Market-Driven Voice Profiling: A Framework for Understanding

- *Nick Couldry* (bio) *and* *Joseph Turow* (bio)

Abstract

This article creates a framework for understanding the philosophical and social implications of an emerging marketing driven business that "robs" people of autonomy over their voice. The focus is on the "voice intelligence industry" whose goal is to analyze people's voices to draw conclusions about their emotions and personalities and, long-term, and to use biometric features that leak from "voice" as the basis for target marketing and influence. These new practices, in effect, turn voice against itself by inferring from a person's sounds and syntax feelings, ideas, or beliefs that have perhaps not actually been expressed, and which that person might not want to acknowledge. We use the concepts of quaternary relationships, seductive surveillance, and habituation to explain how this new industry's influence and power emerged and, on that basis, argue that its outcome is to disrupt the very basis on which voice might be valued in social and democratic life.

Keywords

Alexa, Amazon, autonomy, bioprofiling, consent, digital personal assistants, freedom, Google, listening, marketing, seductive surveillance, smart speakers, voice, voice intelligence

This article concerns today's emerging "voice intelligence industry" whose goal is to analyze people's vocal-chord sounds and speech patterns to draw conclusions about their emotions and personalities and so market successfully to them, often in real time. Workers at customer service centers already discriminate among callers based on what they conclude each person's voice reveals about their emotions, sentiments, and personality. Meanwhile, techniques of voice intelligence are expanding deep into people's homes. Amazon's Alexa and Google Assistant use machine learning to extract voiceprints of people for identification, marketing analysis, and more, while Amazon's Halo wristband presents its wearers with profiles of their alleged emotions based on the sounds they make. In the not-so-distant future the goal may be to use people's weight, height, age, and ethnicity—all things scientists believe leak through from the voice[1]—as the biometric basis for targeted marketing communication. Extensive public debate has emerged about other forms of biometric data extraction, for example, facial recognition,[2] with Facebook recently announcing the ending of its facial recognition function. But to date current and potential voice-profiling activities have received little attention among public advocates or academic researchers, even within the marketing industry. This article explores why not.

The term "voice" and the related term "listening" have generated a large body of literature.[3] Our concern is with a new industrial practice that potentially disrupts the very basis on which voice has traditionally been valued in societies: as one reliable token[4] (among others) of what a person *intends* to express, a token of the account they wish to give of themselves.[5] Physical voice production (the target of the voice intelligence industry) combines three levels of performance. One is the physical production by the body of vocal sounds, which, in part, is beyond the speaker's control. The second is the linguistic patterns in the words uttered through those sounds; those patterns may, in many respects, be beyond the speaker's intentional control. A third level covers the meanings or semantic

content expressed through those words. Voice intelligence uses various forms of machine learning to find patterns in what people say and how (physically) they say it, patterns that lead to inferences about the person whose voice it is. Voice profiling particularly foregrounds the first and second *non-semantic* levels of voice, enabling the industry to transform voice into a fixed biometric and *quasi-biometric* asset from which emotions, sentiments, and even physiology can be inferred. Such inferences may contradict what the voiceholder would intend to say. [6]

Marketers can combine such inferences (accurate or not) with demographic, lifestyle, and company purchasing history information to score people as more or less valuable, show them different products based on that valuation, give them targeted discounts, and vary the treatment they receive when they want help. By making statistical inferences based on training data derived from the vocal production and linguistic patterns of multiple individuals across an indeterminate number of contexts, the voice intelligence industry gives new weight to the non-semantic levels of voice. This potentially lifts voice interpretation out of the context in which voice has primarily mattered until now: human-to-human interaction. As a result, the people about whose voices inferences are generated may have no knowledge of such inferences, and no context from which to infer what they might be. Yet those external inferences get made. In this way, marketers "rob" people of autonomy over their voice. Biometric and quasi-biometric voice production and verbal patterns come into conflict with, and potentially even override, voice as intentional expression by the speaker. In some respects the situation is analogous to how marketers generally draw inferences from digital tracking that are sometimes at odds with people's self-interpretation.[7] But voice intelligence matters specifically because it conflicts with a longstanding belief that voice is as a medium through which people are free to express their relationships to, and participation in, markets and wider society.

Our goal here is to create a framework for understanding the philosophical and social implications of an emerging industry that robs people of autonomy over their voice. We first summarize some recent empirical findings about the voice intelligence industry: how it emerged, the latest techniques, and where it is likely heading. In the second section, we explore voice's significance for why we value democracies and markets, and why nonetheless this potential huge expansion of biometric surveillance has not produced the controversy generated by other forms of biometric surveillance. We argue that one reason is the lack of a theoretical approach to understanding how voice profiling is expanding so effectively and with so little resistance. The interrelated concepts we offer in constructing that approach are *quaternary relationships, seductive surveillance*, and *habituation*. We conclude by considering the wider societal implications if, as societies, we fail to recognize the importance of expanding voice intelligence in daily life.

The empirical basis of this article comprised a wide-ranging investigation of the emerging voice intelligence industry by one of the authors.[8] This involved in-depth interviews of forty-four marketing executives and technology experts; reading over one thousand trade magazine and news articles on the topic; scrutinizing hundreds of pages of US federal and state and EU laws, and dozens of patents;

and exploring the history of voice technologies, especially in marketing. That study provided the starting point for this article's theoretical exploration of why voice intelligence matters, and why so far it has not generated the resistance that other forms of biometric surveillance have.

Voice Profiling: The "New Science" of an Old Practice

While humans have always tried to infer meanings from sounds and inflections of others' voices, in the twentieth century researchers turned the activity into a science of voiceprints. Voiceprints are recorded samples of sounds individuals emit through their vocal tract, analyzed for various characteristics and typically transformed into mathematical expressions with graphical representations. Such explorations have yielded signs that individuals *unknowingly and unintentionally* offer up about themselves when they talk, including gender, weight, height, heart rate, general and specific health issues, and even birth control pill use.[9] To achieve these findings, twentieth-century researchers searched for associative patterns between people's voiceprints and their body features. Twenty-*first*-century investigations do not require human observation to see such relationships directly. Researchers contend that, through machine learning and deep neural networks, computers can discover population-level patterns of voice production not graspable by the human ear. Load a computer with voiceprints, it is proposed, and let them figure out—controlling for age, weight, and many other body features—whether patterned links exist between voiceprints and body characteristics of interest. If analysis finds a general relationship, computers will be able to trace it in individual voiceprints. Computer scientist Rita Singh suggests that in time it should be straightforward to learn whether a person's voice betrays facial appearance, body size, psychiatric illnesses, physical illnesses, age, intellectual capacity, sexual orientation, drug use, eventually perhaps emotion[10]—without any human observation of those properties or interpretation of voice's semantic content.

Scientists and medical practitioners have been listening to bodies for centuries,[11] but this new voice analysis is broader and more fine-grained in its implications than previously. The agenda is still incubating in university computer science labs, psychology programs, and business schools,[12] but powerful external organizations have been paying attention. The emerging voice intelligence industry includes companies such as NICE and CallMiner that provide software for company contact centers; conglomerates such as Amazon, Google, Apple, and Samsung that have developed digital products to profit from processing users' voices; firms such as Spotify, Meta, Pandora, and Bank of America that have patents or adjusted their privacy policies to allow for voice profiling; software firms and that can help with the creation of voice-intelligent apps; and arms of ad agency holding companies such as Publicis and Omnicom that are trying to figure out how to exploit the new developments for their clients. At present, most of these actors' interrogations of voice production draw conclusions about emotions and personality. This section draws on interviews with each type of actor. They lasted from around forty-five minutes to well over an hour.

There is irony in pursuing emotion and personality by analyzing voice production, and not more basic biomarkers such as illnesses and heartbeats.

Singh details how researchers' ability to note certain emotions through their measurable effects on voice production derives from a relationship among nerves in the speaking individual, especially "the vagus nerve" which is "highly implicated in the body's response to emotion." But Singh also underscores that speakers' different cultural interpretations of emotions and the subjective nature of personality make it, at present, difficult to reliably label a voiceprint in that way.[13] Nevertheless, in their urgent quest to sell new ways to discriminate among individuals, voice analytics firms claim to their marketing clients that they identify not just individuals' momentary emotions, but also their stable personality, from voice quality alone or along with word patterns. "Today we're able to generate a complete personality profile," contends the CEO of Voicesense.[14] Firms like his are united in the belief they have discovered signals that people give off unconsciously through their voice production.

Such claims by marketers and voice analysts to track emotion and other personal characteristics via voice need not be proven scientifically valid for them to be concerning. Our argument rather is concerned with social embedding of *assumptions* about the likely success of voice analysis. Marketers are steadily incorporating these ideas into their activities, and over time these will affect the organization of everyday life and, as such, are likely, through the social forces we shall analyze, to change for the long-term our broader social and cultural relationship to voice as a key dimension of freedom. Indeed, one key difference between voice identification and other biometric techniques such as facial recognition is that, for the latter, their social embedding in the organization of daily life have been investigated, in part because of its highly public nature.[15] Until now, the social embedding of voice recognition techniques in the organization of daily life have been very little explored.

The Path to Sales-Oriented Voice AI

Interpreting voice has always played a part in marketing. For millennia the illiteracy of most potential customers meant that speaking was often the only way to gain attention and complete the sale. Many deals were one-on-one, and that meant taking stock of the potential customer's voice and social relationships. Sellers and buyers negotiated prices "within the context of a personal relationship, and through the manoeuvrings of trading and bargaining," according to one historian.[16] Analyzing a shopper's voice for confidence, tentativeness, questions, or other emotions was, and still is, an integral part of the ways one-to-one sellers went about their work to close a deal. Piercing a customer's attempted veil of privacy was recognized as an integral part of the salesperson's job.

While one-to-one selling in stores sometimes took over where peddling and door-to-door salesmen left off, much of the buying in department stores and supermarkets was self-service from the late nineteenth century. This development coincided with the growth of the advertising industry as a major force for pushing products in newspapers, magazines, posters, and other media. Circulation numbers, coupons that readers mailed in to the manufacturer, and opinion surveys were among the early ways print publishers tried to keep track of what people sometimes called the "voice" of their readers.[17]

Radio in the US brought voice back into selling, but not live interactions between buyers and sellers. The peculiarity, for the first time in history, was that while the seller could speak to the audience, the audience couldn't speak back, except indirectly through letters and phone calls. Broadcast television, introduced commercially in the late 1940s, had the same historical oddity. In fact, to assure marketers that their advertising worked, broadcasters hired companies such as Nielsen to survey samples of Americans about what they listened to or watched.[18]

With the rise of the commercial internet in the mid-1990s, mediated vendors increasingly interacted with customers in a manner that went beyond Nielsen's broad social segments. Yet as internet marketing has developed, major problems behind the scenes of clickstream sales interactions have become clear. Demographic, psychographic, and behavioral data may not be up-to-date, profiles may be based on multiple users of a computer or phone, names may be confused, people may lie about age, income, even gender in the hope of confusing marketers. Advertisers are discontented with problems of click fraud by websites (as high as 28% of all web traffic[19]) and ad blocking by web and app users. Parallel to these difficulties is the crescendo of public anger about marketers' surreptitious tracking and targeting,[20] anger that has resulted in wide-ranging laws in the European Union (especially the General Data Protection Regulation or GDPR) and initiatives by US states such as California and Illinois to provide their citizens with more legal leverage with digital marketers than the federal government has done. These regulatory pressures encouraged Google's announcement that it will prohibit cross-site (third-party) tracking cookies in its Chrome browser by 2023 and Apple's decision to require that apps ask for customers' explicit consent to track them via Apple's mobile advertising ID.

Acknowledging these issues, marketing practitioners are today discussing voice intelligence not as a substitute for tracking known people online, on apps, and in stores, but, in the words of marketing consultant Pete Erickson, as a "value-added" to the current personalization regime.[21] Individuals can not only be profiled by what they say and where they say it, but also by underlying linguistic patterns of their talk and the physiology of the sounds their voices make, foregrounding the non-semantic layers of voice. These latter phenomena, marketers believe, cannot lie, so converting the flow of spoken expression into a basis for fixed biometric and quasi-biometric identifiers. However contentious such inferences *might* be, they currently occur inside individual corporations without public scrutiny.

Voice Exploitation by Contact Centers

The customer phone service (or "contact center") business moved first into profiling from individuals' unique voices. A small number of companies such as NICE (the largest), Cogito, and Voicesense have dominated the voice analytics business. They create voice analysis software and rely on technology firms such as AWS, IBM, Dell, SlashDB, and Microsoft to provide other software and/or hardware tools, including storage and networking. Their clients represent a wide gamut of consumer-facing firms such as insurance companies, banks, airlines, consumer package goods firms, and hotel groups. The voice analysis software evaluates a caller's sounds and linguistic patterns for emotion, sentiment, and

personality. It can carry out this inference while the person is speaking to a computer before being sent to a human agent, or while the person is speaking to the agent. In the former situation, the software can draw conclusions about the person's emotions and triage the caller to an agent that research has shown is likely to satisfy those kinds of individuals (for example, angry, logical, worried) and even "upsell" them—lead them to spend more money. When the caller is speaking directly to the human agent, the software can send the representative messages regarding the appropriateness of the agent's speech in view of the caller's emotions, sentiment, or personality.

Not all call center firms use voice analysis. Some analyze a caller's words for clues to emotion. Yet the call business analytics firms that interrogate voice as well as words and syntax boast that by using artificial intelligence on the huge number of customer-agent conversations coming into their systems, they can predict the likelihood a person will recommend the firm to a friend or colleague, the customer's sense of how quickly and easily the company helps solve his or her problems, and a general score of customer satisfaction. Customers likely have no clue that a call center is turning their statements into pro or con viewpoints. Another part of this interaction that is not transparent is the contact center's use of data to identify and respond to each caller's emotional state, personality, and sentiment. The industry defines *sentiment* as a combination of attitude and emotion toward a specific company or its product. The speech analytics procedures of contact center computers explore the text for word- and phrase patterns that in time can generate means for the firm to increase satisfaction.

In the trade press, contact center executives are exuberant about the revenue voice intelligence is generating for their clients and the wider interpretative potential of the activity. Voicesense claims from people's voiceprints to accurately predict loan defaults, likelihood of filing insurance claims, and customers' investment style, among other key indicators.[22] Andy Traba, an executive at voice analytics firm Mattersight (a subsidiary of NICE), predicts that "in the same way that there's information that's associated to an IP address . . . there's going to be information that's associated with my voiceprint, which is unique as my fingerprint."[23] This opens up potential conflicts between such "information" (a fixed biometric identifier) and the information that voice owners *think* they communicate through their words.

Voice Profiling by Digital Assistants

Most of the activities of the contact center business take place outside the public eye. The first steps to introduce voice profiling into everyday life have been through dedicated domestic devices, including "smart speakers," notably offered by Amazon and Google. Nearly 90 million US adults—about one in three—owned smart speakers in early 2020, according to voice industry research firm Voicebot.ai.[24] Speakers collect the voiceprints, and recognize the voices of tens of millions of individuals who speak to them. Amazon and Google assistants also allow their owners to interact with a variety of interconnected devices that turn on lights, set the home's temperature, monitor doorways, and perform other activities. Car companies, home builders, hotels, and even schools are also using

smart devices, smart assistants, or both. Watches and wireless earbuds can send people's voices to various companies (typically via phones), depending on their creators' creativity and the marketers' desires.

The companies are not yet applying these tools for their maximum analytic and marketing potential possibly because they worry about inflaming social worries around this instrumental use of voice. Yet Amazon and Google reserve the right to use voice profiles for their own marketing purposes, and advertising executives we interviewed expected they eventually will.[25] Both companies have staked out numerous patents that suggest broader marketing opportunities. One patent[26] asserts that Amazon's Echo could hear a sniffle in your voice, infer a cold, and offer to deliver aspirin to you within two hours.

An early public implementation of this technology is Amazon's Halo health and wellness wristband. Released for purchase only to certain Amazon customers in fall 2020, the band is sold as having the ability to analyze the tone of its owner's voice for "qualities . . . like energy and positivity."[27] As part of its sales pitch Amazon declares that getting people to consider the emotions that their voice emits will encourage them to adopt healthier communication practices with their loved ones and their bosses. The company asserts that Halo's security features keep its analysis off limits to third parties; the voice profile, too, is explicitly not for use by third parties. But Halo's capability must be seen as a proof of concept for potential wider uses. The voice profiling idea demonstrated here can, as the sniffle patent suggests, be easily ported to marketing and beyond.

Video 1.
Amazon's Halo wristband monitors the wearer's activity, sleep, and voice.[28]

Click to view video

As they look to attract users, Amazon and Google limit advertising on their smart devices to just a few types of voice "apps" (technically called "skills" by Amazon and "actions" by Google). They allow owners of smart speaker apps to ask users to identify themselves, track their activities, and link such data to other information about the individuals. Amazon and Google give app owners transcripts of exactly what individuals say, though they are not yet sharing individuals' voiceprints. Marketing executives worry that Google and Amazon (who keep the voice recordings of what users say to their speakers) might use

people's talk to other companies through their smart speakers to learn competitive information about customers.[29] The concern has led some firms such as Bank of America and the Pandora music service to create *their own* voice assistants that personalize relationships with customers on the web or phone.[30]

Also wary of relying on Amazon and Google, a few major advertising agencies are themselves working on methods to infer customers' inclinations from how they speak.[31] Their executives express confidence that a combination of these voice-profiling approaches will eventually become part of the toolkit for managing their clients' relations to customers across a panoply of smart devices. Parallel to these developments, a number of voice intelligence firms have turned their attention to Wall Street. They create algorithms to help hedge funds and other investment firms profile CEOs' voice and word usage to gauge their "real" feelings about their companies. The investors use the inferences in their calculations about the actual solidity of the firms.[32]

Voice technology is therefore beginning to permeate important areas of personal and business life. What is particularly troubling about such voice profiling is that it acts on information that may be in direct tension with how voice owners choose to express and represent themselves *through* their voices. When voice is used as a fixed biometric or quasi-biometric identifier, it potentially is in tension with the non-fixed, open ways in which we use our voices to express what we mean to say. In the next section, we explore in more detail why this matters, and how, so far, consumers and citizens are being lured into not caring.

Why Does Voice Intelligence Matter and How Is It Spreading?

Voice intelligence converts the sounds made by our vocal cords into a fixed identifier that can be tied to limitless other information about individuals for multiple predictive purposes. As such, voice intelligence is part of the long-term growth of biopolitics[33] and the massive expansion of surveillance to govern populations.[34] In this section, we explore why voice intelligence might matter specifically, and why nonetheless large numbers of people are seemingly welcoming it into their homes.

Voice and Participation

Voice intelligence uses sound production as an entry point for biopolitical governance. While sound studies and STS-inspired approaches to sound and noise have explored various relations between sound and power, they have not generally explored the use of voice as reference point for measurements within a wider toolkit of predictive governance. So Li and Mills[35] explore machine-based speech recognition, but outside of the wider political economy discussed here, while Carmi[36] studied practices of listening-in to conversations as a form of power within communications industries, but without a link to biometric identification. The voice intelligence industry challenges us however to theorize the significance of recent developments whereby voice as individual sound production becomes a direct target of power.

Such developments matter because they are in tension with the way voice's expressive, interactional function has been valued by social and political theorists for more than two millennia. Aristotle distinguished between two uses of humans' vocal powers: to express "pain or pleasure" (as with other animals) and "to indicate what is useful and what is harmful, and so also what is just and what is unjust" (unique to humans).[37] To Aristotle it is from this latter aspect of voice (speech) that politics derives. Hannah Arendt[38] regarded speech—the verbal accounts people give of their lives—as giving substance to their actions, revealing a human being *as* a person. Voice, in short, is a faculty through which people have been understood to participate not just in society, but in democracy.[39] Models of the economy also understand voice as a form of relational self-expression essential to market functioning. Consider Albert Hirschman's[40] theorization of markets as social structures. The exercise of "voice" about products and services is, Hirschman argued, a better outcome than consumers just "exiting" from the market, since, in allowing the customer to express emotions and opinions, markets build "loyalty." For Hirschman, recognition by people of their voice's influence in the marketplace contributes to their sense of self-worth, and so reinforces their economic participation.

Common to all these understandings of expressive voice is a broader notion that underlies why democracy and markets are valued as domains of potential freedom. The continuing assumption has been that people have broad control over their self-expression; otherwise "the revelatory quality of speech"[41] would reveal nothing. We cannot control how others respond, but we can still choose how we, in turn, respond to them. Such symmetry of action and reaction through the expressive exercise of voice is basic to our understanding of social interaction. But in the era of Big Data, the large-scale continuous *asymmetric* extraction and interpretation of data is undermining that symmetry. Through voice intelligence, voice starts to "speak double"—that is, in two potentially conflicting registers, only one of which (expressive voice) can be under the speaker's control. Speakers may believe their intentional expressive participation in the market economy affects how the sellers treat them, while the inferences sellers make about the individuals' desires from their voice may sometimes actually determine particulars of the marketing relationships.

Physical voice is not the only action through which we participate in society. Participation (or "voice" in a more general sense[42]) can involve multiple media, and market participation can comprise data entry and clicking on a screen. But the use of our physical voice remains a key index of free participation in wider society, and yet that same use of voice is now being converted into a biometric identifier, which "speaks" to power in fixed ways. As such, AI-driven voice recognition and identification should matter at least as much as other biometric identifiers such as facial recognition.

Why therefore has there been no outcry about this new use of voice? To understand this, we need to understand better the sorts of social relations through which voice intelligence is being applied in contemporary society. Voice intelligence is certainly part of the imposition of surveillance capitalism[43] and the biopolitical public domain.[44] But both those frameworks downplay the role of consent, with Shoshana Zuboff arguing market surveillance relations are imposed

“typically . . . in the absence of dialogue or consent”[45] and Julie Cohen arguing that changing legal and platform structures “leave[] individual consent “with very little work to do.”[46] There is indeed a degree of market force at work in how, for example, we accept our calls to banks and other service providers being recorded: we need, after all, that service. The expansion of voice intelligence, as we have shown, already goes much further to include the voluntary introduction of devices into the home. We need therefore different theoretical concepts to grasp *how* voice intelligence is spreading in society so effectively, concepts that clarify the type of relations into which populations are being encouraged.

Seducing Voice against Itself

While Aristotle, Arendt, and Hirschman focus on voice relationships between people, we need to better understand individuals’ voice relationships with technologies. Here three concepts (quaternary relationships, seductive surveillance, and habituation) can help us go beyond top-down frameworks of surveillance and informational capitalism.

A neglected essay by Craig Calhoun[47] provides the first building block. Calhoun takes off from the early twentieth-century sociologist Charles Cooley’s idea of primary and secondary relationships (close personal relations and those mediated by a social role) to theorize two new forms of human interrelations via technologies in the digital era: tertiary and quaternary relationships. Tertiary relationships involve people interacting with other people, mediated by impersonal systems such as telecommunications technologies but always with some sense that a human is involved. *Quaternary* relationships move further toward the machine world, involving no interactions with people, only systems that influence people.

Examples are plentiful. A person writing to her bank knows she is communicating via a large communications system within the bank, but somewhere imagines a person reading and responding to the letter (a tertiary relation). A quaternary relationship can emerge from that tertiary interaction. Say you phone a call center, angry about a lack of response to that letter. As you speak to the computer that greets you, that machine sends information about your voiceprint to another computer. It, in turn, tells the first computer you should be connected to someone who knows how to deal with your specific emotion. Quaternary relationships operate “outside of the attention and, generally, the awareness of at least one of the parties to them. They are the products of surveillance and exist wherever a sociotechnical system allows the monitoring of people’s actions.”[48] Nevertheless, Calhoun points out that even quaternary relationships mimic social relations. They “turn” the actions of those who indirectly interact with such systems “*into communication*, regardless of the actors’ intentions,”[49] that is, whether or not people intend to be communicating!

Writing this in the early 1990s, Calhoun extrapolated from late 1980s data banks built from credit card and airline booking transactions. He captured in general terms the sort of social relation we enter when we phone our bank (and learn that our voice is tracked in the background) or install a digital personal assistant like Alexa (and learn that our talk is mined for data). Amazon also tries

to lead us into feeling our interactions via Alexa are like tertiary relationships—relations with a person, albeit an artificial one. Yet Alexa actually is, in Calhoun's terms, the entry point to a quaternary relationship with a computer that learns how to react to us by drawing on algorithms operating across multiple computers. Although its humanoid responses (helpful or not) may invite further interactions, the machine-to-machine nature of the interactions—three steps removed from personal human relationships—are rooted in opaque patterns learned by machines from unknown training sets, and they may extract from our interactions levels of "communication" of which we are not explicitly aware. Knowledgeable consent to the actual operations achieved through such data use would seem to be near-impossible, and yet Google, Amazon, Siri, Pandora, Bank of America and many other companies routinely ask people for permission to use their data and get it. How do they obtain this permission?

One reason is the public's lack of knowledge about how firms use people's data.[50] Another is that large sections of the public are resigned to companies taking their data; two national surveys in the US found that around 60% of Americans said they would like to control the data firms have about them but don't believe they can.[51] Lack of knowledge and resignation are important background factors for the large uptake of smart speakers and personal assistants. Yet the evidence of how quickly voice intelligence is being embedded in daily life suggests there is another, seemingly more positive feature at work. To understand this, we also draw on Pinelopi Troullinou's insight that marketing companies today create an affirmative culture through "seductive surveillance."[52]

As Troullinou defines it, *seductive* surveillance is the activity of building a compelling environment around a technology while playing down the corporate surveillance that is its core rationale. Going further than Troullinou, we suggest that companies carry out seductive surveillance on multiple levels. The most basic seduction is the device: an assistant embedded in home life that can respond to everyday requests made via speaking rather than typing; alternatively, a customer service interface that exists to receive customers' spoken inputs, that is, "hear" their voice. Then there is seduction through the device's own friendly voice. Voice intelligence as data extraction relies on our own positive association between voice and someone's self-expression, even as it exploits it: human beings still want their meanings (indeed their emotions) to be understood, and their voices listened to (that, after all, is one standard way we have understood ourselves to participate in society). Then there is seduction through price: smart speakers and other devices that interact with people's speech (for example, thermostats, light switches, home security cameras) are sold at large discounts on events such as "Amazon Prime Day" or "Cyber Monday." In a 2019 interview with one of the authors, an Amazon executive acknowledged his firm wasn't making money from selling the devices because prices were so low. The same year Google's executive in charge of smart speakers publicly stated the firm wasn't yet profiting from its smart speaker sales.

As a corporate strategy, seductive surveillance uses "intimacy at scale"[53]—the ability to induce quaternary relationships that give individuals the illusion of intimacy, even as they monitor and manage them through automated data collection and analysis. Marketers stand to gain if they can persuade people that

their automated techniques for interpreting our voice production and word-choice patterns mimic human beings' perennial ability to interpret each other's feelings and meanings via speech: for the latter is one way human beings have come to expect their efforts at participation to be socially recognized.

Forces of habit, in turn, are starting to stabilize this emerging intimate relation between consumers and quaternary relationships of data extraction. Bennett and Dodsworth[54] use the term *habituation* to refer to the process by which forces in society cultivate the creation of habits. Habit helps organize behavior individually and collectively without requiring explicit rules.[55] Voice capture and voice analytics are built into mass-market goods (digital personal assistants) that offer new ways of performing habitual acts such as checking the weather or ordering a pizza.[56] It's not hard to see how habituation and seductive surveillance converge. Seductive surveillance is a dual strategy— both persuasion and nudge—by which companies get people into the habit of speaking to a range of devices, in the interests of apparent convenience. Electronics stores, new-home builders, auto manufacturers, hotels, schools, and stores are all integrating digital personal assistants into their activities, with producers foregrounding the seductive, and playing down the surveillant, aspects of these devices, a message reinforced by the commercial and tech press.

Undergirding this expansion of digital personal assistants and background uses of voice intelligence is, we suggest, the deepest seduction of all: the idea that this is done to know human beings better. Until now, voice has been assumed to contribute to human self-knowledge, as an exercise in *self*-expression. Now proponents of voice intelligence claim a new form of social knowledge through which computers *bypass* the awareness and intentions of the human subjects whose voice is captured.[57] This knowledge comes cloaked in the authoritative language of science, even as it disrupts our everyday understandings of why voice matters. This "changing relationship between ways of knowing and forms of power," based on "large-scale strategies of correlation, prediction, and pre-emption" characterizes Big Data techniques generally,[58] but in voice intelligence it finds a distinctive form. Management theorist Jannis Kallinikos[59] captures how such new knowledge, even if it appears intimate, works to override a core feature of traditional voice interpretation, that is, *contextual meaning*. The massive production of data, as in voice intelligence, necessarily detaches potential sources of information from practices of meaning-making in particular social contexts. As Kallinikos notes, "databases . . . contain data but scarcely information, if by information is and should be meant the *living, actively sought semantic content social agents* draw on, in pursuing their objectives."[60] Kallinikos here captures in general terms the tension generated by the voice intelligence industry's new claims of knowledge at a distance, as it overrides the knowledge people normally gain from interpreting each other as they participate in social interaction.

In the quaternary relationships common to the voice intelligence industry, speech as data is asymmetrically, though seductively, extracted, generating value via opaque algorithms from the sound and syntax, as well as the words, the person articulates. This process transforms how voice contributes to social knowledge more generally. When we call out "Alexa!" we invoke the traditional

relation of human voice to meaning; but when Alexa “speaks back,” it is not a human voice speaking, but the interface of an asymmetrical quaternary relationship. As human beings, we are still disposed to interpret this interaction as meaningful, seduced by the meaning that voice as a human faculty still carries for us. But in accepting these new quasi-interactive habits, we are starting to accept unwittingly a new way of instrumentalizing voice that bypasses the human-to-human interpretive contexts that have until now made voice meaningful as social participation. So instrumentalized, a request to Alexa might in the future designate an individual as more obstinate than others and so not worthy of a good discount on a product. An interactive political campaign ad may lead a person to receive on-the-fly messages that suggest sentiments at odds with how the individual sees herself politically. A refugee might not gain entry into the country because algorithms indicate that the individual’s accent and emotional valence reveal the person is not truly at risk. Voice risks being seduced against itself on terms that voice owners can do little to monitor or control.

Voice Intelligence: The Potential Social Costs

In service industries, via digital personal assistants, and in emerging advertising contexts, marketers make voice technologies appear unthreatening and even alluring. All these developments are in their early stages. The contact center business has taken the lead in exploiting voice, and it will take time for other sectors (advertising on smart speakers, for example) to catch up. That said, secrecy and obfuscation within the smart speaker sector make it nearly impossible to determine the extent to which Amazon and Google themselves are *already* taking advantage of the voiceprints they are collecting.[61] We have argued for the long-term significance of voice intelligence in undermining our long-term understanding of how we participate in the marketplace as well as in the larger society. We have also explained how voice intelligence is spreading with little resistance by embedding itself in new relations between device users and corporations. As such, voice intelligence’s power cannot be explained merely as an imposition from above, as theories of surveillance and informational capitalism[62] imply, but needs to be understood through the concepts of quaternary relations, seductive surveillance, and habituation within familiar settings of convenience. The theoretical framework we have laid out allows us to understand how, as an important form of biometric identification, voice intelligence is becoming normalized, and so deepening the surveillance society. As yet, though, there has been little social resistance to voice intelligence, unlike with facial recognition technologies, which many including academics have called to be banned. We conclude by asking about the consequences for society if we continue to accept voice intelligence without resistance.

The voice intelligence industry is potentially creating a biopolitical public domain (Julie Cohen’s term) very different from the market society imagined by Albert Hirschman, let alone the public space of appearances imagined by Hannah Arendt. Our freedom to participate in markets and public space will be increasingly troubled by the fear that, when we speak, we simultaneously risk speaking against, not for, ourselves. Such fear has already been expressed in relation to facial recognition techniques, which transform what it means to move in public space with face uncovered, because “people do not and cannot possess

an appropriate level of knowledge about the substantial threats that facial recognition technology poses to their own autonomy."[63] The parallel risk when our voices are continuously tracked and analyzed has not so far been much noticed, let alone resisted. Many contemporary societies understand themselves through the value of participation in economic markets and political spaces that they allow. Every individual's voice is an index of such participation, even if it is not the only medium of participation. If we allow voice to be reduced to a pervasive biometric identifier whose fixed readings are *normally* at odds with individuals' attempts at self-expression, we risk undermining the core value of participation. As with any practice at odds with underlying values, there is the long-term risk, as legal theorist Nancy Kim puts it, that "the law will arrive too late, after [new] social norms have already been established and when it is much more difficult to reverse society's course."[64]

As Charles Taylor argued, "our notions of freedom—both personal independence and collective self-rule—have helped to define a political identity we share; and one which is deeply rooted in our more basic, seemingly infra-political understandings: of what it is to be an individual, of the person as a being with 'inner' depths."[65] But as Taylor sensed more than three decades ago, "the growth of modern control has involved . . . a dehumanization, an inability to understand and respond to some key features of the human context, those which are suppressed in a stance of thoroughgoing instrumental reason."

The voice intelligence industry, along with the practices of surveillance capitalism more generally, represents an advanced version of this conflict. The potential implications for social discrimination and inequality are serious. Several industry practitioners interviewed[66] said that there is already far more analysis of what people say and how they say it than companies let on. They contend Google, Amazon, and firms they work with are waiting for the "scale" of smart speakers and related devices to grow until voice assistants are integrated into virtually everyone's domestic and professional routines. Then the firms will be able to shift into high gear: people will routinely get personal buying suggestions, search results, map destinations, and ads based on what firms conclude about them through a combination of data points including speech, demographics, behavior, psychographics, and location—all integrated into what we might call Voice+ profiles. We have become used in recent decades to receiving differential offers and opportunities based on being tracked digitally and on various facts about us—such as income, where we live, our race and gender. We have not yet become used to being profiled simply because we open our mouths, based on physiological characteristics of our voice production or linguistic patterns that we typically aren't aware of. Whether or not the inferences marketers make are accurate is far less important than if marketers believe they are accurate and act based on that. What if voice profiling tells a prospective employer that you're a bad risk for a job that you need, or tells a bank that you're a bad risk for a loan? What if a public advocacy organization won't take your donation because its voice-based algorithms profile you as gay? And what if the racial discrimination now shown to be associated with algorithmic practices more generally[67] were to become silently embedded in how voice intelligence operates? There is a longer history of how judgements based on sound have been tied to racial

discrimination,[68] but voice intelligence could embed such ties more insidiously and pervasively than before.

Because voice technologies' seductive sounds and helpful demeanor encourage widespread interest (compared to the visceral anger and concern that facial recognition often evokes), voice profiling's work may provide an effective entry point for institutions interested in getting acceptance of biometric profiling that targets other areas of the body (hands? eyebrows? heartbeat? urine?). As with facial recognition, the debate we wish to encourage does not concern "the West" only. Voice intelligence, like facial recognition, is currently evolving in China too.[69] Both technologies are based on converging global developments in artificial intelligence and machine learning. Wider alliances must be built within and beyond the West that challenge the voice intelligence industry and the risk it poses, if uncontrolled, to the free exercise of voice, and, through this, to the fundamental value of human beings' participation in the cultures of which they are part.[70]

Nick Couldry
Professor of Media Communications