The Acousmatic Question and the Will to Datafy

Otter.ai, Low-Resource Languages, and the Politics of Machine Listening

Jonathan Sterne and Mehak Sawhney

A speaker declaims a lecture for an audience. A group of engineers meet in a conference room to plan the next software release. People conduct transactions through voice authentication. Teachers and students convene online in a pandemic. Phone calls are made to loved ones. Friends send each other voice messages to catch up. Someone complains about the new racist laws being enacted by the local legislature. Activists plan a protest. Journalists talk to sources for their articles. People carry smartphones with always-on voice assistants.

In many parts of the world, people conduct their lives in environments where microphones and speakers outnumber screens. They may make phone calls or send voicemails, but they may also simply be talking in an environment where a microphone is connected to a listening apparatus. In many cases today, that apparatus transforms their speech into data, which is then collected into a vast data lake, whether for a media corporation, a government, or both.[1]

Jonathan Sterne (sterneworks.org) teaches in the Department of Art History and Communication Studies at McGill University. He is the author of *Diminished Faculties: A Political Phenomenology of Impairment* (Durham, NC: Duke University Press, 2021); *MP3: The Meaning of a Format* (Durham, NC: Duke University Press, 2012); *The Audible Past: Cultural Origins of Sound Reproduction* (Durham, NC: Duke University Press, 2003); and numerous articles on media, technologies, and the politics of culture. He is also the editor of *The Sound Studies Reader* (New York: Routledge, 2012) and co-editor of *The Participatory Condition in the Digital Age* (Minneapolis: University of Minnesota Press, 2016). With co-author Mara Mills, he is working on *Tuning Time: Histories of Sound and Speed*, and he has a new project cooking on artificial intelligence and culture, of which this article is a part.

Mehak Sawhney is pursuing her doctorate in communication studies at McGill University. Her doctoral project explores the technological and political aspects of audio surveillance in India. Her research interests lie at the intersection of sound and media cultures of South Asia, with specific focus on sonic surveillance, machine listening, and urban sound. She has previously worked as a researcher at Sarai, the new media program at the Centre for the Study of Developing Societies, Delhi. She holds master of arts and master of philosophy degrees in English literature from the University of Delhi.

Kalfou, Volume 9, Issue 2 (Fall 2022).
ISSN 2151-4712 (print). ISSN 2372-0751 (online).

"Who Is This?"

Nina Sun Eidsheim's *The Race of Sound* outlines the fraught human politics of the acousmatic question—"Who is this?"—which aims to identify and taxonomize the speaker behind a voice. This desire to identify another through their voice can never be fully realized. In scholarship and in everyday talk in the global North, vocality is still far too often understood as a simple proxy for spirit, soul, breath, or essence; but the sound of a voice alone will never be enough to determine another's identity: "We assume that when we ask the acousmatic question we will learn something about an individual. We assume that when we ask the acousmatic question we inquire about the essential nature of a person."[2] While ultimately arguing against the reduction of voice to timbre,[3] and therefore against the likes of the data-based reasoning outlined in this article, Eidsheim also uses data to challenge prevailing conceptions of essential vocal gender (in the case of Jimmy Scott) and vocal race (in the case of Vocaloid).[4]

But what happens when it is not just human beings asking the acousmatic question? What happens when the question is delegated to machinery? Eidsheim offers a preliminary answer in her chapter on Vocaloid: "While the reader may still be somewhat hesitant about accepting my argument that race, as thought to be heard in vocal timbre, has no essential origin, this chapter shows that even when assembling zeros and ones, listeners continue to produce and reify notions of racialized vocal timbre."[5] While *The Race of Sound* explores how listeners attribute race to voice in a musical context, vocal identification occurs in many other situations. In this article, we focus on the acousmatic question as it is applied to human speech in machine listening. Drawing on the recent boom in artificial intelligence (AI) and more specifically machine learning (ML), a growing machine-listening industry has arisen that is predicated on mass surveillance and the expropriation of data from people without their knowledge or full consent. As some examples in the latter half of this article demonstrate, consent loses its meaning because the existing power relations do not realistically offer people the choice to opt out of the extractive processes of data collection. It may well be the case that most data extraction happens in situations where it is exceptionally difficult for people to *not* consent.

When the acousmatic question is transferred from humans to machines, processes of voice identification and recognition take an entirely different form. We use the phrase *machine listening* to refer to a broad set of automated actions that computers perform on audio signals, whether it be music, voice, or ambient sound.[6] When applied to the human voice, machine listening can be undertaken with the goals of speech recognition, voice identification, emotion analysis, lie detection, or even medical diagnosis. While machine listening is often *unsuccessful* in its desired applications—there is no evidence that it is effective at detecting affect or lies—it is often deployed as if it were.

For the purposes of recognizing speech and identifying speakers, machine listening requires large amounts of voice data. Speech corpora—databases of audio files and their corresponding text transcriptions—are used for developing voice recognition and identification systems in various languages. In the last few years there has been a discursive shift in the AI industry toward making AI more ethical, inclusive, and accessible. Major corporations like Facebook and Google now have "ethics" boards, and "AI Ethics" institutes have sprouted up at major universities across the planet. Often, calls for more ethical AI from these quarters involve the expansion of the datasets on which machine listening depends: more machine-recognizable languages and accents, more people, more reach. Like technological ethicists before them, AI ethicists rarely say "no."[7]

By transporting the acousmatic question to the context of machine listening, this article aims to critique the emerging discourse of ethical and inclusive AI in both the global South and global North. While the question of accessibility is becoming central to building better AI tools, we argue that the promise of inclusion masks the prevailing logic of datalogical extraction, or the *will to datafy*, that fuels AI by promoting its inevitability in contemporary times. It is not accessible technologies but the extractive logic of AI that we question. The first section of the article explores the acousmatic question in machine listening in greater detail. The second section delves into a critique of ethical AI and offers a reflection on its techno-colonial *will to datafy*. The third and fourth sections investigate specific examples of speech data collection—"low-resource" languages in the global South and Otter.ai in the global North—to expose the entanglement of accessibility and data extraction in both contexts.

The Acousmatic Question in Machine Listening

Today, vocal identification is part and parcel of a big data project that crosses government, military, and industrial operations. While its work is unfinished, it is part of a long history of classifying speakers by race, gender, ethnicity, age, accent, and ability, a project undertaken by people who believe it is possible to definitively answer the acousmatic question. Eidsheim offers three correctives regarding "what we identify when we identify voice: Voice is not singular, it is collective. Voice is not innate, it is cultural. Voice's source is not the singer, it is the listener."[8] With the latter point, Eidsheim displaces the process of vocal evaluation from the speaker to the listener. Her definition of voice is that "it does not exist a priori"[9] and that it is the processes of auto-listening and listening more broadly that assign value to voice. But how do machines make sense of what Eidsheim calls a complex and "thick vocal event"?[10]

In order to identify voices or transcribe speech, machines quantify, classify, and predict human speech.[11] The technical processes involved in speech

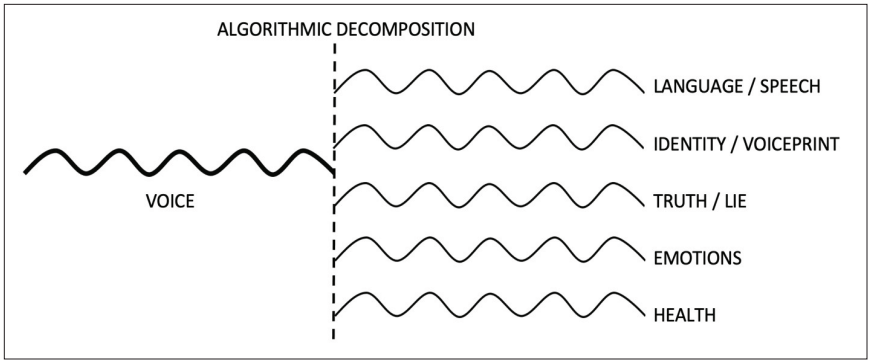


Figure 1. Algorithmic decomposition of human voice. (Image created by Mehak Sawhney.)

recognition can broadly be divided into two parts: signal processing and data analysis. While digital signal processing converts a human voice into a mathematical, analyzable form as a series of zeroes and ones, data analysis trains the machine to recognize, identify, or evaluate speech by classifying human voice into patterns or averages. In the case of speech recognition, it also predicts what speakers might say next based on what they have said already. Both these processes require large amounts of recorded voice data. Classification operates not only at the technical but also at the social level as machines, corporations, and states indulge in *over*hearing to profile and identify the person behind the voice.

Machine listening entails the perception of *voice as a composite of information*. It frames the human voice as a repertoire of information that can be algorithmically decomposed to obtain multiple kinds of data. Figure 1 depicts a voice signal passing through an algorithmic process that can decode it to make guesses about speech and language, the identification of a person, a person's emotional state, the truthfulness of their speech, or even the status of their health. Emerging from interviews with computer scientists in India, the image is an interpretation of how human voice can be analyzed or decomposed to obtain information about the aforementioned parameters. The informational and statistical treatment of human speech reduces listening to a mathematical average, making speech recognition a process that thins the thick vocal event. One of the most basic, and most flawed, suppositions of this model is that the voice *must* contain information. But this is not necessarily always the case. Further, it is highly unlikely that machine listening will be successful at guessing whether a speaker is lying or telling the truth, healthy or ill, or what they are feeling. These kinds of questions represent a twenty-first-century variety of phrenology.[12]

Within the domain of sound studies, machine listening is also subject to critiques of listening that are emerging as the field has begun to intersect with other areas such as Indigenous and disability studies. For instance, in his writ-

ing on Indigenous sound studies, Dylan Robinson understands listening as multisensory while also arguing against any fixed definition of it: "Decolonizing musical practice involves becoming no longer sure what LISTENING is."[13] In their work on deafness in Bangalore and Bangkok, Michele Friedner and Benjamin Taussig argue not only that listening is variable but also that sensory capacities are not biologically determined; rather, they "emerge within social, political, and economic contingencies."[14]

The conception of voice as a repertoire of information also implies the necessity of classifying and identifying it. What happens when a human voice is subjected to machine analysis and assessed in such a way that it *must* possess a particular emotion, language, race, or gender, or a certain degree of truthfulness? We come full circle to Eidsheim's acousmatic question—who is this?—and the idea that it is listening that assigns value to voice. The output of voice analysis is always a probabilistic guess, a statistical approximation of the emotion or identity the voice supposedly possesses; it is never 100 percent certain. When it comes up, this lack of certainty is often framed as a problem of an algorithm that needs more tuning, or a result of incomplete data or "bias." But the problem is not statistical ambiguity or the search for a nonexistent nonbiased dataset. The problem is the corporate and state will to datafy and classify human voice in the first place.

Despite what some of the more optimistic commentary might suggest, it is not possible to entirely delegate listening to computer systems. Algorithms may seek to identify speakers automatically, but the work of voice identification is ultimately for deeply human purposes: sorting, classification, prediction. Eidsheim's claim that the voice alone is unable "to be unique and yield precise answers" is doubly important when trying to understand what machine listening operations can and cannot do. While we enthusiastically embrace Eidsheim's call for a "performative approach" to denaturalize the acousmatic question,[15] corporations, states, and a large swath of the machine learning subfield of academic computer science do not. Researchers and institutions invested in machine listening claim that it is possible to answer the acousmatic question using machine listening techniques, at least probabilistically. In so doing, they claim to be able to fix the identities of speakers, track them, connect their vocalizations with other data, and thereby build a bigger, more sophisticated profile of the whole person, including the person's emotional states, desires, and intentions. The critical literature on AI describes this datafied self as a "data double," a "data body," or a "digital exhaust."[16] While each term means something slightly different, they all refer to the techniques that data science uses to construct a profile of a person across multiple data points and datasets, with the goal of tracking and predicting their behavior. In contexts of heavy policing and sanctioned state surveillance, medicalized surveillance of disabled bodies, or corporate surveillance of employees or users, the data body crashes into the

flesh-and-blood subject, often with violent implications for the people who have been datafied.

While the technology in use today is digital, the impulses behind it are not. As Jennifer Stoever has shown in *The Sonic Color Line*, U.S. sound culture is built on practices of sorting voices into different racialized containers for the purposes of reinforcing white supremacy. "The sonic color line posits racialized subject positions like 'white,' 'black,' and 'brown' as historical accretions of sonic phenomena and aural stereotypes that can function without their correlating visual signifiers and often stand in for them," she asserts; the *listening ear* is the "ideological filter" shaped in relation to the sonic color line.[17] Though racialized peoples have long challenged the classifications attributed to them, or even used them to build vital aesthetic and political traditions—as Fred Moten and Daphne Brooks have shown—the power dynamic remains.[18] In machine listening, that ideological filter can operate through data processing protocols, masking its working through the opacity of algorithms and the secrecy of corporations and governments. As Ruha Benjamin explains, the universalizing language to describe code and data processing can hide the discriminatory aspects of design, mystifying complex and contingent assemblies of institutions, practices, and technologies as "just how the system works."[19]

Further, while it has become much easier to acquire and process voice data, states, corporations, and individuals have sought access to voices for analysis for centuries. The *eave* in *eavesdropping* is a border or threshold, and the purpose of eavesdropping is to transgress or overcome it; as Brian Hochman writes, "Civil War generals traveled with professional telegraph tappers in the 1860s, law enforcement agencies began planting telephone taps in the 1890s, and corporate communications giants tacitly sanctioned state and federal eavesdropping programs of various sorts for most of the twentieth century." Karin Bijsterveld's ongoing research on the Stasi's listening practices in the former German Democratic Republic (GDR) shows that long before today's data science, GDR technicians developed their own sonic skills to identify speakers and register political patterns in their talk.[20] The politics and urges behind machine listening are as old as the acousmatic question itself. Therefore, the political problem of machine listening is not foremost, or only, a technological problem. Beware of simple oppositions between authentic human and inauthentic machine listening.

As mentioned earlier, proponents of ethical AI recognize the variability of human speech and push for the diversification of speech corpora through the inclusion of more accents and languages to make speech recognition accessible to a wider audience. The next section reveals that despite the seeming resonance of ethical AI with principles of inclusivity and access that critical scholars have long propounded, the diversification of the dataset, the rhetoric of inclusion, and the machinic recognition of variable speech can also contribute to the very problems they claim to overcome.

AI's Ethical Contradictions: Access versus the Will to Datafy

In the critical literature on AI, discussions of structural racism, sexism, or ableism are most frequently framed as problems of "bias" and "ethics": for instance when facial recognition engines do not respond to Black faces, or when speech recognition engines cannot perceive female voices or varying English accents as efficiently as the white U.S. English accent.[21] Over the past few years, several books and hundreds of articles have critiqued automated decision making for its bias in relation to marginalized peoples.[22] Authors have variously identified the causes of algorithmic discrimination as biased developers, datasets, or design. This often leads to calls for more inclusive datasets. But even the attempts to mitigate oppression, as Benjamin argues, sometimes feed back into reinforcing inequality. As a result, she warns readers against focusing on big data as the site of inclusion.[23] In other words, the most dangerous belief is not simply a biased assumption that must be overcome. Rather, it is the belief that an algorithmic system—which is itself always a social system—could be unbiased.

To this end, Kate Crawford argues that AI is neither artificial nor intelligent. In *Atlas of AI*, she describes AI as an extractive industry that runs on the systematic exploitation of natural resources and human labor, one that reduces complex human and social experiences to a set of classifications. Thus, our task is not simply to make machine learning more inclusive but to question the very discourse around necessitating the use of AI. We should understand that the expansion of machine learning systems is a kind of technological manifest destiny, a digital colonialism, a future that is sold as *necessary* and *logical* but is in fact backed by financial and military force.[24]

Calls for less bias, more inclusion, or greater access to machine listening systems are often built on assumptions that naturalize the particular corporate and state structures of the machine learning enterprise. In fact, these assumptions help reinforce some of the most dangerous and extractivist aspects of the machine listening project, to the point that noble demands for live transcription of Zoom events or better datafication of languages spoken by less literate populations help prop up the extractivist project of AI (Figure 2). We do not dispute the value of voice recognition for specific tasks or access needs people may have. However, granting the potential usefulness of the technology does not require also accepting the current industrial or political structure of AI. By arguing that data must be made better, calls for ethical AI are calls for generating more data. Scholars have compared this propensity to datafy with the colonial gaze, which converts the world into a profitable resource for the gains of a select few.[25] They have also demonstrated the continuities between colonial forms of data collection such as censuses and fingerprint repositories, as well as other modes of record-keeping, and the contemporary expansion of such in-

formation regimes.[26] It is this "will to datafy"—that is, *the techno-colonial and techno-capitalist logic that aims to turn the world into a resource rendered as data*—that we question. Consider the following quote by Kenneth Cukier and Viktor Mayer-Schönberger:

> To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed. . . . The IT Revolution is evident all around us, but the emphasis has mostly been on the T, the technology. It is time to recast our gaze to focus on the I, the information. In order to capture quantifiable information, to datafy, we need to know how to measure and to record what we measure. This requires the right set of *tools*. It also necessitates a *desire* to quantify and to record. Both are prerequisites of datafication.[27]

It is also important to distinguish between the *will to datafy* and the *processes of datafication*. While the former is akin to the logic and "*desire* to quantify and to record," the latter includes the material practices of data genesis, or the *tools* of measuring and recording. Our concern is not the success or failure, ease or difficulty, with which processes of datafication take place. Rather, we interrogate the extractivist ambitions of machine listening, which aim for ubiquitous recognition in the service of ubiquitous computing. The narrative of AI's inevitability, and the assumption that it is the best or necessary solution, supports a *will to datafy*, which supports *wills to classify and identify*. We focus on the *desire* or *logic* that governs extractive practices of datafication, as there are many contexts, especially in the global South, where human life is not yet continuously trackable due to the unavailability of internet service or smartphones. The desire for data extraction might, however, manifest in different forms, such as calls to increase inclusion and access, which is similar to the civilizational and developmental narratives of European colonialism for mining what they perceived as untapped worlds. This unbridled and multifaceted desire for data extraction

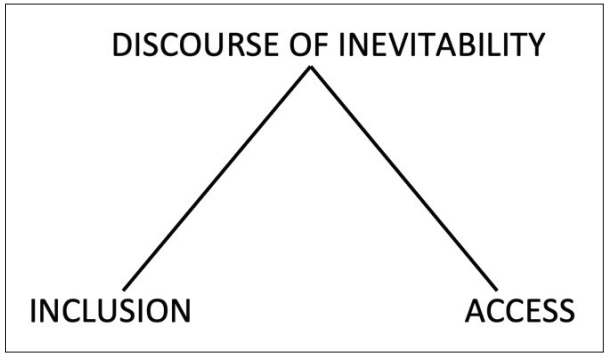


Figure 2. Demands for inclusion and access prop up the discourse of inevitability. (Image created by Jonathan Sterne.)

and classification masked by discourses of inclusion, access, and inevitability of AI is what we call the *will to datafy*.

In the context of both the global South and global North, scholars have argued that the meaning and power of data reside in the various materialities, infrastructures, practices, and relationalities that define it—from the way data gets attached to multiple media forms such as paper and screens to the mediating influence of the many scientists, analysts, brokers, and other intermediaries who collect, clean, analyze, or sell it.[28] Beyond the immediate profitability of data, there is also significant speculation; the AI industry attaches value to the accumulation of data, even if it is not immediately monetizable. The will to datafy rests on certain discursive formations that legitimize incessant datafication, the discourse on inclusive and accessible AI being one such instance. In a nutshell, the will to datafy refers to the techno-capitalist desire of perpetual datafication, where data might be monetizable at a future point in time, while this drive rests on narratives such as those of access and inclusion. In the remaining sections of this article, we compare how the will to datafy plays out in global South and global North contexts, highlighting the ways in which progressive projects of inclusion and access are co-opted into extractive datafication projects.

Data Genesis in a Low-Resource Context: "Underrepresented" Languages in India

In Natural Language Processing, that is, the computational processing of language data such as speech and text, languages that are not well recognized by speech recognition engines or other language technologies such as text-to-speech or optical character recognition are called "low-resource" or "underrepresented" languages. If machine learning depends on data abundance, then the data scarcity of a language prompts the need for greater datafication to make the language machine-recognizable. In the case of data genesis for these so-called underrepresented languages, language and speech are rendered datafiable in the name of access, welfare, and even the preservation of endangered languages. The ethical discourse of making corporate speech technologies in underrepresented languages more accessible replaces the erstwhile Western discourse of mitigating technological backwardness through tech-developmental projects in the global South.

In the context of India, the question of making speech technologies accessible manifests in two goals: to include as many languages as possible, and to facilitate the use of internet and phones through speech due to pervasive nonliteracy in the country. Multiple voice technologies have been developed to these ends, both by corporations and the state, while many others have been in the pipeline for years or failed due to bureaucratic hurdles, lack of resources, or scalability concerns. An initiative by Google called the Next Billion Users,[29] with

its mantra of expanding digital dominance through the "3Vs"—voice, video, and vernacular[30]—merits some elaboration here. In this project, Google aims to bring people from developing countries, especially those in South Asia, Southeast Asia, and Africa, online by making technology accessible for them through the 3Vs. The taglines of the project include "Building for everyone, everywhere"; "People are at the centre of everything we build"; and "Building helpful, inclusive products is a global effort." This exemplifies how the idea of inclusion, as proposed by the ethical AI school, actually rests on extractive processes of datafication in the name of access and welfare.

A few examples of data collection for "low-resource languages," mostly funded by the Indian government or tech giants such as Microsoft, illustrate the problem. A recent paper titled "Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Groups"[31] details the collection of speech in Marathi from three different groups: "1) low-income rural Marathi speakers in Amale, a tribal village in rural Maharashtra, 2) low-income urban Marathi speakers in the slums of Kolhapur, a small city in Maharashtra, and 3) university students in Mumbai, who are the typical target population of crowdsourced data collection."[32] The authors demonstrate that the speech samples collected from low-income groups are of comparable quality to those from high-income groups. Originally, data collection was primarily an urban phenomenon; as it has extended into rural areas, its extractive features have been enacted through the guise of an employment program for lower-income people. But there is also a dimension of soft coercion. The fact of unemployment motivates participation in the study: when you are hungry, you are probably not thinking about, or prioritizing, the implications of giving up the rights to your voice data. People who have more access to resources may be more likely to refuse the work. The comparison between the speech samples of high- and low-income groups as well as urban and rural communities is also problematic in itself, because it reinstalls wealth and regional inequality as the basis for two different categories in a dataset, and then uses that difference as the basis for qualitative comparison. Further, the fruits of the research will never travel back to the research subjects themselves: the scientists conducting the study mention that some of the participants did not have any access to smartphones or the internet.

Linguistic data collection in India for ASR (Automatic Speech Recognition) also involves the digitization of books and handwritten manuscripts in multiple languages because in addition to a speech corpus, a text corpus is also needed for speech recognition. Proponents argue that the digitization of physically dispersed books and archives can fulfill the purpose of datafication for machine learning and can also provide public access to these resources for research and reading. The main aim of these digitization drives, however, is the collection of textual data rather than the promotion of public research or the public ownership of digitized resources.[33] Another motivation for building multilingual

speech datasets, computer scientists argue, is that it leads to the preservation of endangered languages, many of which do not have standard orthographies or systems of record keeping. But the creation of speech datasets does not guarantee that the recorded material will be returned to the community, as the history of ethnographic sound recording amply demonstrates. It is only in recent years that colonial sound archives have begun really wrestling with their own extractive legacies.[34]

The collection of speech and textual data from poor villagers, print repositories, and Indigenous communities—albeit with added advantages of income support, digitized libraries, and linguistic preservation—is similar to the philological enterprises during British colonization, German anthropometry and phonography during World War I, and the American Cold War project of global domination that aimed to know the native-others through their languages.[35] The datafication of low-resource languages can hence be placed within the longer history of datafying the other, which "extracts the vitality of black and brown bodies but also enrolls them as the data-labourers to assemble the global assemblages of datafication."[36] The problem with the collection of speech data by tech giants like Google or Microsoft is therefore twofold: their aim is either to increase their user base in developing contexts such as India, which will provide these corporations with their "next billion users," or to acquire data for a growing data lake that is projected to be useful at some later date. As Halcyon Lawrence's work also demonstrates, Amazon's or Apple's support for indigenized or nonnative versions of English in places like India and Singapore are premised not on inclusionary goals but on commercial motives to target emerging profit centers in the global South.[37] In many cases, then, these technologies do not benefit the specific people upon whose data they are built.

The idea here is not to deny the need for accessible speech technologies in India; rather it is to ask which technologies might best provide access for people who need it. Initiatives such as Mobile Vaani (Mobile Speech) are social enterprises in specific parts of India that aim to build voice technologies using IVRS (Interactive Voice Response Systems), which cater to the access needs of nonliterate and low-income groups. They also do not assume the universal availability of smartphones and internet access.[38] IVRS projects combine automation with telephony, for instance in customer service systems where the keypad is used to navigate a menu; they do not necessarily datafy speech. To access these IVRS-based speech technologies, users simply have to telephone a particular number and then hang up. They receive a call back and can access the information they need using a speech-based navigation menu. Though voice-based technologies are extremely important in a low-literacy context such as India, the extractive collection of voice data and machine analysis is not necessarily useful for these populations. According to its manifesto, Mobile Vaani "sustains itself through service fees from primarily non-profit organizations to use the

platform for different development objectives."[39] While it does partner with corporations for some advertisement revenue and suggests that a hybrid for-profit financial model might be important for sustainability, most of its revenue, aside from government funding and philanthropic contributions, comes from running social messaging campaigns on issues such as early marriage, education, livelihood, maternal health, and governance in partnership with various civil society organizations. Most importantly, Mobile Vaani adheres to its objective of social good and retains a user base of low-income and historically marginalized communities while also steering away from surveillance and extraction masquerading as access.[40]

In contrast to other social media platforms such as Facebook, Mobile Vaani "does not even require users to register on the platform and it does not collect any personal details. Even on content contributed on the platform, groups are free to shape practices of whether or not users should reveal their identity when recording messages on the platform."[41] Its terms of use also prohibit the commercial use of the data available on its website or app.[42] Enterprises such as Mobile Vaani provide a glimpse into what access technologies might look like when technological development emerges from the social realities of the users and speaks to their needs first.

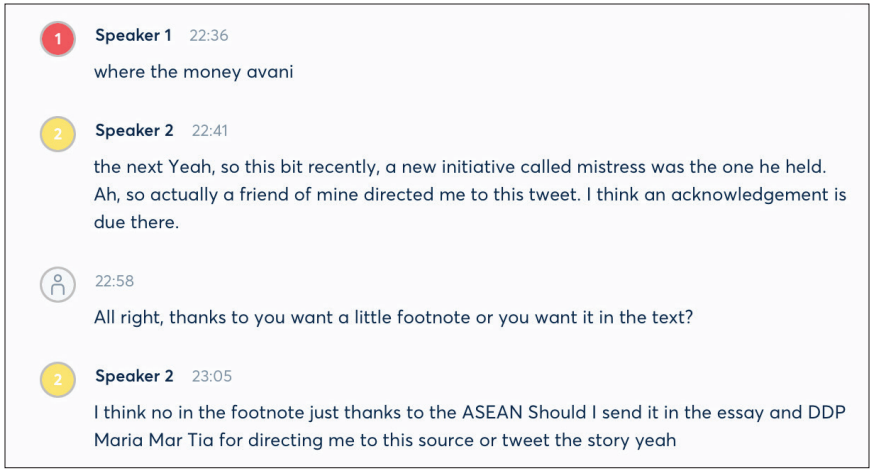
Data Genesis in a High-Resource Context: Otter.ai and Zoom

If the push for access in the global South has focused on datafying low-resource languages, the narrative of inclusion in the global North has made use of the COVID-19 crisis. The pandemic has occasioned one of the great data heists in human history: the mass harvesting of voiceprints. A voiceprint refers to certain measurable qualities of human voice that can purportedly help in uniquely identifying an individual, just like a fingerprint. In fact, voiceprints are not nearly as developed a science as fingerprinting, and within forensic science, there is still some disagreement as to what even constitutes them. The haziness of the science around voiceprinting provides further evidence for Eidsheim's claim that the acousmatic question cannot be fully answered. Nevertheless, the desire to voiceprint is very much part of the will to datafy. Otter.ai, through partnerships with Zoom, has created profiles of millions of speakers, or more accurately, extracted them from Zoom conversations that would otherwise not be retained or copyrighted and claimed ownership over them. Located in Silicon Valley, and still run off venture capital (over US$63 million as of April 2021), Otter.ai's main business is what it calls automated transcription—of meetings, speeches, talk, any audio that can get turned into data.[43] There are two ways this can happen. One option is automatic transcription in real time. At McGill University, we have been using Zoom for our meetings and courses. As of this writing, if the person who starts a meeting enables automatic transcription, then the audio from the

meeting is uploaded to Otter.ai's servers (which are actually provided by Amazon Web Services), and Otter.ai's algorithms perform speech recognition, which not only renders the speech as text but also compares it with nearby speech that has been converted into text, which is why sometimes corrections in transcription are visible as the transcription appears on the screen. Another option is to upload a finished recording to Otter.ai's website.[44] Sometimes Otter.ai will take considerably longer than the duration of the recording to transcribe. This may be because of processing bottlenecks, but it is more likely a result of what Mary Gray and Siddharth Suri call "the paradox of automation's last mile": "The great paradox of automation is that the desire to *eliminate* human labor always *generates* new tasks for humans. . . . '[T]he last mile' is *the gap between what a person can do and what a computer can do*."[45] In Otter.ai's case, it is likely that the company hires digital piece workers, paid at well below minimum wage, to manually listen to recordings and manually correct transcripts. Since Otter.ai has not patented its processes and does not disclose the inner workings of its technology, we cannot be certain, but based on studies of how other AI-based businesses currently run, it is quite likely human beings are involved in some of the transcription work.[46]

As whole sectors of culture and industry moved to online meetings through Zoom, Otter.ai gained access to a massive trove of voices. Otter.ai transcripts routinely show that the company asks the acousmatic question of every bit of speech it processes. Every time the engine labels someone as "Speaker 1" or "Speaker 2," the system is attempting to identify different voices. As is illustrated in Figure 3, the system tries to guess *what* the speakers are saying and *who* is speaking. In this particular error-ridden case, taken from one of our meetings about this article (with Jonathan in Montreal and Mehak in Kolkata), it not only misattributes words but is unsure of how many people are speaking.

Otter.ai's business model is also currently secret, and as they are running off venture capital, they do not yet have to make money. Current business practices may therefore be based on speculation rather than actual profitability. Yet there are clues to how they view the value of voices in their work. According to their terms of service (as of April 2021), customers retain all ownership rights to the "User Content processed using the service," and users retain the right to permanently delete their recordings from Otter.ai at any time. In other words, if Jonathan uploads one of his Zoom lectures to Otter.ai, he retains the authorship rights to what he said. This suggests that, as with other Silicon Valley businesses, Otter.ai is not interested in the "content" of speech, or particular speech acts. Compare this with Otter.ai's explanation of its machine learning operations in the same terms of service: "Nothing in these Terms gives you any rights in or to any part of the Service or the Machine Learning generated by Company or the Machine Learning generated in the course of providing the Service."[47] This implies that the real value is generated from the process of datafying voices. Most distressingly, this language suggests that while Otter.ai clearly develops



Speaker 1 22:36
where the money avani

Speaker 2 22:41
the next Yeah, so this bit recently, a new initiative called mistress was the one he held. Ah, so actually a friend of mine directed me to this tweet. I think an acknowledgement is due there.

[Unattributed] 22:58
All right, thanks to you want a little footnote or you want it in the text?

Speaker 2 23:05
I think no in the footnote just thanks to the ASEAN Should I send it in the essay and DDP Maria Mar Tia for directing me to this source or tweet the story yeah

Figure 3. An Otter.ai transcript that may or may not be automatically generated.

profiles of the different speakers whose speech it analyzes, those speakers have no rights to their own profile. In other words, not only does Otter.ai ask—and partially automate—the acousmatic question; it claims, in the legalistic writing of its user agreement, that it owns the answer and the speaker does not. Imagine not having the rights to even see or know your own fingerprints, and imagine that those fingerprints are not even necessarily a good representation of the ridges on your fingertips. According to their terms of service, Otter.ai wants a voice . . . and nothing more.

Among the problems raised by the mass privatization of voice data is the contradiction between the very real potential harms of a private company collecting millions of voiceprints and the very real need for automatic transcrip-

tion. Real-time transcription of online meetings is not simply a useful service: it is a crucial access technology. For people with disabilities such as hardness-of-hearing and ADHD, real-time transcription is a lifeline in a sometimes difficult and alienating online environment. It is also a crucial access technology for nonnative speakers of the language being spoken on the call. Rightfully, activists and advocates have thus pushed hard for real-time transcription to be a default part of Zoom meetings. While, for instance, hiring someone to do live transcription would be a better option to serve Deaf participants in such meetings (and possibly also a sign language interpreter), many institutions go for the less expensive solution of automatic transcription.[48]

The will to datafy thus manifests as an inevitability and a necessity. Calls for access and inclusion actually propel the project further. Yet none of this is a true necessity: consider that the Zoom interface also has an option for a human transcriber for meetings. It would certainly be possible for institutions to hire or designate people to do this work. But that would involve recognizing it as valuable labor, negotiating the politics of transcription (rather than mystifying them in a technical system), and paying people fairly for their work. How is it that such a scenario seems more difficult and far-fetched in 2021, in the middle of a pandemic that has occasioned mass unemployment, than simply giving millions of people's voiceprints to a private company—and paying for the privilege?

Conclusion

Since machine listening is premised on the conception of voice as a repertoire of information and the necessity of classifying or identifying it, Eidsheim's analysis of the acousmatic question is both useful and necessary for critiquing AI and machine perception more broadly. It can also be applied to other technologies such as face recognition, risk assessment, biometrics, and social listening (social media analytics) to question the necessity of attaching classifiers and identifiers to human bodies, emotions, and expressions. In its current industrial form, machine listening is very closely tied to corporations and states. While the acousmatic question has long been asked by both corporations and states, machine listening operates on the tripartite logic of the *will to datafy, classify, and identify* the speaker. When voice technologies are presented as inclusive and accessible, they may serve some needs for inclusion or access, but only in service of a greater extractivist project. So long as we do not directly address the political consequences of data extractivism, it will remain easy for corporations and states to subordinate the project of making AI more inclusive or accessible to the will to datafy.

Some activists and scholars adopt a more radical stance, arguing for the "abolition" or "refusal" of big data. Yeshimabeit Milner released a pamphlet

called "Abolish Big Data" urging readers to "dismantle the structures that concentrate the power of Big Data into the hands of a few,"[49] and Joanna Radin calls on the big data community to accept Indigenous communities' "refusal" to provide data for technological or scientific research.[50] As we have seen with the collection of speech data in rural India and the transcription of Zoom speech in Canada and the United States, not all communities are equally able to refuse data collection due to their marginal position; and not all circumstances permit the adoption of alternative sources for accessible information. In most cases this inability to escape data extraction is actually built into business plans: by eliminating alternatives and naturalizing extractivist relationships; by obfuscating or mystifying the actual shape of the relationship between business, user, and technology in user agreements; and by defining the scenario in terms of a set of technical operations ("application of machine listening") rather than beginning from the needs of a population—access to information, cultural preservation, real-time captioning.

Even so, our task as critical scholars is to expose the extractive logics that AI rests upon while pointing toward actually existing alternatives, such as the employment of fairly compensated human transcribers instead of a combination of automated transcription and digital piecework, or the use of IVRS instead of datafied speech for exchanging and accessing information. In the machine listening world, the acousmatic question should be replaced with genuinely inclusive questions and truly accessible answers. In the meantime, when assessing the politics of machine listening, we only need to modify the acousmatic question a little and ask: *Who is this for?*

NOTES

Acknowledgments: We would like to thank Sadie Couture, Burç Kostem, Carrie Rentschler, Magnus Schaefer, Andy Stuhl, Ravi Sundaram, Angus Tarnawsky, Ravi Vasudevan, and all our colleagues at Sarai and in the Culture and Technology Discussion and Working Group (CATDAWG) at McGill.

1. A data lake is a large collection of data whose purpose and utility are not yet fully specified. Amassing data lakes is one of the core business functions of companies like Amazon, Google, and Facebook; for instance, the Amazon Web Services side of Amazon is considerably larger than the consumer-facing side of the company.

2. Nina Sun Eidsheim, *The Race of Sound: Listening, Timbre, and Vocality in African American Music* (Durham, NC: Duke University Press, 2019), 2.

3. Timbral discrimination can be understood as the sonic equivalent of discrimination according to skin color or hair texture. See Eidsheim, *The Race of Sound*, 4.

4. Eidsheim, *The Race of Sound*, 6, 91–113, 115–150. Jimmy Scott was an African American singer with Kallmann syndrome, which meant his voice never dropped from going through puberty. Eidsheim uses the story of Scott's career, alongside an analysis of the pitches he and other singers could actually reach, in order to challenge prevailing notions of vocal gender. She also uses the case of Vocaloid, a voice synthesis program, to show how the acousmatic question applies not only to human voices but also to synthesized voices. She demonstrates how listeners aim to achieve consonance between vocal and visual identity markers even for digital voices.

5. Eidsheim, *The Race of Sound*, 116.

6. There is widespread debate within various machine learning communities as to whether computers "listen" at all. Here, we use the term simply to denote the processing of sonic data, and to highlight the ways in which that processing is implicated in social practice. Listening should always be understood as a relational practice, whether or not computers are involved.

7. In this respect, AI ethics follow well-worn patterns developed in fields like engineering ethics, bioethics, and journalism ethics. See Langdon Winner, testimony to the Committee on Science of the U.S. House of Representatives on the Societal Implications of Nanotechnology, April 9, 2003, https://www.govinfo.gov/content/pkg/CHRG-108hhrg86340/html/CHRG-108hhrg86340.htm: "Although the new academic research in this area would be of some value, there is also a tendency for those who conduct research about the ethical dimensions of emerging technology to gravitate toward the more comfortable, even trivial questions involved, avoiding issues that might become a focus of conflict. The professional field of bioethics, for example (which might become, alas, a model for nanoethics) has a great deal to say about many fascinating things, but people in this profession rarely say 'no.'"

8. Eidsheim, *The Race of Sound*, 9.

9. Ibid., 28.

10. Ibid., 9.

11. For a history of vocal identification and speech recognition, see Xiaochang Li and Mara Mills, "Vocal Features: From Voice Identification to Speech Recognition by Machine," *Technology and Culture* 60, no. 2 (2019): S129–S160.

12. Kate Crawford, *Atlas of AI* (New Haven, CT: Yale University Press, 2021), 123–128.

13. Dylan Robinson, *Hungry Listening: Resonant Theory for Indigenous Sound Studies* (Minneapolis: University of Minnesota Press, 2020), 47.

14. Michele Friedner and Benjamin Taussig, "The Spoiled and the Salvaged: Modulations of Auditory Value in Bangalore and Bangkok," in *Remapping Sound Studies*, ed. Gavin Steingo and Jim Sykes (Durham, NC: Duke University Press, 2018), 169.

15. Eidsheim, *The Race of Sound*, 3, 43.

16. Kevin Heggerty and Richard Ericson, "The Surveillant Assemblage," *British Journal of Sociology* 51, no. 4 (2000): 606, 615; Joanna Radin, "Digital Natives: How Medical and Indigenous Histories Matter for Big Data," *Osiris* 32 (2017): 47.

17. Jennifer Stoever, *The Sonic Color Line: Race and the Cultural Politics of Listening* (New York: New York University Press, 2016), 11, 13.

18. Fred Moten, *The Universal Machine* (Durham, NC: Duke University Press, 2018), esp. 65–118; Daphne Brooks, *Liner Notes for the Revolution: The Intellectual Life of Black Feminist Sound* (Cambridge, MA: Harvard University Press, 2021), esp. 65–123.

19. Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data and Society* 3, no. 1 (2016): 1–12.

20. James Parker and Joel Stern, "Eavesdropping," *Eavesdropping: A Reader*, ed. James Parker and Joel Stern (City Gallery Wellington, Melbourne Law School and Liquid Architecture, 2020), 8–41; Brian Hochman, "Eavesdropping in the Age of the Eavesdroppers; or, The Bug in the Martini Olive," *Post45*, February 2016, https://post45.research.yale.edu/2016/02/eavesdropping-in-the-age-of-the-eavesdroppers-or-the-bug-in-the-martini-olive/; Karin Bijsterveld, "Slicing Sound: Sonic Skills and Speaker Identification at the Stasi, 1966–1989," *Isis* 112, no. 2 (2021): 215–241.

21. Alex Najibi, "Racial Discrimination in Face Recognition Technology," *Science in the News*, October 24, 2020, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/; Joan Palmiter Bajorek, "Voice Recognition Still Has Significant Race and Gender Biases," *Harvard Business Review*, May 10, 2019, https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases#.

22. See Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown Publishers, 2016); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Pun-*

ish the Poor (New York: St. Martin's Press, 2018); Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Cambridge: Polity Press, 2019).

23. Benjamin, *Race after Technology*, 126.

24. Crawford, *Atlas of AI*, 7–9; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the Frontier of Power* (New York: Public Affairs, 2020), 221–225. For her critique of attempts to eliminate bias, see 128–136.

25. See Nick Couldry and Ulises A. Mejías, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (Stanford, CA: Stanford University Press, 2019).

26. See Tiziana Terranova and Ravi Sundaram, "Colonial Infrastructures and Technosocial Networks," *E-flux Journal* 123 (December 2021): n.p.

27. Kenneth Cukier and Viktor Mayer-Schönberger, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (New York: Houghton Mifflin Harcourt, 2013), 78, emphasis ours.

28. Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron* (Cambridge, MA: MIT Press, 2013); Sandeep Mertia, "Introduction: Relationalities Abound," in *Lives of Data: Essays on Computational Cultures from India*, ed. Sandeep Mertia (Amsterdam: Institute of Network Cultures, 2020), 9–25; Dylan Mulvin, *Proxies: The Cultural Work of Standing In* (Cambridge, MA: MIT Press, 2021).

29. See the website Next Billion Users, https://nextbillionusers.google, accessed April 27, 2021.

30. *Business Line*, "Video, Voice and Vernacular—3Vs to Triumph Digital: Google's Rajan Anand," January 17, 2019, https://www.thehindubusinessline.com/info-tech/video-voice-and-vernacular-3vs-to-triumph-digital-googles-rajan-anandan/article26015545.ece.

31. Basil Abraham, Danish Goel, Divya Siddharth, et al., "Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers," *Proceedings of the 12th Conference on Language Resources and Evaluation, European Language Resources Association* (May 11–16, 2020): 2819–2826.

32. Ibid., 2819.

33. Pratik Joshi, Christain Barnes, Sebastin Santy, et al., "Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities," *Microsoft Research Bangalore, India, and Stanford University* (December 2019): n.p., https://arxiv.org/pdf/1912.03457.pdf.

34. Pierre Godard, Gilles Adda, Martine Adda-Decker, et al., "A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments," *Language Resources and Evaluation Conference* (May 2018): 3366–3370; Aaron Fox, "Repatriation as Reanimation through Reciprocity," *The Cambridge Handbook of World Music*, ed. Philip Bohlmann (Chicago: University of Chicago Press, 2013), 522–554; Trevor Reed, "Reclaiming Ownership of the Indigenous Voice: The Hopi Music Repatriation Project," *The Oxford Handbook of Musical Repatriation*, ed. Frank Gunderson, Robert C. Lancefield, and Bret Woods (New York: Oxford University Press, 2019), 627–654.

35. Manan Ahmed Asif, "Technologies of Power—From Area Studies to Data Sciences," *Spheres: Journal of Digital Cultures* 5 (November 20, 2019): https://spheres-journal.org/contribution/technologies-of-power-from-area-studies-to-data-sciences/; Anette Hoffman and Phindezwa Mnyaka, "Hearing Voices in the Archive," *Social Dynamics* 41, no. 1 (2015): 140–165.

36. Noopur Raval, "An Agenda for Decolonizing Data Science," *Spheres: Journal for Digital Cultures* 5 (November 20, 2019): https://spheres-journal.org/contribution/an-agenda-for-decolonizing-data-science/.

37. Halcyon Lawrence, "Siri Disciplines," in *Your Computer Is On Fire*, ed. Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip (Cambridge, MA: MIT Press, 2021), 189–190.

38. See Mobile Vaani: A Gram Vaani Initiative, http://mobilevaani.in/vaani/#/1/home, accessed April 29, 2021.

39. Gram Vaani, "The Mobile Vaani Manifesto," https://gramvaani.org/?p=3901, accessed June 10, 2021.

40. Aaditeshwar Seth, "The Elusive Model of Technology, Media, Social Development, and Financial Stability," in *Socio-tech Innovation: Harnessing Technology for Social Good*, ed. Latha Poonamallee, Joanne Scillitoe, and Simy Joy (Cham: Palgrave Macmillan, 2020), 73–102; Aaditeshwar Seth, email message to Mehak Sawhney, June 11, 2021.

41. Gram Vaani, "The Mobile Vaani Manifesto."

42. Mobile Vaani, "Mobile Vaani Legal Info," http://mobilevaani.in/vaani/#/1/legal, accessed June 11, 2021.

43. Kyle Wiggers, "Otter.ai Raises $50 million for AI Transcription," *Venture Beat*, February 25, 2021, https://venturebeat.com/2021/02/25/otter-ai-raises-50-million-for-its-ai-transcription-service/; *Crunchbase*, "Otter.ai: Funding, Financials, Valuation & Investors," https://www.crunchbase.com/organization/aisense-inc/company_financials, accessed April 27, 2021.

44. It is actually unclear whether Zoom users even have an option *not* to share their voice data with Otter.ai once automatic transcription is a possibility.

45. Mary Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (New York: Houghton Mifflin, 2019), xxii.

46. Otter.ai's official line is that its service is provided entirely by machine learning systems. CEO Sam Liang claimed in 2019 that Otter.ai's system is completely original, though it made use of some previous academic work in the area. Otter.ai's approach to AI is part of a larger field of Natural Language Processing (NLP). But it is extremely unlikely that machines are doing everything; every major critical study of what appears to be automation has revealed a heavy reliance on low-paid click workers. See Crawford, *Atlas of AI*; Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven, CT: Yale University Press, 2018); Gray and Suri, *Ghost Work*. For more on Otter.ai's self-representation, see Jeb Su, "CEO Tech Talk: How Otter.ai Uses Artificial Intelligence to Automatically Transcribe Speech to Text," *Forbes*, June 19, 2019, https://www.forbes.com/sites/jeanbaptiste/2019/06/19/ceo-tech-talk-how-otter-ai-uses-artificial-intelligence-to-automatically-transcribe-speech-to-text/?sh=59a087ec3872; *Crunchbase*, "Otter.ai: Tech Stack, Apps, Patents & Trademarks," https://www.crunchbase.com/organization/aisense-inc/technology, accessed April 27, 2021; Otter.ai, "Subprocessors," https://otter.ai/subprocessors, accessed April 27, 2021. For more blogs about Otters, see e.g., *Discourse on the Otter*, https://discourseontheotter-blog.tumblr.com/post/133791821995/simone-de-beauvoir, accessed April 30, 2021.

47. Otter.ai, "Otter.ai Terms of Service," https://otter.ai/terms, accessed April 27, 2021.

48. See, for example, the University of Colorado's "Zoom Accessibility Best Practices," https://www.colorado.edu/accessible-technology/resources/zoom-accessibility-best-practices, accessed April 27, 2021. Even knowing all of the problems with Otter.ai's approach (and after explaining it to students), Jonathan went ahead and enabled automatic transcription in his courses in 2020–2021.

49. Yeshimabeit Milner, "Abolish Big Data." *Medium*, July 9, 2019, https://medium.com/@YESHICAN/abolish-big-data-ad0871579a41.

50. Radin, "Digital Natives," 59.