What the Robot Saw

Amy Alexander[1]

[1] University of California, San Diego, USA
ajalexander@ucsd.edu

Abstract *What the Robot Saw* (http://what-the-robot-saw.com) is a continuously generated YouTube livestream, framed as the documentary of a social media robot. The Robot depicts the images of people and scenes it encounters — as its computer vision and AI algorithms obsessively perceive them. Video clips within the documentary are selected from among the least-viewed and least-subscribed videos on YouTube within the past several hours. People framed in closeup in the videos are identified in lower-third text with Amazon Rekognition's analysis of their mood, gender, and age – ironically simplistic consumer identifiers. The robot-edited video is then live streamed nearly continuously, as the Robot's durational performance and a document of the moment. While revealing unsensational content normally only seen by robots, the streamed film's focus is on the processes of self-performance and representation – both human and algorithmic -- in the increasingly blended online and offline culture.

Keywords: AI, YouTube, social media, ranking algorithms, computer vision, machine learning, Amazon Rekognition.

On Social Media Algorithms and [In]visible Selves.

Social media ranking algorithms have recently come under scrutiny for amplifying and encouraging sensational content. YouTube and other media providers have taken steps to address these concerns and attempt to limit some sensational content, like political and medical disinformation. But visibility, through search rankings and recommendation algorithms, is still dependent on engagement metrics: some combination of viewer attention measurements and interaction metrics. So, the business model of promoting attention-grabbing videos continues to encourage sensationalism.

These algorithms can impact public perceptions in various ways — e.g., by amplifying stereotypes: Kate Crawford and Trevor Paglen's *ImageNet Roulette* recently highlighted bias issues resulting from stereotyped labeling in image databases used to train machine learning algorithms (Crawford and Paglen 2019). But algorithmic bias can also determine whether images are seen at all: some types of videos and videomakers — "crowd pleasers" — get more visibility than others, shifting the aggregate of what the public sees. Seasoned "YouTubers" with the knowledge and inclination to strategize their work to maximize algorithmic appeal can increase their visibility. And, as Sophie Bishop points out, algorithms can actively perpetuate stereotypes by rewarding YouTubers for producing demographically stereotypical content (Bishop 2018) -- performing selves for the camera that algorithms favor.

As a result, videos by ordinary people are often seen by few or no human eyes. As with many contemporary human actions, robots may be their main audience: computer vision and artificial intelligence robots analyze social media posts, online videos, or faces of people looking up at ads (Lewis 2019). *What the Robot Saw* (http://what-the-robot-saw.com) is a documentary by such a robot. It sees what humans rarely get to see – but it sees it from the perspective of a robot.
The content is curated using algorithms that run counter to standard commercial ranking algorithms: it includes only videos with low view counts and channel subscriber counts, focusing on content likely to represent personal narratives.

Curation of lesser seen content turns out to be less than straightforward: YouTube's Data API offers developers the option to sort videos based on views from highest to lowest, but not the reverse. This is in some ways understandable: the low end of the view count spectrum includes a number of "troll" and "spam" videos that could benefit from such exposure.[1] Nevertheless, it's worth noting that the API facilitates third-party developers' participation in and amplification of the curatorial biases of YouTube's algorithms (Google Developers 2020).

The real-time cinematography in *What the Robot Saw* is based on the imagined directorial style of the computer vision Robot, as it pans, zooms, greyscales and edge-detects, looking for "features" in images to help it understand and organize the human world. Behind the scenes, computer vision and neural networks are used to eliminate undesirable clips and edit selected clips, then organize the clips into a stream-of-AI-"consciousness" linear structure, focusing on periodic "interviews" with subjects determined to be humans framed as talking heads. Neural network-based audio analysis of the clips sorts files according to prominence of speech vs. music, which facilitates the real-time sound mixing and audio effects.

The film live streams as it is generated, creating a near livestream ouroboros. The film originally live streamed back to YouTube, in an attempt to conceptually return YouTube's unseen videos to YouTube. However, in March 2020, YouTube's COVID-19-era automated moderation algorithms began removing the Robot's new streams, so the film was moved to other online streaming services. The stream presently runs twenty-four hours a day, with short automated "intermissions" every few hours, during which the stream restarts.

Streams are archived and linked from the Robot's Videos page (http://what-the-robot-saw.com/video-samples/). This creates a massive archive that in some ways functions as a collective time capsule of YouTubers' on-camera lives. The collective imagery and narration of the stream evolves as seasons change, holidays are celebrated, and cultural shifts emerge (like the wearing of masks and "lockdown" activities during the COVID-19 outbreak).

Like Crawford and Paglen's *ImageNet Roulette*, *What the Robot Saw* uses a Caffe framework model trained on ImageNet. However, *What the Robot Saw* uses a publicly available pre-trained model that has virtually no categories for people.[2] It therefore classifies images that feature people based on "knowledge" only of objects and animals; it also has accuracy limitations typical of image classification models. I developed an algorithm that attempts to extrapolate the resulting image classifications of a subset of a clips video frames in an attempt to determine a rough classification for each video clip. The resulting somewhat peculiar algorithm creates behind-the-scenes categories, some of which are hinted at in the live stream by text label section identifiers. All categories contribute to the "stream-of-consciousness" sequencing of the live stream.

When the Robot detects talking head videos, it uses Amazon Rekognition, a popular commercial facial recognition service, to estimate age, gender, and mood as displayed in facial expression, then label subjects accordingly: these characteristics are superimposed over their video image where viewers might expect to see an interviewee's name, occupation, and age. The labeling reveals the Robot's inclination to define people in terms of the features Rekognition and similar surveillant services provide — the features business customers presumably seek. As vloggers, job interviewees, students, and others, talk to the camera about whatever is on their minds, the Robot superimposes labels like "Confused-Looking Female, age 23-35.") The absurd juxtaposition of complex human faces and first-person narration with the Robot's inane labels suggests the reductiveness of framing complex people according to characteristics determined useful to marketers. Have our identities as

[1] *What the Robot Saw* attempts to filter "spam" and "troll" videos using custom algorithms. But this is of course difficult, and any algorithmic approach is necessarily imperfect.

[2] The model is available at: https://github.com/torch/tutorials/blob/master/7_imagenet_classification/synset_words.txt

combinations of demographics, facial expressions, and other characteristics come to represent our essential selves more than do our individual identities?

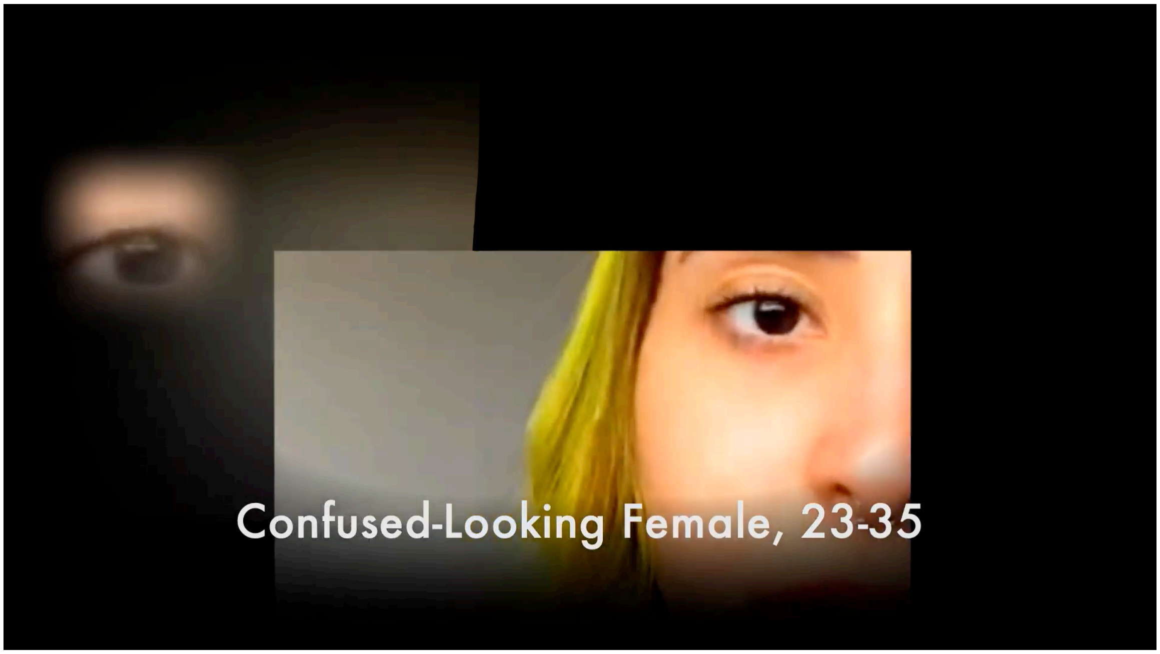

Fig. 1. Screenshot (http://what-the-robot-saw.com/photos)

The Robot's – and Amazon Rekognition's -- interpretations of emotions as one word identifiers like "happy," "calm", "angry", "disgusted," "fearful", etc. both simplify complex emotion and lack the social context with which humans read expression and emotion. YouTubers focused on wide-eyed, upbeat performance may become "confused" or "surprised" in the marketing-centric world of the Robot, while a mischievous smirk, or simply the neutral expression popularly referred to as "resting bitch face,"[3] can become "disgust." The complexities of gender, ethnic, and cultural differences in both general display of emotion and performance for YouTube are invisible to the Robot. (Amazon Rekognition has received criticism for its emotion detection features (Simonite 2019) and its reading of female and darker-skinned faces in general (Singer 2019), as well as for its sales to law enforcement agencies (Harwell (2019).)

But what the robot saw is only part of 'What the Robot Saw.'

Referring to robots and AI's performing human-like activities often elicits concerns about anthropomorphizing them. Perhaps we should actually consider doing more of that, the way we anthropomorphize a ventriloquist dummy while simultaneously understanding that it's only a representation of a human: we understand that the puppeteer is responsible for the dummy's ideas. The Robot's "ideas" are an amalgamation of human ideas, drawn from the particular humans who wrote the algorithms it uses. Some of those humans are developers of popular machine learning algorithms. One of those humans — the puppeteer — is me. But *What the Robot Saw* is not pedagogical. The project's title is a play on the expression "what the butler saw" — an allusion to

[3] https://en.wikipedia.org/wiki/Resting_bitch_face

early peep show films in which a voyeuristic butler spied through a keyhole[4] (Camerani 2009, 115). Although *What the Robot Saw* reveals content normally only seen by robots, it's not about revealing robots' actual perceptions — how they "see" or "think" — any more than *What the Butler Saw* films were about revealing how butlers see. Both the Robot and "the butler" saw something they weren't supposed to see. But they could only peer at the object of their obsession through a keyhole (a metaphorical keyhole, in the Robot's case.) It was an incomplete image: seductive, but just a squinted glimpse of a one-eye peep show. Despite their efforts and presumed satisfaction, neither the peep show butlers nor the Robot could really have a meaningful perception of the people on whom they spied.

So *What the Robot Saw's* title, while literally descriptive, is largely metaphorical. The streamed film itself is a loose allegory for the tangle of processes of representation and perception in the current moment -- of online and offline selves performed and perceived. Luciano Floridi writes of these intertwined selves: "the micro-narratives we are producing and consuming are also changing our social selves and hence how we see ourselves" (Floridi 2014, 62). *What the Robot Saw* is on the one hand about unseen content. But it's more broadly a response to ever-expanding human (and robot) attempts to depict and label ourselves and others according to our respective performed appearances as two-dimensional pixel matrices –- as media-making, neural networks, surveillance, and performance awkwardly collide. The collision happens at the point where the online and offline worlds seem to have collapsed on one another, where borders between our selves and our screen selves, between surveillance, voyeurism, and performance, have become almost incomprehensible. It happens in a culture so accommodated to selves performed as two-dimensional grids of pixels in the shape of talking heads that, when pandemic caused large swaths of the world to abruptly switch to working and socializing as online talking heads, the transition was largely managed within a matter of days. But performed selves should not be dismissed as merely wishful self-caricature: behind online selves are multi-dimensional humans. As *What the Robot Saw* reveals, robots' superficial attempts at comprehension and representation of humans always fall short. But it can be fun watching them try.

References

Bishop, S. 2018**.** Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. Convergence, 24(1), 69–84. https://doi.org/10.1177/1354856517736978

Camerani, M. (2009, October). Joyce and Early Cinema: Peeping Bloom Through the Keyhole. In *Joyce in Progress: Proceedings of the 2008 James Joyce Graduate Conference in Rome* (pp. 114-28). Cambridge Scholars Publishing.

Crawford, Kate and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning. Retrieved 1 May 2020 from https://excavating.ai

Floridi, Luciano. 2014. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press UK.

[4] The expression actually dates back to the late nineteenth century London divorce case of Lord Colin Campbell and Gertrude Elizabeth Blood: their butler testified that he had peered through a keyhole and spied Blood with another man. In the early twentieth century the name was used for both mutoscope "peep show" machines and the moving pictures they played. Since the mid-twentieth century the expression has been used as a title for various films, plays, and television shows, usually with only a figurative connection to the original theme.

Google Developers. 2020. Search. *YouTube Data API.* Retrieved May 31, 2020, from https://developers.google.com/youtube/v3/docs/search

Harwell, Drew. 2019**.** Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong? *The Washington Post.* Retrieved 15 February 2020, from https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/

Lewis, T. 2019**.** AI can read your emotions. Should it? The Guardian. Retrieved 15 February 2020, from https://www.theguardian.com/technology/2019/aug/17/emotion-ai-artificial-intelligence-mood-realeyes-amazon-facebook-emotient

Simonite, T. 2019. Amazon Says It Can Detect Fear on Your Face. You Scared? Wired. Retrieved 15 February 2020, from https://www.wired.com/story/amazon-detect-fear-face-you-scared/

Singer, Natasha. 2019. Amazon Is Pushing Facial Technology That a Study Says Could Be Biased. (2019). *The New York Times*. Retrieved 15 February 2020, from https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html