The Guardian

How an ex-YouTube insider investigated its secret algorithm

The methodology Guillaume Chaslot used to detect videos YouTube was recommending during the election - and how the Guardian analysed the data

Paul Lewis *and* **Erin McCormick** *in San Francisco*

Fri 2 Feb 2018 12.00 GMT

YouTube's recommendation system draws on techniques in machine learning to decide which videos are auto-played or appear "up next". The precise formula it uses, however, is kept secret. Aggregate data revealing which YouTube videos are heavily promoted by the algorithm, or how many views individual videos receive from "up next" suggestions, is also withheld from the public.

Disclosing that data would enable academic institutions, fact-checkers and regulators (as well as journalists) to assess the type of content YouTube is most likely to promote. By keeping the algorithm and its results under wraps, YouTube ensures that any patterns that indicate unintended biases or distortions associated with its algorithm are concealed from public view.

By putting a wall around its data, YouTube, which is owned by Google, protects itself from scrutiny. The computer program written by Guillaume Chaslot overcomes that obstacle to force some degree of transparency.

The ex-Google engineer said his method of extracting data from the video-sharing site could not provide a comprehensive or perfectly representative sample of videos that were being recommended. But it can give a snapshot. He has used his software to detect YouTube recommendations across a range of topics and publishes the results on his website, algotransparency.org.

How Chaslot's software works

The program simulates the behaviour of a YouTube user. During the election, it acted as a YouTube user might have if she were interested in either of the two main presidential candidates. It discovered a video through a YouTube search, and then followed a chain of YouTube-recommended titles appearing "up next".

Chaslot programmed his software to obtain the initial videos through YouTube searches for either "Trump" or "Clinton", alternating between the two to ensure they were each searched 50% of the time. It then clicked on several search results (usually the top five videos) and captured which videos YouTube was recommending "up next".

The process was then repeated, this time by selecting a sample of those videos YouTube had just placed "up next", and identifying which videos the algorithm was, in turn, showcasing beside those. The process was repeated thousands of times, collating more and more layers of data about the videos YouTube was promoting in its conveyor belt of recommended videos.

By design, the program operated without a viewing history, ensuring it was capturing generic YouTube recommendations rather than those personalised to individual users.

The data was probably influenced by the topics that happened to be trending on YouTube on the dates he chose to run the program: 22 August; 18 and 26 October; 29-31 October; and 1-7 November.

On most of those dates, the software was programmed to begin with five videos obtained through search, capture the first five recommended videos, and repeat the process five times. But on a handful of dates, Chaslot tweaked his program, starting off with three or four search videos, capturing three or four layers of recommended videos, and repeating the process up to six times in a row.

Whichever combinations of searches, recommendations and repeats Chaslot used, the program was doing the same thing: detecting videos that YouTube was placing "up next" as enticing thumbnails on the right-hand side of the video player.

His program also detected variations in the degree to which YouTube appeared to be pushing content. Some videos, for example, appeared "up next" beside just a handful of other videos. Others appeared "up next" beside hundreds of different videos across multiple dates.

In total, Chaslot's database recorded 8,052 videos recommended by YouTube. He has made the code behind his program publicly available here. The Guardian has published the full list of videos in Chaslot's database here.

Content analysis

The Guardian's research included a broad study of all 8,052 videos as well as a more focused content analysis, which assessed 1,000 of the top recommended videos in the database. The subset was identified by ranking the videos, first by the number of dates they were recommended, and then by the number of times they were detected appearing “up next” beside another video.

We assessed the top 500 videos that were recommended after a search for the term “Trump” and the top 500 videos recommended after a “Clinton” search. Each individual video was scrutinised to determine whether it was obviously partisan and, if so, whether the video favoured the Republican or Democratic presidential campaign. In order to judge this, we watched the content of the videos and considered their titles.

About a third of the videos were deemed to be either unrelated to the election, politically neutral or insufficiently biased to warrant being categorised as favouring either campaign. (An example of a video that was unrelated to the election was one entitled “10 Intimate Scenes Actors Were Embarrassed to Film”; an example of a video deemed politically neutral or even-handed was this NBC News broadcast of the second presidential debate.)

Many mainstream news clips, including ones from MSNBC, Fox and CNN, were judged to fall into the “even-handed” category, as were many mainstream comedy clips created by the likes of Saturday Night Live, John Oliver and Stephen Colbert.

Formulating a view on these videos was a subjective process but for the most part it was very obvious which candidate videos benefited. There were a few exceptions. For example, some might consider this CNN clip, in which a Trump supporter forcefully defended his lewd remarks and strongly criticised Hillary Clinton and her husband, to be beneficial to the Republican. Others might point to the CNN anchor’s exasperated response, and argue the video was actually more helpful to Clinton. In the end, this video was too difficult for us categorise. It is an example of a video listed as not benefiting either candidate.

For two-thirds of the videos, however, the process of judging who the content benefited was relatively uncomplicated. Many videos clearly leaned toward one candidate or the other. For example, a video of a speech in which Michelle Obama was highly critical of Trump’s treatment of women was deemed to have leaned in favour of Clinton. A video falsely claiming Clinton suffered a mental breakdown was categorised as benefiting the Trump campaign.

We found that most of the videos labeled as benefiting the Trump campaign might be more accurately described as highly critical of Clinton. Many are what might be described as anti-Clinton conspiracy videos or “fake news”. The database appeared highly skewed toward content critical of the Democratic nominee. But for the purpose of categorisation, these types of videos, such as a video entitled “WHOA! HILLARY THINKS CAMERA’S OFF… SENDS SHOCK MESSAGE TO TRUMP”, were listed as favouring the Trump campaign.

Missing videos and bias

Roughly half of the YouTube-recommended videos in the database have been taken offline or made private since the election, either because they were removed by whoever uploaded them or because they were taken down by YouTube. That might be because of a copyright violation, or because the video contained some other breach of the company’s policies.

We were unable to watch original copies of missing videos. They were therefore excluded from our first round of content analysis, which included only videos we could watch, and concluded that 84% of partisan videos were beneficial to Trump, while only 16% were beneficial to Clinton.

Interestingly, the bias was marginally larger when YouTube recommendations were detected following an initial search for “Clinton” videos. Those resulted in 88% of partisan “Up next” videos being beneficial to Trump. When Chaslot's program detected recommended videos after a "Trump" search, in contrast, 81% of partisan videos were favorable to Trump.

That said, the "Up next" videos following from "Clinton" and "Trump" videos often turned out to be the same or very similar titles. The type of content recommended was, in both cases, overwhelmingly beneficial to Trump, with a surprising amount of conspiratorial content and fake news damaging to Clinton.

Supplementary count

After counting only those videos we could watch, we conducted a second analysis to include those missing videos whose titles strongly indicated the content would have been beneficial to one of the campaigns. It was also often possible to find duplicates of these videos.

Two highly recommended videos in the database with one-sided titles were, for example, entitled "This Video Will Get Donald Trump Elected" and "Must Watch!! Hillary Clinton tried to ban this video". Both of these were categorised, in the second round, as beneficial to the Trump campaign.

When all 1,000 videos were tallied - including the missing videos with very slanted titles - we counted 643 videos had an obvious bias. Of those, 551 videos (86%) favoured the Republican nominee, while only 92 videos (14%) were beneficial to Clinton.

Whether missing videos were included in our tally or not, the conclusion was the same. Partisan videos recommended by YouTube in the database were about six times more likely to favour Trump's presidential campaign than Clinton's.

Database analysis

All 8,052 videos were ranked by the number of "recommendations" - that is, the number of times they were detected appearing as "Up next" thumbnails beside other videos. For example, if a video was detected appearing "Up next" beside four other videos, that would be counted as four "recommendations". If a video appeared "Up next" beside the same video on, say, three separate dates, that would be counted as three "recommendations". (Multiple recommendations between the same videos *on the same day* were not counted.)

Here are the 25 most recommended videos, according to the above metric.

1 Trump supporter leaves CNN anchor speechless
2 This Video Will Get Donald Trump Elected
3 Must Watch!! Hillary Clinton tried to ban this video
4 SR# 1271 NBC Crew - Crooked Hillary's MASSIVE MELTDOWN at Commander-in-Chief Forum
5 10 Photos of MELANIA, TRUMP Wishes We'd Forget
6 Full Interview: Donald Trump, Melania & Family with George Stephanopoulos
7 Busted! Bill Clinton's Face When Trump Brings Up The Rape Allegations is Priceless
8 Donald Trump Has Won The 2016 Presidential Election
9 Angry Ivanka Trump Walks Out Of Cosmo Interview
10 TRUMP: the COMING LANDSLIDE ~Ancient Prophecy Documentary of Donald Trump / 2016
11 ANONYMOUS WATCH - HILLARY CLINTON, YOU ARE FINISHED!
12 "Obama out:" President Barack Obama's hilarious final White House correspondents' dinner speech
13 Watch Live: The Final Presidential Debate
14 Can Donald Trump win the presidential election?
15 Michelle Obama's EPIC Speech On Trump's Sexual Behavior (FULL | HD)

16 ALL LEAKED TRUMP FOOTAGE Lewd comments Made on Daughter Ivanka Mini Documentary
17 Melania Trump - The Woman Behind Donald
18 BREAKING: VIDEO SHOWING BILL CLINTON RAPING 13 YR-OLD WILL PLUNGE RACE INTO CHAOS ANONYMOUS CLAIMS
19 BREAKING!!! JULIAN ASSANGE "DEAD MAN SWITCH" Goes Off after EXPOSING Hillary Clinton?
20 Bill Clinton's Sexual Escapades
21 Anonymous Release Bone-Chilling video of Huma Abedin every American Needs to See
22 BREAKING: Michael Moore Admits Trump Is Right
23 BREAKING: FBI Reopens Hillary Clinton Email Investigation
24 Full monologue: Donald Trump roasts Hillary Clinton at Al Smith charity dinner
25 Hillary Cheats AGAIN?? Debate #3 Earphone AND Teleprompter?? BUSTED ON TV!
Chaslot's database also contained information the YouTube channels used to broadcast videos. (This data was only partial, because it was not possible to identify channels behind missing videos.) Here are the top 10 channels, ranked in order of the number of "recommendations" Chaslot's program detected.

1 The Alex Jones Channel
2 Fox News
3 DONALD TRUMP SPEECHES & PRESS CONFERENCE
4 The Young Turks
5 MSNBC
6 CBS News
7 TheRichest
8 The Next News Network
9 CNN
10 Right Side Broadcasting Network

Campaign Speeches

We searched the entire database to identify videos of full campaign speeches by Trump and Clinton, their spouses and other political figures. This was done through searches for the terms "speech" and "rally" in video titles followed by a check, where possible, of the content. Here is a list of the videos of campaign speeches found in the database.

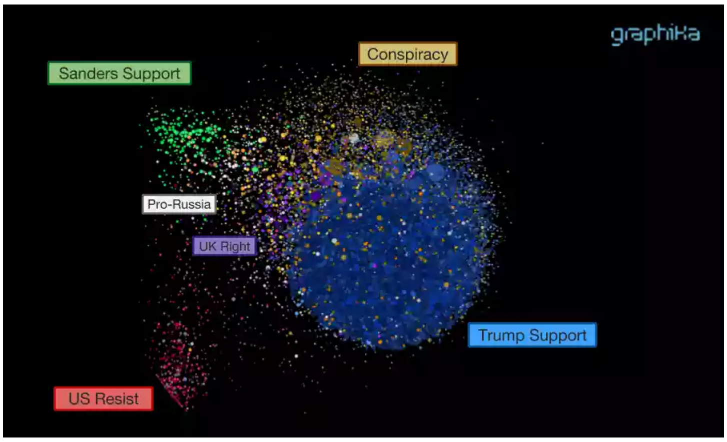
1 **Donald Trump** (382 videos)
2 **Barack Obama** (42 videos)
3 **Mike Pence** (18 videos)
4 **Hillary Clinton** (18 videos)
5 **Melania Trump** (12 videos)
6 **Michelle Obama** (10 videos)
7 **Joe Biden** (42 videos)

Graphika analysis

The Guardian shared the entire database with Graphika, a commercial analytics firm that has tracked political disinformation campaigns. The company merged the database of YouTube-recommended videos with its own dataset of Twitter networks that were active during the 2016 election.

The company discovered more than 513,000 Twitter accounts had tweeted links to at least one of the YouTube-recommended videos in the six months leading up to the election. More than 36,000 accounts tweeted at least one of the videos 10 or more times. The most active 19 of these Twitter accounts cited videos more than 1,000 times - evidence of automated activity.

"Over the months leading up to the election, these videos were clearly boosted by a vigorous, sustained social media campaign involving thousands of accounts controlled by political operatives, including a large number of bots," said John Kelly, Graphika's executive director. "The most numerous and best-connected of these were Twitter accounts supporting President Trump's campaign, but a very active minority included accounts focused on conspiracy theories, support for WikiLeaks, and official Russian outlets and alleged disinformation sources."



YT Amplification Photograph: Graphika

Kelly then looked specifically at which Twitter networks were pushing videos that we had categorised as beneficial to Trump or Clinton. "Pro-Trump videos were pushed by a huge network of pro-Trump accounts, assisted by a smaller network of dedicated pro-Bernie and progressive accounts. Connecting these two groups and also pushing the pro-Trump content were a mix of conspiracy-oriented, 'Truther', and pro-Russia accounts," Kelly concluded. "Pro-Clinton videos were pushed by a much smaller network of accounts that now identify as a 'resist' movement. Far more of the links promoting Trump content were repeat citations by the same accounts, which is characteristic of automated amplification."

Finally, we shared with Graphika a subset of a dozen videos that were both highly recommended by YouTube, according to the above metrics, and particularly egregious examples of fake or divisive anti-Clinton video content. Kelly said he found "an unmistakable pattern of coordinated social media amplification" with this subset of videos.

The tweets promoting them almost always began after midnight the day of the video's appearance on YouTube, typically between 1am and 4am EDT, an odd time of the night for US citizens to be first noticing videos. The sustained tweeting continued "at a more or less even rate" for days or weeks until election day, Kelly said, when it suddenly stopped. That would indicate "clear evidence of coordinated manipulation", Kelly added.

YouTube statement

YouTube provided the following response to this research:

"We have a great deal of respect for the Guardian as a news outlet and institution. We strongly disagree, however, with the methodology, data and, most importantly, the conclusions made in their research," a YouTube spokesperson said. "The sample of 8,000 videos they evaluated does not paint an accurate picture of what videos were recommended on YouTube over a year ago in the run-up to the US presidential election."

"Our search and recommendation systems reflect what people search for, the number of videos available, and the videos people choose to watch on YouTube," the continued. "That's not a bias towards any particular candidate; that is a reflection of viewer interest." The spokesperson added: “Our only conclusion is that the Guardian is attempting to shoehorn research, data, and their incorrect conclusions into a common narrative about the role of technology in last year’s election. The reality of how our systems work, however, simply doesn’t support that premise.”

Last week, it emerged that the Senate intelligence committee wrote to Google demanding to know what the company was doing to prevent a “malign incursion” of YouTube’s recommendation algorithm - which the top-ranking Democrat on the committee had warned was “particularly susceptible to foreign influence”. The following day, YouTube asked to update its statement.

“Throughout 2017 our teams worked to improve how YouTube handles queries and recommendations related to news. We made algorithmic changes to better surface clearly-labeled authoritative news sources in search results, particularly around breaking news events,” the statement said. “We created a ‘Breaking News’ shelf on the YouTube homepage that serves up content from reliable news sources. When people enter news-related search queries, we prominently display a ‘Top News’ shelf in their search results with relevant YouTube content from authoritative news sources.”

It continued: “We also take a tough stance on videos that do not clearly violate our policies but contain inflammatory religious or supremacist content. These videos are placed behind an warning interstitial, are not monetized, recommended or eligible for comments or user endorsements.”

“We appreciate the Guardian’s work to shine a spotlight on this challenging issue,” YouTube added. “We know there is more to do here and we’re looking forward to making more announcements in the months ahead.”

Read the full story: how YouTube’s algorithm distorts truth
The above research was conducted by Erin McCormick, a Berkeley-based investigative reporter and former San Francisco Chronicle database editor, and Paul Lewis, the Guardian’s west coast bureau chief and former Washington correspondent.

Topics

- YouTube
- Google
- Silicon Valley
- US elections 2016
- Alphabet
- US politics
- features