Dockray, Sean. “Learning from YouTube.” *Rivers of Emotion, Bodies of Ore*. Oslo: Uten Tittel (Not Yet Titled Press) in collaboration with Kunsthall Trondheim, 2018

Learning from YouTube

Drifting across YouTube, I seized upon one video that caught my eye: “Why Audio Analytics?”[1] It was an advertisement uploaded by a company called Louroe Electronics for a product that looked like an upturned smoke detector capable of analyzing and detecting sounds “through advanced algorithms.” The video imagined several scenarios to demonstrate the kinds of sounds it’s capable of recognizing: glass breaking at night in the showroom of an automobile dealership, a gunshot in a school hallway, and aggression in a public space. Putting aside the differences in these scenarios — isn’t aggression much more dependent on interpretation and an understanding of context than whether or not a gun is fired? — they are all examples of *machine listening* and mark both a departure from and an expansion of *speech recognition* of the sort built into Siri, Alexa, and Google Assistant.

In speech-to-text, an audio signal containing human speech is converted to a textual representation of the words that are spoken. With the more general “audio event recognition,” however, *all sounds* are mapped to descriptive categories. To give a concrete example what this means: Google has created an *ontology*[2] that defines the conceptual space of 632 possible sound categories. 13 of these are ”human voice” sounds, such as “sigh” or “wail, moan.” Only one is “speech.”

Although this range of sounds is very wide, the early commercial applications of machine listening tend to be in the security and surveillance industries. Audio Analytic, who maintain their own proprietary audio dataset called Alexandria, develop software that is implemented by smart home devices to listen for alarms or break-ins. Shooter Detection Systems provides technology for early detection of active shooter situations, with marketing material claiming that their Guardian System “removes the ’human factor’ so that nothing is left to interpretation and costly delays can be avoided.”[3] Wendy Hui-Kyong Chun discusses another example of using the digital to circumvent the human in

---

[1] “Why Audio Analytics?”, *YouTube*, accessed 3 April 2018, https://www.youtube.com/watch?v=fxg6ZfkgpM8

[2] “Ontology” is the name used by the engineers at Google, drawing from a longer history of use in information science, to establish a “fixed, controlled vocabular[y]” to model some aspect of the world. “Ontology (information science)”, *Wikipedia*, last modified 4 March 2018, https://en.wikipedia.org/wiki/Ontology_(information_science)

[3] ”The Guardian Indoor Active Shooter Detection System”, *Shooter Detection Systems*, accessed 3 April 2018, http://shooterdetectionsystems.com/products/guardian/

her discussion of face-recognition technology in the aftermath of September 11, 2001. The technology "corrects for visual subjective bias by inhumanly bypassing rationalization and deduction,"[4] identifying terrorists by correlating patterns of camera data. Chun also quotes promotional media on the subject: "There is no chance for human error or 'racial profiling' because there is no need for a human operator to fixate on a particular person. The camera does it all automatically."[5]

Part of the argument that Lourou Electronics makes in its video is that by sensing aggression in the environment, it is possible to intervene before a crime even happens. This argument, when combined with the predictive promises of Artificial Intelligence, suggests that it is possible to change the future. Beyond simply preventing an event from occurring, in a preemptive system an individual would never feel directly constrained, but rather would be guided into contexts where undesirable behavior is least probable.[7] Louroe Electronics' detection of aggression and broken glass brings to mind the "broken windows" theory of policing, which asks how to "identify neighborhoods at the tipping point… where a window is likely to be broken at any time, and must quickly be fixed if all are not to be shattered." At the time that the theory was introduced in 1982, the police did not have "ways of systematically identifying such areas,"[8] a limitation overcome by algorithmic surveillance. Palantir Technologies, founded by the Silicon Valley libertarian Peter Thiel in 2004, is a data mining company in intelligence and national security that was recently found to be using the New Orleans Police Department as a testing ground for predictive policing. The preemptive logic of the broken windows theory — hypothesizing that urban disorder cultivates actual, serious crime — is reinforced by these algorithms, which aim to intervene "before an incident turns into a violent outbreak,"[9] mobilizing a fear and anxiety that tends to be oriented towards specific

---

[4] Wendy Hui-Kyong Chun, *Control and Freedom: Power and Paranoia in the Age of Fiber Optics* (Cambridge, Mass: MIT Press, 2006), 263.

[5] Chun, *Control and Freedom*, 262. While I couldn't find the same *New York Times Magazine* source as Chun, I located the same quote in another news item, CBSNEWS.COM STAFF, "Facial Recognition Technology May Screen for Terrorists," *CBS News*, January 2002, https://www.cbsnews.com/news/facial-recognition-technology-may-screen-for-terrorists/.

[7] Antoinette Rouvroy and Thomas Berns, "Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation ?" trans. Elizabeth Libbrecht, *Réseaux* 177, no. 1 (2013): IX, doi:10.3917/res.177.0163.

[8] George L. Kelling and James Q. Wilson, "Broken Windows," *The Atlantic*, no. March (1982), https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/.

[9] "Why Audio Analytics?"

groups of people, like "panhandlers, drunks, addicts, rowdy teenagers, prostitutes, loiterers, the mentally disturbed."[10]

This responsive security environment of sensing surveillance devices is prefigured in Felix Guattari's imagined electronic access card, which Gilles Deleuze recounts in his famous essay, "Postscript on the Societies of Control." The control mechanism – the card – could track the position of its holder, locating an individual in space and time, thereby allowing or prohibiting access somewhere based on some set of rules, which are themselves potentially changing in real time. William Burroughs, whom Deleuze acknowledges[12] for "naming the monster"[13] of control, commented that a sense of free will was *necessary* for control to be effective. If "the workers have become machine-like tape recorders"[14] then they are merely being *used*, not controlled. For Burroughs, control requires incompleteness, or a gap between the controller and the controlled, which is not quite a direct performance of the wishes of the controller, *but almost*.

When Vilém Flusser talks about control, however, he doesn't mean it as a mode of power, but rather as something that has been *lost*, namely freedom: "The crisis of authority has not led to the emancipation of society, but as it allows for an apparent freedom of choice, it has led to the cybernetic totalitarianism programmed by apparatus."[15] In Flusser's post-industrial society, people don't work, they are occupied. Work is left to automatic machines that manipulate the material world into mass-produced objects. Being occupied means that people are *functionaries* and *programmers* involved in the processing of symbols, like a white-collar worker who sends memos and fills in spreadsheets. At a superficial level, programmers are the ones who write the programs and functionaries are the ones who use them. But if we look more deeply, the two collapse into each other. Computer programmers program by pushing buttons in order to manipulate symbols. Every choice made in every keystroke is, however, a choice made within another program, a *metaprogram*. "And this regression from meta- to meta-,

---

[10] Kelling and Wilson, "Broken Windows."

[12] Gilles Deleuze, "Having an Idea in Cinema," in *Deleuze & Guattari: New Mappings in Politics, Philosophy, and Culture*, ed. Eleanor Kaufman and Kevin Jon Heller, trans. Eleanor Kaufman (Minneapolis: University of Minnesota Press, 1998), 17. "the term put forth by William Burroughs… societies of control".

[13] Gilles Deleuze, "Postscript on Control Societies," in *Negotiations, 1972-1990*, European Perspectives (New York: Columbia University Press, 1995), 179.

[14] William S. Burroughs, "The Limits of Control," ed. Sylvere Lotringer, *Semiotext(e): Schizo-Culture* III, no. 2 (1978): 38.

[15] Vilém Flusser, *Post-History*, ed. Siegfried Zielinski, trans. Rodrigo Maltez Movaes (Minneapolis, MN: Univocal Publishing, 2013), 86.

from the programmers of programmers of programmers, proves to be infinite."[16] At every level, programmers are simultaneously functionaries, and vice-versa. This paradox is epitomized on social media platforms, which are both stages for mass-individualized self-expression and highly scripted, addictive frameworks that compel participation.

Social media platforms operate both as sites of control and as machines for aggregating data that can be utilized for future forms of control. In 2017, members of the Sound and Video Understanding team[17] at Google announced *Audio Set*, a dataset of 2 million YouTube videos that aspired to "substantially stimulate"[18] the development of machine listening algorithms. This announcement was accompanied by relatively little fanfare because, rather than news media, it was published on the Google research blog and in an academic paper. It was — and still is, at the time of this writing — an esoteric development, primarily of interest to programmers and machine learning enthusiasts. And crucially, it is temporally *prior* to any particular artificial intelligence, or machine-learning application, that will be developed from the dataset, be that home automation, workplace monitoring, or automated policing. This means it appears to be pre-political, free from the inequality and bias that only seems to become apparent after when it is discovered that an automated system has been, for example, targeting black people.

The videos in *Audio Set* have been randomly selected, so it is unlikely that any of the uploaders know that their content is being used in this way. It's just as unlikely, however, that they would care. Who knows how many ways a video has already been sliced to inform recommendations and advertisements? On YouTube, videos live a double-life as entertainment for a human audience *and* as data for an algorithmic audience and it is the continuous invention of new algorithms that watch in new ways that makes old videos new again. Uranium, after all, was observed in mountains for centuries before it was deliberately mined for radium. Data will gather in server farms for years before it is exploited most profitably.

The mass of YouTube videos in *Audio Set* are akin to the cropped centerfold of *Playboy* model, Lena Söderberg, which was used as a test image for digital image compression research and has been an industry standard for testing imaging algorithms from the JPEG format to Photoshop effects ever since. In the age of machine learning, the test image becomes a massive dataset. Near the end of 2016, the Google announced YouTube-8M, a dataset of 8 million categorized YouTube videos (of which, the aforementioned *Audio Set*

---

[16] Vilém Flusser, *The Shape of Things: A Philosophy of Design* (London: Reaktion, 1999), 93.

[17] Part of the Google Machine Perception Team

[18] Jort F. Gemmeke et al., "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events" (IEEE, 2017), 776, doi:10.1109/ICASSP.2017.7952261.

is a subset) in order to accelerate breakthroughs in machine learning and machine perception.[19] Not long afterwards, Sundar Pichai, Google's CEO, shifted the corporation's strategy to be "AI first." Suddenly, Google's decision to acquire YouTube in 2006 seemed to be less about the human audience than the algorithmic one. It was at this moment that the videos' uploaders had been retroactively automated, crowdsourced without realizing it, becoming memories for an algorithm with unknown politics. Google refashions the past with its corporate machinations and the future through the predictive capacities of its AI work.

The degree to which different kinds of automation abound in the acquisition of data and training of neural networks anticipates the way that artificial intelligence automates certain jobs, including police surveillance. "Broken Windows" was written at a moment of cuts to police forces across the U.S. and should be read, in part, as a strategy for reorganizing policing when budgets no longer allow for foot patrols. Networked surveillance cameras allow few people to monitor many different locations from a distance. But human labor could be reduced even further: each image is confined to the zoom, focus, and orientation of a particular camera, and each image depends on an operator to see what it displays, whereas an omnidirectional microphone covers a much larger area, including spaces outside the frame of the image. Moreover, it is unnecessary, even impossible, for a human operator to listen to all of the audio, so it is instead monitored by algorithm. Not only does this further the conversion of the body of the policeman into electronics and code — much the same way that the 19th century officer has been absorbed into the 21st century traffic control systems[20] — but it enables a kind of just-in-time policing that short-circuits labor-intensive criminal investigations and legal deliberation by preempting criminal acts.

The drive towards automation and control is not limited to policing, even if the effects are often more visible there. Google recently trialed a neural network to predict when a patient in hospital will die, allegedly with 95 percent accuracy. Will data companies triage the sick and infirmed before any medical staff sees them? Maybe there aren't doctors and nurses any more, but technicians. Microsoft believes that signals from web searches can be used to predict cancer and Target has determined pregnancies from purchase histories. It is one thing to make these predictions as if "being right" were the goal, but these predictions are used to reconstruct worlds around individuals, most visibly

---

[19] Sudheendra Vijayanarasimhan and Paul Natsev, "Announcing YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research," *Google Research Blog*, September 2016, https://research.googleblog.com/2016/09/announcing-youtube-8m-large-and-diverse.html.

[20] Sean Dockray, Steve Rowell, and Fiona Whitton, "Blocking All Lanes," *Cabinet*, no. 17: Laughter, accessed April 4, 2018, http://cabinetmagazine.org/issues/17/blocking.php.

in YouTube’s recommendation algorithms or Facebook’s targeted advertisements. The walls closing in are not in the form of a prison cell but molded to the shape of our own bodies. If the future is foreclosed, it is to the trajectory that we would have chosen anyway. Flusser wrote, “the human being can only want what the robot can do,”[21] describing a future in which the machines don’t exactly become more sentient, but that sentience becomes more machinic. Of course, this process is never complete. It doesn’t terminate with the production of a neural network and some predictions. Rather, it enables the extraction of further data and undoubtedly training further AIs, and furthering the regression from meta- to meta- to meta-

[21] Flusser, *The Shape of Things*, 48.