(https://www.theatlantic.com/technology/archive/2024/05/elevenlabs-ai-voice-cloning-deepfakes/678288/, retrieved 5/4/2024)

My Journey Inside the Voice-Clone Factory

ElevenLabs has made some of the internet's most convincing AI voices. Are its creators ready for the chaos that is coming?

Charlie Warzel,[1] May 4, 2024, 7 AM ET

My voice was ready. I'd been waiting, compulsively checking my inbox. I opened the email and scrolled until I saw a button that said, plainly, "Use voice." I considered saying something aloud to mark the occasion, but that felt wrong. The computer would now speak for me.

I had thought it'd be fun, and uncanny, to clone my voice. I'd sought out the AI start-up ElevenLabs, paid $22 for a "creator" account, and uploaded some recordings of myself. A few hours later, I typed some words into a text box, hit "Enter," and there I was: all the nasal lilts, hesitations, pauses, and mid-Atlantic-by-way-of-Ohio vowels that make my voice mine.

It was me, only more pompous. My voice clone speaks with the cadence of a pundit, no matter the subject. I type *I like to eat pickles*, and the voice spits it out as if I'm on *Meet the Press*. That's not my voice's fault; it is trained on just a few hours of me speaking into a microphone for various podcast appearances. The model likes to insert *um*s and *ah*s: In the recordings I gave it, I'm thinking through answers in real time and choosing my words carefully. It's uncanny, yes, but also quite convincing—a part of my essence that's been stripped, decoded, and reassembled by a little algorithmic model so as to no longer need my pesky brain and body.

Using ElevenLabs, you can clone your voice like I did, or type in some words and hear them spoken by "Freya," "Giovanni," "Domi," or hundreds of other fake voices, each with a different accent or intonation. Or you can dub a clip into any one of 29 languages while preserving the speaker's voice. In each case, the technology is unnervingly good. The voice bots don't just sound far more human than voice assistants such as Siri; they also sound better[2] than any other widely available AI audio software right now. What's different about the best ElevenLabs voices, trained on far more audio than what I fed into the machine, isn't so much the quality of the voice but the way the software uses context clues to modulate delivery. If you feed it a news report, it speaks in a serious, declarative tone. Paste in a few paragraphs of *Hamlet*, and an ElevenLabs voice reads it with a dramatic storybook flare.

ElevenLabs launched an early version of its product a little over a year ago, but you might have listened to one of its voices without even knowing it. Nike used the software to create a clone of the NBA star Luka Dončić's voice[3] for a recent shoe campaign. New York City Mayor Eric Adams's office cloned the politician's voice so that it could deliver[4] robocall messages in Spanish, Yiddish, Mandarin, Cantonese, and Haitian Creole. The technology has been used to re-create the voices[5] of children killed in the Parkland school shooting, to lobby for gun reform. An ElevenLabs voice might

[1] Charlie Warzel is a staff writer at The Atlantic and the author of its newsletter Galaxy Brain, about technology, media, and big ideas.

[2] https://www.theverge.com/23864878/ai-voice-clones-podcastle-elevenlabs-personal-voice

[3] https://elevenlabs.io/customers/luka-doncic

[4] https://ny1.com/nyc/all-boroughs/politics/2023/10/17/mayor-eric-adams-mandarin-ai-robocalls-different-languages-yiddish

[5] https://www.wsj.com/tech/ai-brings-back-voices-of-children-killed-in-shootings-7d72cb8d

be reading this article to you: *The Atlantic* uses the software to auto-generate audio versions of some stories, as does *The Washington Post*.

It's easy, when you play around with the ElevenLabs software, to envision a world in which you can listen to all the text on the internet in voices as rich as those in any audiobook. But it's just as easy to imagine the potential carnage: scammers targeting parents by using their children's voice to ask for money, a nefarious October surprise from a dirty political trickster. I tested the tool to see how convincingly it could replicate my voice saying outrageous things. Soon, I had high-quality audio of my voice clone urging people not to vote, blaming "the globalists" for COVID, and confessing to all kinds of journalistic malpractice. It was enough to make me check with my bank to make sure any potential voice-authentication features were disabled.

I went to visit the ElevenLabs office and meet the people responsible for bringing this technology into the world. I wanted to better understand the AI revolution as it's currently unfolding. But the more time I spent—with the company and the product—the less I found myself in the present. Perhaps more than any other AI company, ElevenLabs offers a window into the near future of this disruptive technology. The threat of deepfakes is real, but what ElevenLabs heralds may be far weirder. And nobody, not even its creators, seems ready for it.

In mid-November, I buzzed into a brick building on a London side street and walked up to the second floor. The corporate headquarters of ElevenLabs—a $1 billion company—is a single room with a few tables. No ping-pong or beanbag chairs—just a sad mini fridge and the din of dutiful typing from seven employees packed shoulder to shoulder. Mati Staniszewski, ElevenLabs' 29-year-old CEO, got up from his seat in the corner to greet me. He beckoned for me to follow him back down the stairs to a windowless conference room ElevenLabs shares with a company that, I presume, is not worth $1 billion.

Staniszewski is tall, with a well-coiffed head of blond hair, and he speaks quickly in a Polish accent. Talking with him sometimes feels like trying to engage in conversation with an earnest chatbot trained on press releases. I started our conversation with a few broad questions: *What is it like to work on AI during this moment of breathless hype, investor interest, and genuine technological progress? What's it like to come in each day and try to manipulate such nascent technology?* He said that it's exciting.

We moved on to what Staniszewski called his "investor story." He and the company's co-founder, Piotr Dabkowski, grew up together in Poland watching foreign movies that were all clumsily dubbed into a flat Polish voice. Man, woman, child—whoever was speaking, all of the dialogue was voiced in the same droning, affectless tone by male actors known as[6] *lektors*.

They both left Poland for university in the U.K. and then settled into tech jobs (Staniszewski at Palantir and Dabkowski at Google). Then, in 2021, Dabkowski was watching a film with his girlfriend and realized that Polish films were *still* dubbed in the same monotone *lektor* style. He and Staniszewski did some research and discovered that markets outside Poland were also relying on *lektor*-esque dubbing.

The next year, they founded ElevenLabs. AI voices were everywhere—think Alexa, or a car's GPS—but actually *good* AI voices, they thought, would finally put an end to *lektors*. The tech giants have hundreds or thousands of employees working on AI, yet ElevenLabs, with a research team of just seven people, built a voice tool that's arguably better than anything its competitors have released. The company poached researchers from top AI companies, yes, but it also hired a college dropout who'd won coding competitions, and another "who worked in call centers while exploring audio research as a side gig," Staniszewski told me. "The audio space is still in its breakthrough stage,"

[6] https://www.wsj.com/articles/SB119215016517556740

Alex Holt, the company's vice president of engineering, told me. "Having more people doesn't necessarily help. You need those few people that are incredible."

ElevenLabs knew its model was special when it started spitting out audio that accurately represented the relationships between words, Staniszewski told me—pronunciation that changed based on the context (*minute*, the unit of time, instead of *minute*, the description of size) and emotion (an exclamatory phrase spoken with excitement or anger).

Much of what the model produces is unexpected—sometimes delightfully so. Early on, ElevenLabs' model began randomly inserting applause breaks after pauses in its speech: It had been training on audio clips from people giving presentations in front of live audiences. Quickly, the model began to improve, becoming capable of *ums* and *ahs*. "We started seeing some of those human elements being replicated," Staniszewski said. The big leap was when the model began to laugh like a person. (My voice clone, I should note, struggles to laugh, offering a machine-gun burst of "haha"s that sound jarringly inhuman.)

Compared with OpenAI and other major companies, which are trying to wrap their large language models around the entire world and ultimately build an artificial human intelligence, ElevenLabs has ambitions that are easier to grasp: a future in which ALS patients can still communicate in their voice after they lose their speech. Audiobooks that are ginned up in seconds by self-published authors, video games in which every character is capable of carrying on a dynamic conversation, movies and videos instantly dubbed into any language. A sort of Spotify of voices, where anyone can license clones of their voice for others to use—to the dismay of professional voice actors. The gig-ification of our vocal cords.

What Staniszewski also described when talking about ElevenLabs is a company that wants to eliminate language barriers entirely. The dubbing tool, he argued, is its first step toward that goal. A user can upload a video, and the model will translate the speaker's voice into a different language. When we spoke, Staniszewski twice referred to the Babel fish from the science-fiction book *The Hitchhiker's Guide to the Galaxy*—he described making a tool that immediately translates every sound around a person into a language they can understand.

Every ElevenLabs employee I spoke with perked up at the mention of this moonshot idea. Although ElevenLabs' current product might be exciting, the people building it view current dubbing and voice cloning as a prelude to something much bigger. I struggled to separate the scope of Staniszewski's ambition from the modesty of our surroundings: a shared conference room one floor beneath the company's sparse office space. ElevenLabs may not achieve its lofty goals, but I was still left unmoored by the reality that such a small collection of people could build something so genuinely powerful and release it into the world, where the rest of us have to make sense of it.

ElevenLabs' voice bots launched in beta in late January 2023. It took very little time for people to start abusing them. Trolls on 4chan used the tool to make deepfakes of celebrities saying awful things. They had Emma Watson reading[7] *Mein Kampf* and the right-wing podcaster Ben Shapiro making racist comments about Representative Alexandria Ocasio-Cortez. In the tool's first days, there appeared to be virtually no guardrails. "Crazy weekend," the company tweeted,[8] promising to crack down on misuse.

ElevenLabs added a verification process for cloning; when I uploaded recordings of my voice, I had to complete multiple voice CAPTCHAs, speaking phrases into my computer in a short window of time to confirm that the voice I was duplicating was my own. The company also decided to limit its

[7] https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs?utm_source=reddit.com

[8] https://x.com/elevenlabsio/status/1620032627075063811?s=20

voice cloning strictly to paid accounts and announced a tool that lets people upload audio to see if it is AI generated. But the safeguards from ElevenLabs were “half-assed,” Hany Farid, a deepfake expert at UC Berkeley, told me—an attempt to retroactively focus on safety only after the harm was done. And they left glaring holes. Over the past year, the deepfakes have not been rampant, but they also haven’t stopped.

I first started reporting on deepfakes in 2017, after a researcher came to me with a warning[9] of a terrifying future where AI-generated audio and video would bring about an “infocalypse” of impersonation, spam, nonconsensual sexual imagery, and political chaos, where we would all fall into what he called “reality apathy.” Voice cloning already existed, but it was crude: I used an AI voice tool to try to fool[10] my mom, and it worked only because I had the halting, robotic voice pretend I was losing cell service. Since then, fears of an infocalypse have lagged behind the technology’s ability to distort reality. But ElevenLabs has closed the gap.

The best deepfake I’ve seen was from the filmmaker Kenneth Lurt, who used ElevenLabs to clone Jill Biden’s voice for a fake advertisement[11] where she’s made to look as if she’s criticizing her husband over his handling of the Israel-Gaza conflict. The footage, which deftly stitches video of the first lady giving a speech with an ElevenLabs voice-over, is incredibly convincing and has been viewed hundreds of thousands of times. The ElevenLabs technology on its own isn’t perfect. “It’s the creative filmmaking that actually makes it feel believable,” Lurt said[12] in an interview in October, noting that it took him a week to make the clip.

“It will totally change how everyone interacts with the internet, and what is possible,” Nathan Lambert, a researcher at the Allen Institute for AI, told me in January. “It’s super easy to see how this will be used for nefarious purposes.” When I asked him if he was worried about the 2024 elections, he offered a warning: “People aren’t ready for how good this stuff is and what it could mean.” When I pressed him for hypothetical scenarios, he demurred, not wanting to give anyone ideas.

A few days after Lambert and I spoke, his intuitions became reality. The Sunday before the New Hampshire presidential primary, a deepfaked, AI-generated robocall went out to registered Democrats in the state. “What a bunch of malarkey,” the robocall began.[13] The voice was grainy, its cadence stilted, but it was still immediately recognizable as Joe Biden’s drawl. “Voting this Tuesday only enables the Republicans in their quest to elect Donald Trump again,” it said, telling voters to stay home. In terms of political sabotage, this particular deepfake was relatively low stakes, with limited potential to disrupt electoral outcomes (Biden still won in a landslide). But it was a trial run for an election season that could be flooded with reality-blurring synthetic information.

Researchers and government officials scrambled to locate the origin of the call. Weeks later, a New Orleans–based magician confessed[14] that he’d been paid by a Democratic operative to create the robocall. Using ElevenLabs, he claimed, it took him less than 20 minutes and cost $1.

[9] https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news

[10] https://www.buzzfeednews.com/article/charliewarzel/i-used-ai-to-clone-my-voice-and-trick-my-mom-into-thinking

[11] https://twitter.com/MericasFunniest/status/1715277968031822082?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1715277968031822082%7Ctwgr%5E5dd35a849986bd2bfabac0bf4dc052e8ac45fd0d%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fventurebeat.com%2Fai%2Fconfessions-of-an-ai-deepfake-propagandist-using-elevenlabs-to-clone-jill-bidens-voice%2F

[12] https://venturebeat.com/ai/confessions-of-an-ai-deepfake-propagandist-using-elevenlabs-to-clone-jill-bidens-voice/

[13] https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984

[14] https://www.nbcnews.com/politics/2024-election/biden-robocall-new-hampshire-strategist-rcna139760

Afterward, ElevenLabs introduced a “no go”–voices policy, preventing users from uploading or cloning the voice of certain celebrities and politicians. But this safeguard, too, had holes. In March, a reporter[15] for *404 Media* managed to bypass the system and clone both Donald Trump’s and Joe Biden’s voices simply by adding a minute of silence to the beginning of the upload file. Last month, I tried to clone Biden’s voice, with varying results. ElevenLabs didn’t catch my first attempt, for which I uploaded low-quality sound files from YouTube videos of the president speaking. But the cloned voice sounded nothing like the president’s—more like a hoarse teenager’s. On my second attempt, ElevenLabs blocked the upload, suggesting that I was about to violate the company’s terms of service.

For Farid, the UC Berkeley researcher, ElevenLabs’ inability to control how people might abuse its technology is proof that voice cloning causes more harm than good. “They were reckless in the way they deployed the technology,” Farid said, “and I think they could have done it much safer, but I think it would have been less effective for them.”

The core problem of ElevenLabs—and the generative-AI revolution writ large—is that there is no way for this technology to exist and not be misused. Meta and OpenAI have built synthetic voice tools, too, but have so far declined[16] to[17] make them broadly available. Their rationale: They aren’t yet sure how to unleash their products responsibly. As a start-up, though, ElevenLabs doesn’t have the luxury of time. “The time that we have to get ahead of the big players is short,” Staniszewski said. “If we don’t do it in the next two to three years, it’s going to be very hard to compete.” Despite the new safeguards, ElevenLabs’ name is probably going to show up in the news again as the election season wears on. There are simply too many motivated people constantly searching for ways to use these tools in strange, unexpected, even dangerous ways.

In the basement of a Sri Lankan restaurant on a soggy afternoon in London, I pressed Staniszewski about what I’d been obliquely referring to as “the bad stuff.” He didn’t avert his gaze as I rattled off the ways ElevenLabs’ technology could be and has been abused. When it was his time to speak, he did so thoughtfully, not dismissively; he appears to understand the risks of his products. “It’s going to be a cat-and-mouse game,” he said. “We need to be quick.”

Later, over email, he cited the “no go”–voices initiative and told me that ElevenLabs is “testing new ways to counteract the creation of political content,” adding more human moderation and upgrading its detection software. The most important thing ElevenLabs is working on, Staniszewski said—what he called “the true solution”—is digitally watermarking synthetic voices at the point of creation so civilians can identify them. That will require cooperation across dozens of companies: ElevenLabs recently signed an accord[18] with other AI companies, including Anthropic and OpenAI, to combat deepfakes in the upcoming elections, but so far, the partnership is mostly theoretical.

The uncomfortable reality is that there aren’t a lot of options to ensure bad actors don’t hijack these tools. “We need to brace the general public that the technology for this exists,” Staniszewski said. He’s right, yet my stomach sinks when I hear him say it. Mentioning media literacy, at a time when trolls on Telegram channels can flood social media with deepfakes, is a bit like showing up to an armed conflict in 2024 with only a musket.

[15] https://www.404media.co/elevenlabs-block-on-cloning-bidens-voice-easily-bypassed/

[16] https://www.google.com/url?q=https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices&sa=D&source=docs&ust=1713460259396578&usg=AOvVaw1BzgcOpdaVGM0kApLpXvPV

[17] https://ai.meta.com/blog/audiobox-generating-audio-voice-natural-language-prompts/

[18] https://www.aielectionsaccord.com/

The conversation went on like this for a half hour, followed by another session a few weeks later over the phone. A hard question, a genuine answer, my own palpable feeling of dissatisfaction. I can't look at ElevenLabs and see beyond the risk: *How can you build toward this future?* Staniszewski seems unable to see beyond the opportunities: *How* can't *you build toward this future?* I left our conversations with a distinct sense that the people behind ElevenLabs don't want to watch the world burn. The question is whether, in an industry where everyone is racing to build AI tools with similar potential for harm, intentions matter at all.

To focus only on deepfakes elides how ElevenLabs and synthetic audio might reshape the internet in unpredictable ways. A few weeks before my visit, ElevenLabs held a hackathon, where programmers fused the company's tech with hardware and other generative-AI tools. Staniszewski said that one team took an image-recognition AI model and connected it to both an Android device with a camera and ElevenLabs' text-to-speech model. The result was a camera that could narrate what it was looking at. "If you're a tourist, if you're a blind person and want to see the world, you just find a camera," Staniszewski said. "They deployed that in a weekend."

Repeatedly during my visit, ElevenLabs employees described these types of hybrid projects—enough that I began to see them as a helpful way to imagine the next few years of technology. Products that all hook into one another herald a future that's a lot less recognizable. More machines talking to machines; an internet that writes itself;[19] an exhausting, boundless comingling of human art and human speech with AI art and AI speech until, perhaps, the provenance ceases to matter.

I came to London to try to wrap my mind around the AI revolution. By staring at one piece of it, I thought, I would get at least a sliver of certainty about what we're barreling toward. Turns out, you can travel across the world, meet the people building the future, find them to be kind and introspective, ask them all of your questions, and still experience a profound sense of disorientation about this new technological frontier. Disorientation. That's the main sense of this era—that something is looming just over the horizon, but you can't see it. You can only feel the pit in your stomach. People build because they can. The rest of us are forced to adapt.

[19] https://nymag.com/intelligencer/2024/01/new-ai-powered-google-chrome-browser-end-of-human-internet.html