ebd.).

Dass Veränderung in der realen Welt eine Umdeutung der digitalen Welt voraussetzt, um aus regressiven stereotypisierenden Zuschreibungen wie normativen binären Setzungen herauszukommen, manifestiert sich in einigen Texten. Die queer-feministische Autorin Legacy Russell vertritt in ihrem Manifest *Glitch Feminism* genauso die Position, etwas am Maschinencode zu verändern und damit die Welt draußen, „away from the keyboard (‚AFK')," wie sie es nennt, um die eingefahrene Gewichtung von online und IRL neu zu denken. Etwas am Maschinencode zu verändern ist deswegen bedeutsamer als sich hinter Verschlüsselungstechnik und Datenschutz weiter zu verriegeln, weil die Menschen ohnehin schon mitten in den Maschinen stehen und sich nicht entziehen können (Russell 2020: 141). Das Mittendrin sein begründet sie mit den vielen Automatismen, die sich in das digitale Leben eingeschlichen haben: Autofill, Autocomplete, Autoplay. Ihre Idee von „reduce the way our bodies can be read" bedeutet, „to throttle the predictability of autoplay" (Ebd.: 75). *Glitch* ist hier Metapher und Methode, Projektion und Aktion, um die vom Plattformkapitalismus eingeforderte Performance zu stören oder zu verweigern: „glitch is a form of refusal."

Wie nah kann also die maschinelle Stimme der menschlichen kommen und vice versa? Mit welchen Interventionen lassen sich die Gewohnheiten, Zuschreibungen und Markierungen im Sinne von Russells Manifest unterbrechen? Wie könnten die Computer und andere Maschinen sprechen, um nicht den Uncanny Valley-Effekt[17)] zu reproduzieren und mit menschlichen Stimmen verwechselt zu werden? Die künstlich wirkenden Stimmen, die einsame User*innen in *BigBlueButton* begrüßen und der Text, den sie sprechen, so scheint es, verraten sie schnell. Die menschliche Stimme imitierte bestenfalls Modulation, Affektarmut und Einsilbigkeit der ersten Generationen von Computerstimmen.

17) Mit dem Uncanny Valley-Effekt ist das irritierende, ins Unheimliche tendierende Erleben von KI-Technologie, wie es der japanische Robotikforscher Masahiro Mori 1970 anhand von zu menschlich und daher unheimlich wirkenden Robotern und Handprothesen erforscht und in einer Diagrammkurve als „Uncanny Valley" markiert hatte (Mori, Masahiro (2019): Das Unheimliche Tal [Bukimi no Tani Genshō, 1970]. Aus dem Japanischen von MacDorman, Karl F.; Schwind, Valentin. In: Haensch, Konstantin Daniel; Nelke, Lara; Planitzer, Matthias (Hg.), Uncanny Interfaces. Hamburg, Textem Verlag 2019. S. 212–219). Die Affinität zwischen Mensch und Maschine nimmt durch ausgeprägt anthropomorphes Design stark ab. Daher plädierte Mori dafür, „durch nichtmenschliche Designs Maschinen zu entwerfen, deren Gegenwart Affinität im Menschen erzeugt" (Ebd.: 218). AI-Anwendungen begleiten also scheinbar weiterhin dieselben Probleme, wenngleich sie in andere Anwendungsbereiche fallen.

Denn Stimme und Raum bedingen sich gegenseitig. Doch im Digitalen bleiben die Welten fragmentarisch. Was dort fehlt, sind unter anderem „extra-linguistic elements of communication: the soundings, gestures and affective transmissions that make up our different relations“ (Kanngiesser 2011). Die Möglichkeiten in den Code zu intervenieren und gegenwärtige Cyborg-Prozesse künstlicher Stimmen hörbar und verhandelbar zu machen oder zu stören, sind dennoch gegeben, wie die hier vorgestellten Beispiele verdeutlichen. Feminist AI haben es exemplarisch mit Testen alternativer Stimmen demonstriert. *Q* ist ein weiterer Vorschlag, anders über die konventionellen AI-Stimmen und ihre Wirkungen in der realen Welt zu nachzudenken. Smartness sollte diverser ausgelegt werden, andernfalls hängen die digitalen Ichs vieler Nutzer*innen im Uncanny Valley fest. You are not the only person in this conference.

// Literaturverzeichnis

AdHominem (Username), How do I eliminate “You are the only person in this conference”, Februar 2012. https://community.freepbx.org/t/how-do-i-eliminate-you-are-currently-the-only-person-in-this-conference/13348 (22.02.2021)

Barrera, Pablo (Google Research); Stirnberg, Florian (DeepMind) (2020): Improving Audio Quality in Duo with WaveNetEQ, Google AI Blog, 1. April 2020. https://ai.googleblog.com/2020/04/improving-audio-quality-in-duo-with.html (22.02.2021)

Bucher, Taina (2012): Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. In: New Media & Society 14(7), S. 1164–1180.

Descript (2021): Ultra-realistic voice-cloning. https://www.descript.com/overdub?lyrebird=true (22.02.2021)

Dolar, Mladen (2014). His Master’s Voice. Eine Theorie der Stimme. [OR, 2003], Frankfurt/Main, Suhrkamp.

Dyson, Frances (2014), The Tone of Our Times. Sound, Sense, Economy, and Ecology. Cambridge, MA und London, The MIT Press.

Feminist AI (2016): Thoughtful Voice Design. https://www.feminist.ai/thoughtful-voice-design_ (22.02.2021)

Galloway, Alexander R. (2012), The Interface Effect. Cambridge, Polity Press.

Galloway, Alexander R. (2017), Intervening Infrastructures: Ad Hoc Networking and Liberated Computer Language. In: Caygill, Leeker, Schulze (Hg.), Interventions in Digital Cultures. Lüneburg, meson press 2017, S. 61–72.

Haraway, Donna (1984): Lieber Kyborg als Göttin! Für eine sozialistisch-feministische Unterwanderung der Gentechnologie. In: Lange, Bernd-Peter / Stuby, Anna Maria (Hg.), Neunzehnhundertvierundachtzig, Argument-Sonderband 105. Berlin, Argument-Verlag, S. 66–84.

Harwell, Drew (2018): A Google program can pass as a human on the phone. Should it be required to tell people it’s a machine? In: The Washington Post, 9. Mai 2018. https://www.washingtonpost.com/news/the-switch/wp/2018/05/08/a-google-program-can-pass-as-a-human-on-the-phone-should-it-be-required-to-tell-people-its-a-machine/ (22.02.2021)

Kanngieser, Anja (2011): A sonic geography of voice: Towards an affective politics. In: Progress in Human Geography, 10 November 2011, DOI: 10.1177/0309132511423969.

Lieb, Matthias (2020): Bildtelefonie – Vorläufer der Videokonferenz. Museumsstiftung Post und Telekommunikation. 24. März 2020. https://www.lebenx0.de/bildtelefonie-videokonferenzen/?fbclid=IwAR0ZG4wGINZJIhs9kgh0w2UtxUaBm_LEHufmzIJ-BMYkI1giFcvD9FqrVL0 (22.02.2021)

MIT Technology Review (2020): Google’s auto-complete for speech can cover up glitches in video calls. 6. April 2020. https://www.technologyreview.com/2020/04/06/998410/google-artificial-intelligence-autocomplete-internet-voice-speech-glitches-video-call/ (22.02.2021)

Mrkšić, Nikola (2020): Our Voice Assistant Spoke to Google Duplex. Here’s What Happened… 23. September 2020. https://www.polyai.com/

our-voice-assistant-spoke-to-google-duplex-heres-what-happened/ (22.02.2021)
Munn, Luke (2018), Ferocious Logics. Unmaking the Algorithm. Lüneburg, meson press.
Myers West, Sarah (2020), AI Now Institute, Ethical Implications of Using Artificial Intelligence and Automated Decision Systems. New York City Council, Committee on Technology, November 13, 2020. https://ainowinstitute.org/ai-now-city-council-testimony-fair-shot-act.pdf (22.02.2021)
Nunes, Mark (2011), Error. Glitch, Noise, and Jam in New Media Cultures. New York und London, Continuum Books.
Russell, Legacy (2020): *Glitch Feminism. A Manifesto.* London, New York, Verso.

// Abbildungsverzeichnis

Abb. 1: Screenshot „Meet Q the genderless Voice“

// Angaben zur Autorin

Vera Tollmann ist Kulturwissenschaftlerin und Dozentin für Theorie der digitalen Medien an der Universität Hildesheim. Im September 2020 hat sie an der HFBK Hamburg promoviert. Zu ihren jüngsten Veröffentlichungen gehören „Poor Connections. A long history of videotelephony“ (Cabinet Magazine, 2020), „Wow, that's so postcard!“ in *Planet Earth* (Humboldt Books, 2019) und „Proxies“ (mit Wendy Hui Kyong Chun und Boaz Levin) in *Uncertain Archives* (MIT Press 2021). Sie ist Ko-Kuratorin der Ausstellung „Sensing Scale,“ die im Frühjahr 2021 in der Kunsthalle Münster zu sehen sein wird.

// FKW wird gefördert durch das Mariann Steegmann Institut und Cultural Critique / Kulturanalyse in den Künsten ZHdK

Sigrid Adorf / Kerstin Brandes / Edith Futscher / Kathrin Heinz / Anja Herrmann / Marietta Kesting / Marianne Koos / Mona Schieren / Kea Wienand / Anja Zimmermann // www.fkw-journal.de

// Lizenz