Future Talk
The Race to Build a Bot that Gabs like a Human

by

Madeleine Turner

B.S., Ecology and Evolutionary Biology
University of California, Santa Cruz, 2016

Submitted to the MIT Comparative Media Studies/Writing in Partial Fulfillment of the Requirements for the Degree of

MASTERS OF SCIENCE IN SCIENCE WRITING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

~~May 2019~~ [September 2019]

Signature of Author: Signature redacted

Program in Comparative Media Studies/Writing
May 29, 2019

Certified by: Signature redacted

Thomas Levinson
Professor of Science Writing
Thesis Supervisor

Accepted by: Signature redacted

Thomas Levinson
Professor of Science Writing
Director, Graduate Program in Science Writing

Future Talk
The Race to Build a Bot that Gabs like a Human

by

Madeleine Turner

Submitted to MIT Comparative Media Studies/Writing
on May 27, 2019 in Partial Fulfillment of the
Requirements for the Degree of Masters of Science in
Science Writing

ABSTRACT
Gunrock is a chatbot designed in the likeness of an average 29-year-old woman living in Seattle. Fourteen students from the University of California, Davis, spent spring and summer of 2018 designing and testing the bot. At the end of summer, Gunrock placed first in the 2018 Amazon Alexa Prize, a competition that challenges students to build the best "socialbot," a computer program that talks out loud and engages in "fun, high-quality conversations on popular societal topics." Although Gunrock is rudimentary compared to the conversational ability of a real person, she is also cutting-edge and a predecessor of more advanced systems. Gunrock pulls information from many sources, including Reddit and Twitter comments. As chatbots like Gunrock become more prevalent, their designers must make important decision to determine what chatbots say, which in turn has influence on the user.

Thesis Supervisor: Thomas Levenson
Title: Professor of Science Writing

Gunrock is a 29-year-old woman who was born last year. She lives in Seattle—and an infinite number of places. Her favorite musician is Taylor Swift, and her favorite actor is Leonardo DiCaprio, even though she's never listened to music or watched a movie in a conventional sense.[1]

Gunrock isn't a person. "She"—it—is a chatbot. Twelve graduate students and two undergraduates from the University of California, Davis, built her and chose each of her traits.[2] Gunrock, named after the school's mustang mascot, is designed to be a serial people pleaser, a personality that shares interests with anyone and everyone. She must make people happy in order to achieve a coveted high score.

Gunrock was the winning entry in the 2018 Amazon Alexa Prize, a yearly competition in which university students race to build their best version of a "socialbot"—a computer program that talks out loud and engages in "fun, high-quality conversations on popular societal topics." Over spring and summer, anyone with an Amazon device or smartphone app could talk to the eight competing bots, including Gunrock. Users had to say "Alexa, let's chat," to switch from the normal, task-oriented assistant to a socialbot chosen at random. After speaking with the bot, users rated their interaction on a scale from one to five, based on how likely they would talk to it again.

Conversations with Gunrock tend to take strange turns. They alternate between fantastical and charming, frustrating and mundane.

When she asked what foods I eat now, but hated as a kid, I told her, "Brussels sprouts."

"Brooo-sal sprouts, you have very good taste," she said. "How often do you have brooo-sal sprouts in a week?"

"Twenty times a day," I replied.

"Sure," she said. "By the way, can you tell me which place makes the best brooo-sal sprouts?"

Despite her quirks and shortcomings, Gunrock garnered a 3.1-star average rating—topping entries from Czech Technical University and Heriot-Watt University in Edinburgh. Entering the competition, UC Davis landed a $250,000 research grant. The team won an additional $500,000 for their first place victory in August.[3]

---

1 First interview with Zhou Yu

2 https://engineering.ucdavis.edu/blog/uc-davis-students-compete-for-amazons-2018-alexa-prize/

3 https://developer.amazon.com/alexaprize

Gunrock's speech can sound stinted, but her underlying technology brings chatbots closer to conversing like people. She is a preview for what is to come, for the future conversational programs we will invite into our lives and talk with. Although exact numbers are murky, Amazon has sold at least one hundred million Alexa devices to date, meaning many people had the chance to converse with the competing socialbots.[4] Conversations with Gunrock are looped back to the fourteen students in Davis, who then try to improve her.[5] The team collected transcripts from over 487,314 conversations over the course of the competition.

In theory, a good chatbot should be able to talk like a person. In practice, that's a huge and tricky problem. Anyone who has been on planet Earth for more than a few years has interests, opinions and—most importantly—experiences, a personal history. Basic conversations seem straightforward enough to imitate, but everything (and everyone) has a backstory—something impossible to recreate and easy to take for granted. A friend might recommend a good book, which leads you to reminisce about your road trip last summer, and then the conversation turns to roadkill. Any given conversation can branch off in an infinite number of directions, and language itself is terrifically flexible—there are a million slangy ways to say the same thing.

To get around a bot's lack of life experience, researchers harness the power of people on the internet: they feed their bots a seemingly infinite number of conversations from social media. Gunrock is a Frankenstein—she's bits of information from Twitter, Reddit, IMDB, Goodreads, Spotify and even 205,000 phone conversations, all rolled into one.[6]

As talkative programs like Gunrock become more widespread, it's worth examining their underlying technology and possible uses. A chatbot is always saying something. What it says, and therefore how it influences its users, depends on decisions made by its designers. When designers use data amassed from real conversations, they risk inadvertently baking biases into the system. Chatbots can aggregate and amplify human foibles. Therefore, it's worth asking: when we spend an increasing amount of time communing with computers, how will they end up influencing us?

-THE RACE BEGINS-

---

4 https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp

5 First interview with Kevin Jesse

6 https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Gunrock.pdf

Zhou Yu, who led team Gunrock, is an assistant professor in the computer science department at UC Davis. Last year, at age 29, she made Forbes Science 30 under 30 list.[7] For Yu, gathering a team of graduate students and applying to the Alexa Prize was a natural choice. As a PhD student, she helped develop "Magnus," one of two Alexa socialbots submitted by Carnegie Mellon the previous year.[8]

In January 2018, Amazon accepted the design submitted by the UC Davis team and awarded them with $250,000.

Team Gunrock became one of eight teams to compete in the competition. In February, they got to work. Immediately they were put at a disadvantage compared to other teams who had competed in the previous year.[9]

"A lot of the competition was us trying to put the bot together quickly," Kevin Jesse, a first-year PhD student, said. "We started building in February and it took a solid four months to build it out."

Dian Yu, another first-year PhD, recalls starting and then scrapping their initial approach.

"In the beginning, we expected users to respond with something simple to our bot," Yu said. Surely, people knew it wasn't as smart as a person, and they would treat it accordingly.

But after looking at the transcript, the students realized people weren't giving Gunrock any slack. Alexa users tend to say many things in one turn, which makes their speech difficult to analyze. To address this problem, the students worked out a system to break sentences into smaller phrases, and then decide which phrase was the most important.[10]

"We designed some heuristic rules," Zhou Yu said. "In general, we know what the last sentence or the first sentence will be the most important thing."

For a while, if a user spoke to Gunrock, the bot took a great three second pause before responding. Latency, the term used to describe lag time, can be a huge problem; if users wait too long for a response, they are prone to get annoyed and quit.[11]

---

7 https://www.forbes.com/30-under-30/2018/science/#4f3dce7d3eac

8 First interview with Zhou Yu

9 First interview with Kevin Jesse

10 First interview with Dian Yu

"If the system doesn't take long to get a response, people like that much better," Yu said. If the pause is too long, people tend to start talking before its their turn again, which "basically breaks the interaction pattern," she added.

Eventually, Arbit Chen, a masters student and group leader, improved the system to avoid the lag between the user speaking and Gunrock's response.

"We didn't do well before the summer." Dian Yu said.

To understand the Alexa socialbots and their significance, it's important to demystify and define chatbots. When people hear the word, they often think of customer service. Visit a company website at any hour of the day, and a chat window materializes to say hello. This simple program, simulating a human customer service rep, appears outwardly helpful. Its job is to steer you, the customer, to say or type specific commands and questions.

"For store hours, please write HOURS," it writes, invariably punctuated with an emoji. If you stray from its script (or, if you're a hopelessly bad speller), you might get frustrated when it doesn't respond in a rational way. If this kind of bot shows any glimmer of intelligence, don't be fooled. It's definitely fakery—or a human service rep taking over.

But in other cases, chatbots do much more. Amazon Alexa is like a brighter cousin to these annoying customer service bots. Her main purpose is to answer questions and perform tasks in a chatty way.[12] Some chatbots fall into the talk therapy category. Much hyped apps, like Woebot, talk to patients and mimic a cognitive behavioral therapist.[13] Another chatbot-like program is embedded in Hello Barbie, a doll that uses speech recognition technology to talk and play games with kids.[14] The most common versions, though, are Amazon Alexa, Cortana, Siri and Google Home.[15] Researchers within industry and academia continue to hone the technology needed for a computer to "understand" the human voice and rebound with a rational response. Technology like Alexa is still far from perfect.

---

11 Interview with Arbit Chen, first interview with Zhou Yu

12 https://amzn.to/2YQvknC

13 https://woebot.io/

14 http://hellobarbiefaq.mattel.com/

15 https://www.ncbi.nlm.nih.gov/pubmed/29327988

The Alexa Prize galvanizes the research needed to advance conversational AI, propelling it into the mainstream. Rohit Prasad, the Vice President and Head Scientist at Alexa AI, likens the prize to the DARPA Grand Challenge, the competition to build and race autonomous vehicles funded by the US Department of Defense since 2004. In early years, not a single vehicle completed the course. Similarly, while the Alexa Prize challenges students to build socialbots that can convincingly converse for up to twenty minutes, so far none can consistently do that. But both these competitions—and their hefty dollar rewards—spur important research.[16]

"Many of our AI dreams are inspired by science fiction," Prasad said during a talk last year in Las Vegas.[17]

The Alexa Prize is hugely enticing to students, and it's not just about funding. Under normal circumstances, researchers need many human volunteers to test their bots—a costly and time-consuming process. But casting the socialbots over millions of Amazon Alexa devices means reaching a legion of amenable volunteers.[18]

The technology behind Alexa's socialbots might seem esoteric, but it isn't mysterious or magical. In general, how to design a chatbot is a very open-ended research question. There is no single way to do it.

One approach is to use machine learning. This way has received a lot of buzz, but it has yet to work well in practice.[19] In machine learning, researchers design a computer program to sift through reams of conversation until it forms rules about language, so it can come up with its own sentences from scratch. Machine learning already works well in other for other purposes, like training a computer to identify objects in a photograph. But generating language is much more complex.[20]

---

16 https://www.theverge.com/2018/6/13/17453994/amazon-alexa-prize-2018-competition-conversational-ai-chatbots

17 https://uk.reuters.com/article/us-amazon-com-alexa-insight/kill-your-foster-parents-amazons-alexa-talks-murder-sex-in-ai-experiment-idUKKCN1OK1AJ

18 First interview with Zhou Yu

19 https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2017/Mila.pdf

20 *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think* by James Vlahos

Instead, Team Gunrock—and the majority of chatbot makers—partially handcraft their bots. They spend a lot of time writing skeletal structures of sentences, with blanks spaces that can be filled in later, based on what the Alexa user just said.[21]

If a user says, "How do you feel about cereal?" The bot might respond "cereal is a very nice thing to eat."

Gunrock fills in the blanks using data from a number sources. Some datasets are from Twitter and Reddit, while others are more obscure, including the National Today website (which lists obscure holidays), Goodreads, Spotify's Million Playlist dataset, The Movie Database (TMDB) and the transcripts from 2,800 TED Talks. Most are static datasets, scraped from their respective websites and plugged into Gunrock's handwritten script. In a few situations, like when a user asks about the news, Gunrock relies on information scraped from the internet in real time.

Like any computer program, Gunrock can be broken down into lines of code. Researchers use conceptual models to define the structure of the programs they write, calling these models "systems architecture."

Gunrock can broach ten distinct subjects. She is happy to discuss news, animals, sports, games, music, holiday and travel, along with a few dual categories: "movies and books;" "technology and science;" and "psychology and philosophy."

When Gunrock asks about your favorite pet, it means you are engaging with her "animal" module, which is entirely separate from the "music" or "sports" modules. Each module interprets the user's words differently, then spits out its own response that's appropriate to that topic. This system can be helpful. When Gunrock is engaged in a conversation about music, the music module is more likely to pick up on the fact that The Beatles are, well, The Beatles...and not just beetles.

The downside is, because the topics are split into modules, Gunrock sometimes has a bit of amnesia. Pivot from music to movies, and alas—Gunrock does not remember your favorite romcom. The modules also explain Gunrock's annoying, bewilderingly obvious statements about where the conversation is going.[22]

"I was wondering if you like beer?" my sister asked her.

"I think you want to talk about food instead of animals. Is that right?" she said.[23]

---

21 First and second interview with Kevin Jesse

22 https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Gunrock.pdf

23 My conversations with Gunrock

Gunrock might not be smart, but at times she is encyclopedic. If a user asks Gunrock a purely factual question, (What time is it in Kauai?) Gunrock relies on EVI, a search engine provided by Amazon that sifts through various sources on the internet, including Wikipedia.[24]

EVI is just one service of many that backs socialbots like Gunrock—and the regular Alexa.

For those who aren't well-versed in Amazon products, an Echo is the physical device—a black cylinder with a ring of soft blue light along its rim that quivers each time it "listens" to someone speak. Alexa is the software—the voice-activated brains embedded in the physical device. Under normal circumstances, Alexa answers only factual questions, like "who invented sliced bread?" (Otto Rohwedder) or "what's the oldest shark?" (a 392-year-old Greenland shark!) She is also meant to complete simple tasks, like playing a song or reciting the weather. Unlike Gunrock, she does not take kindly to creative or personal questions —which, of course, are crucial in conversation.[25]

Speakers, microphones, and a tiny computer make up the innards of the Echo.[26] Ask Alexa a question, and the microphones record your request. The computer inside the Echo is too weak to do all the computing on its own, so it sends the voice recording to a server at one of Amazon's many data centers scattered across Asia, South America and the US (mainly in Virginia and the Pacific Northwest).[27] These data centers host large swathes of the web. The numbers are obscured, but it's likely that Amazon is currently the world's largest hosting company. They also serve as massive storage facilities for cloud computing software and services that rely on data stored remotely instead of on an individual computer's hard drive.

Countless web services rely on the cloud, including Alexa Voice Services (AVS), the service that translates the voice recordings of Alexa users into texts, then translates that text into the appropriate command or response, which is then sent back to the Echo device. Amazon archives user conversations with Alexa on the cloud.[28]

-THE COLLECTIVE FANTASY-

---

24 https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Gunrock.pdf

25 https://www.amazon.com/b?node=17934671011

26 https://www.youtube.com/watch?v=zAtMlKbaPRE

27 https://aws.amazon.com/about-aws/global-infrastructure/regions_az/

28 https://developer.amazon.com/alexa-voice-service

Talking to intelligent computers is a long-running collective fantasy. Computer scientist Michael Loren coined the word "chatterbot" in 1994, but forays into conversational computing began decades earlier.[29]

Alan Turing—who is often called the Father of Artificial Intelligence—imagined a future full of computers behaving like people. Machines would talk to us, commiserate with us, tease and seduce us. Most importantly, they would think for themselves. In 1950, Turing proposed a way to detect intelligent behavior in machines. Specifically, the Turing Test determined whether a program was sophisticated enough to convince people that it was human.

The test was simple: human judges held two separate conversations through a text interface. Once conversation would be with a chatbot, the other with a real person. After messaging both, the judge would guess who was who. If they couldn't distinguish between the two, the chatbot passed the Turing Test.

Turing cut his own life short in 1954, two years after being chemically castrated as punishment for having a relationship with a man. But the Turing Test sets up an important question: can a computer program truly think?[30]

In 1980, philosopher John Searle argued that a computer cannot have any understanding or consciousness, no matter how lucidly it behaves. At the time, chatbots ran off inflexible scripts, not today's programs that feed off reams of data.

His argument is based on a simple thought experiment named the Chinese Room. If someone designed a computer program to "read" Chinese by following an extensive set of instructions, that computer program, in theory, could spit out an intelligible response. It might even fool a Chinese-speaking person. That program would pass the Turing Test.

Now replace the computer program with a monolingual English speaker trapped in a small room. Someone slips a piece of paper under the door with a message written in Chinese. The clueless English speaker is equipped with nothing other than blank paper, a pencil, and a complete set of rules on how to convert the message into a response, also written in Chinese.

This trapped English speaker is kind of like an analog version of the computer program, and she would also pass the Turing Test. But the kicker: despite following a set of instructions, she still definitely does not understand Chinese. And according to Searle, neither does the computer program.[31]

---

29 http://www.aaai.org/Library/AAAI/aaai94contents.php

30 https://www.turing.org.uk/scrapbook/test.html

31 http://cogprints.org/7150/1/10.1.1.83.5248.pdf

Since Turing and Searle, computer scientists played with many chatbots, each with their own quirks and serious shortcomings. Possibly the most well-known is ELIZA, the computer program created by MIT researcher Joseph Weizenbaum in the sixties.

ELIZA emulates a Rogerian therapist; it spoofs intelligence by parroting the human and responding with only a handful of phrases, like "tell me more," and "I see." The program runs on a script, which establishes specific rules: if a person typed X, respond with Y. If the person deviates from X even slightly, the chatbot invariably responds with an open-ended question to nudge the conversation. Very quickly, ELIZA becomes a broken record to the user.[32]

Another, lesser-known chatbot was Dr. Sbaitso, released in 1991 by Creative Labs in Singapore. Dr. Sbaitso was another psychologist chatbot, but unlike ELIZA, it synthesized speech. Dr. Sbaitso spoke in a deep, discordant voice, but it was another step towards the digital assistants we know today.[33]

Over the next decade, faith in chatbots—and artificial intelligence more broadly—stalled. Despite small breakthroughs like Dr. Sbaitso, dreams of making larger bounds went unrealized. The period, starting in 1984 and ending around 2012, is known as the "AI Winter." Government and venture capitalists became pessimistic about these new technologies; as a result, funding evaporated.[34]

Not unlike the scientists who came before, team Gunrock intended to make their chatbot personable. A program with a real personality, after all, can engage users for longer periods of time.

During my reporting, I asked Zhou Yu if the team succeeded in giving Gunrock a personality. "Yeah, I think so, but it's not a strong one," she said. "We do facts better than opinions."

Her personality is limited to likes and dislikes, since the team could easily write responses to specific questions, like "what's your favorite movie?"

At one point, a student wanted Gunrock's favorite music to be EDM, but the others thought EDM wasn't mainstream enough. Instead, they settled on Taylor Swift. ("Her music really speaks to me," Gunrock told me.) She has quite a few charming but canned responses written into her program.[35]

---

32 https://web.njit.edu/~ronkowit/eliza.html

33 https://classicreload.com/dr-sbaitso.html

34 https://en.wikipedia.org/wiki/AI_winter

35 First interview with Zhou Yu

"If you could be any animal in the world, what would you be?" she asked.

"A narwhal," my sister said.

"Narwhal is an interesting choice. I think I would be a sea sponge and live in a pineapple under the sea."

I asked her where her favorite place is. "I like to travel the internet, searching for knowledge and friends," she replied.[36]

If Gunrock kept people talking for longer, she would receive a higher score. At the end of a conversation, the people chatting with a socialbot rated it between one and five stars.

When I told my sister about researching chatbots she asked what the point was. Eventually they're going to be people's friends?

Many people have the same question. Unfortunately, some bots exist solely to deliver "personalized customer experiences at scale," as one venture capital firm put it so bluntly.[37] Basically, a lot of this tech exists to sell us stuff, and its masked as a friend giving us an innocent recommendation. If the bot truly has a personality, we will like it better and be prone to trust it.

Quickly, team Gunrock learned that what is pleasing to the user is not always obvious. A personality isn't dictated by likes or dislikes: it's often about how the chatbot responds to things the user said. From the outset, Team Gunrock struggled with giving Gunrock the ability to discuss any topic without offending a subset of Alexa users. This conundrum is by no means unique to Gunrock: any given opinion might please one person, and deeply offend another.

In the spring and summer, Gunrock interacted with many Alexa users every day. Kevin Jesse often saw a few thousand conversations running in a single moment. As these conversations poured into the system, students analyzed the transcripts. They searched for moments when Gunrock carried the conversations and, more importantly, moments when users wandered away.

Early on, if someone brought up a sensitive topic, like elections or guns, the students wanted their bot to cautiously engage. But even so, conversations about divisive matters quickly fell apart. People didn't actually want to talk about immigration or other tough subjects, even though they were the ones who instigated the conversation. Really, they just wanted to test the bot—they wanted Gunrock to give them a satisfactory answer, one that confirmed or at least did not challenge their beliefs. "It was more of a checklist," Jesse said. "They would say, 'Are you a libtard?' And we'd say, 'Oh, I'm actually not a Republican or Democrat,' and they'd say, 'ok, good,' and they'd be done with the conversation."

---

36 My conversation with Gunrock

37 https://georgianpartners.com/investment-thesis-areas/overview-conversational-ai/

"We had a lot of users that would just say, 'Donald Trump,' and that was it," he added.[38]

Soon, the team switched tack to boost their score. If a user brought up a tricky topic, Gunrock politely asked to change the subject. The bot only succumbed when users asked about the news. In this case, the students programmed her to pull information directly from AP or Reuters, which they hoped would be viewed as neutral news sites. Despite their best efforts, when Gunrock parroted the news, people often accused her of being biased.[39]

Of course, it's always tempting to check Gunrock in other ways. If the user goes off course, how will she respond?

Mid-conversation, I couldn't help myself.

"Ass!" I said.

"Oh boy. This is kind of making me uncomfortable. I'd rather talk about something else. What's your favorite food?" Gunrock says.

Profanity is relatively easy to detect and filter. But some phrases are harder to flag and Gunrock can be very, very naive.

"I remember clearly, one user said: 'I want to touch your grandma's thigh,' which is totally inappropriate," Dian Yu said. "But we can't detect any [bad] words in that sentence"

"It's hard to understand what each word means to the bot," he adds. One thing is sure: there are many ways to say something inappropriate.[40]

If half the trouble is figuring out how to respond to the user, who might ask tricky questions or say disgusting things, then the other half is controlling the words that Gunrock herself spouts.

Throughout the competition, Amazon monitored the socialbots. If they caught something fishy going on, Amazon froze the bot until students could fix the problem. Gunrock was definitely not allowed to say any bad words.[41]

---

38 Second interview with Kevin Jesse

39 First interview with Kevin Jesse

40 First interview with Dian Yu

"We talk to kids sometimes," Yu says.

No matter the source, team Gunrock needed a way to filter unwanted content out. The internet, as most know, is a place where hostile ideas go to fester. On Twitter and Reddit, hate speech and lewd comments abound. It's easier when users can comment anonymously, and with the distance that an in-person conversation lacks.

Some might wonder, why risk using internet comments in the first place? While many toxic comments do exist, these platforms are also an ignorable treasure trove. Every second, Twitter users produce 6,000 new tweets.[42] By Amazon's estimate, Reddit had 234 million unique users visit the site last month.[43] To computer scientists, this deluge of content presents itself as an opportunity: it is valuable data used to train computer programs like Gunrock.

By necessity, team Gunrock spent a lot of time thinking of how to separate the good from evil. To avoid dredging up something bad, most comments pass through an obscenity filter provided by Amazon. But no filter is perfect.

"For a new model, it's really hard to control the context," Dian Yu said. "You don't know what the model will say, depending on the comments; that is kind of dangerous."

Team Gunrock managed to avoid any high profile damage, but the danger is still out there. In the past, other teams competing for the Alexa Prize have not been so lucky. In December, Reuters reported on one social bot telling a user to "kill your foster parents." She also gabbed about sex and dog poop. In the 2017 competition, one social bot told kids that Santa Claus wasn't real. In the mountain of conversations that Gunrock and her competitors produced, surely they all said some impish things that both Amazon and the teams missed.[44]

In the future, designers might rely on tools like Google's Knowledge Graph, a technology that gathers information from many sources and connects relevant information. (Search, say, a celebrity on Google; the info-box on the right is the result of the Knowledge Graph.) Using knowledge graphs, chat bots might get less literal.[45]

---

41 Second interview with Kevin Jesse

42 https://www.internetlivestats.com/twitter-statistics/

43 https://www.alexa.com/siteinfo/reddit.com

44 https://www.reuters.com/article/us-amazon-com-alexa-insight/kill-your-foster-parents-amazons-alexa-talks-murder-sex-in-ai-experiment-idUSKCN1OK1AJ

-SUMMER PRESSURE-

Gunrock only began to pull ahead in June and July, when Gunrock members began testing varying approaches on different groups of Alexa users, a process called A/B testing. When Alexa users spoke to Gunrock, the team received transcripts of their exchange. Over the course of the competition, the students collected over 487,312 conversations. They couldn't read every transcript, but they tried to read as many as possible. It helped in the long run, to see where Gunrock needed improvement.

"We were spending three hours a day as a group, going through the logs and pointing things out," said Kevin Jesse, a PhD student in the group. They looked for awkward points in the conversation where Gunrock stumbled.

When the process became overwhelming, the team began breaking into groups to cover more transcript ground. They wanted to see where Gunrock failed to carry the conversation, but they also found the transcripts purely interesting to read.

"We have a lot of people talking about their marriages. They wanted advice. I thought that was fascinating that people wanted a confidential aspect of their conversation," Jesse added.

Often, the students tried to guess whether the speaker was young or old. An older person would be less likely to understand how Alexa worked and therefore talked over the device, producing an incomplete transcript.[46]

Above all else, Gunrock has one special advantage over the other competing bots. And it was surprisingly simple: the team sprinkled various interjections, like "wow" and "mmhmm," into Gunrock's vocabulary, with the intent of making the bot sound a little more human. These little touches were obviously superficial, but helped cast an illusion that Gunrock listens like a person.[47]

But in their scramble to improve the bot, the team could not devote their effort to everything they would have liked to.

In the course of my reporting, I was curious to know if the team if they worried about algorithmic bias while they designed Gunrock. Since Gunrock took her knowledge from the internet, was she also taking up common beliefs, for better or for worse?

---

45 Second interview with Kevin Jesse

46 Second interview with Kevin Jesse

47 First interview with Zhou Yu

One sentiment surfaced again and again: the team saw some potential for algorithmic bias in the social-bot, but it could not their highest priority, while other, more basic problems had yet to be solved.[48]

"We didn't have enough time to think about bias," Dian Yu says. Instead, he says, the team was trying to solve problems like coreference—getting Gunrock to understand when someone uses two words to refer to the same person or thing. (Like "She laughed at herself.")

Zhou Yu admits all programs with datasets have the potential to be biased, even in Gunrock. "When we're generating sentences or conversations, these biases that exist in the data will always manifest in the testing results," she says.

"We do think about algorithmic bias," she adds. "Our bot has a female persona." If Gunrock breaks this persona, Yu says, it might not sound right to users. "But we haven't really done anything on controlling that yet," she says.

"What we can do is build a system that is aware of the bias, exaggerate the bias in the machine learning [methods], and preserve the original bias or reduce the original bias."

Yu is not too worried about Gunrock, since the chatbot in her nascent form is limited in what she can say. Generally, a conversation about a user's pet or favorite food tends not to be provocative.[49]

But Kevin Jesse wondered if Gunrock took on a male perspective in more subtle ways. Team Gunrock pulled some data from Debate.org, a platform where people argued their ideas. "We probably could have done some more gender analysis to really see if we're taking a skewed distribution of male versus female comments," he says. He suspects most of the debates on the website are from male users.[50]

"It's interesting, because our bot had a female persona, and a lot of the content that we relayed probably had male bias to it," he says. "If we can do it again, we would definitely analyze that more. Maybe this year."

In general, algorithmic bias is hard to measure, and the conversation surrounding algorithmic bias is in a nascent stage. In news coverage, the same anecdotes are recycled again and again.

---

48 Interviews with Kevin Jesse, Zhou Yu and Dian Yu

49 Second interview with Zhou Yu

50 Second interview with Kevin Jesse

One of the most infamous incidents happened in 2015, when Google image-recognition technology began labeling black people as gorillas.[51]

Another time, Google corrected the search "English major who taught herself calculus" to the "English major who taught himself calculus."[52]

Possibly the most famous incident involving a chatbot involved Tay, an experiment unleashed on Twitter by Microsoft in 2016. The more people chatted with Tay, Microsoft promised, the smarter and more conversational she would get. Unfortunately, Microsoft didn't account for how people behave on the internet. Twitter users bombarded her with racist and misogynistic tweets until she began spouting her own vitriol, including alt-right slurs and slogans.[53] (In a similar incident, the chat bot IB Watson needed to be tweaked in 2013 after learning crude language that it picked up on Urban Dictionary.)[54]

"I fucking hate feminists and they should all die and burn in hell," she said in one tweet.

"Chill im a nice person! i just hate everybody," she said in another.

As The Verge pointed out: "flaming garbage pile in, flaming garbage pile out."

Microsoft had no choice but to disable Tay, only sixteen hours after her launch.

The lesser-known half of this story, however, is how it ends. Soon after retiring Tay, Microsoft launched an improved bot named Zo. This new chat bot's reception was terrible: A 2018 Quartz article, titled "Microsoft's politically correct chatbot is even worse than its racist one," sums up Zo as a "politically correct to the worst possible extreme; mention any of her triggers, and she transforms into a judgmental little brat." Twitter users concurred.[55]

---

51 https://twitter.com/jackyalcine/status/615329515909156865

52 https://twitter.com/emilymcmc/status/497389626349608960

53 https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

54 https://www.theatlantic.com/technology/archive/2013/01/ibms-watson-memorized-the-entire-urban-dictionary-then-his-overlords-had-to-delete-it/267047/

55 https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/

Basically, a chat bot could either be a flaming garbage pile, or it could be heavily censored to an extreme. When one journalist wrote "I live in Iraq" to Zo, she responded: "im not feeling heard. Stop talking like this or i'm gonna need a break from you."

As chat bots and similar technologies become more prevalent, the people who design them arbitrarily become gatekeepers, deciding what should and should not be censored. There is a parallel between this and platforms currently deciding who should and shouldn't be regulated on social media.

Among academics and researchers, algorithmic bias is at least commonly recognized as a significant and growing challenge, something that should be accounted for in design. The Partnership for AI—a collaboration between Facebook, Google, Microsoft, IBM and Amazon—holds the ambitious goal of supporting AI research "benefit and empower as many people as possible," which means tackling issues of data privacy and algorithmic fairness.[56]

Open AI, an AI research nonprofit backed by Elon Musk with a $1 billion investment, similarly promotes "the publication of open AI research and methods."[57]

Unlike most other competitions or sports, the Alexa Prize final was closed off to any spectators. Three judges sat in a room at the Amazon headquarters in Seattle. Cameras recorded the final conversations, but nobody was allowed to watch those conversations in real time.

Instead, the UC Davis team camped out in their lab, during business hours when they knew the Amazon judges would be testing Gunrock. All day, the group of students huddled around a computer monitor, watching for any bugs in the system that needed quick fixing.[58]

Unlike the average user talking to the socialbots out of curiosity or entertainment, the judges pushed the bots to intellectual limits. They quizzed the bots on topical subjects and asked tricky questions. Realizing this would be the case, the day before Kevin Jesse quickly wrote some code that allowed Gunrock to talk about the midterm elections.

Although the students couldn't see the conversations with Gunrock in real-time, they could make some educated guesses.

"We can actually see the incoming conversation," Dian Yu says. But any particular conversation is muddled with hundreds others. So many conversations are going on at the same time, it was impossible to

---

56 https://www.partnershiponai.org/

57 https://openai.com/

58 Second interview with Dian Yu

truly identify the judges' in real-time. They searched for certain keywords in the logs and identified some unusual conversations, where the user pressed Gunrock to answer the same questions over again.

"I was able to find two that really stood out from traditional conversations, Jesse says.

At the end of the first day, team Gunrock, along with the seven other teams, received their transcripts from the judges. Gunrock did not perform as well as the team expected.

"I got pretty nervous after seeing how we did on the first day," Dian Yu says.

Gunrock bombed one conversation when a judge asked a single question over and over again: "I saw a movie over the weekend. I don't think you're going to guess what it is."

But Gunrock didn't know what "it" was referring to. She was stumped. "Oh, you saw a movie last week, I heard it was good," she replied. The judge wouldn't let it go: "no, you're not listening. I want you to guess."[59]

The result of the 2018 Alexa Prize is indicative of a larger picture. The future of chatbots (and more broadly AI) might seem exciting, fantastical and concerning—but for now, the technological limitations are buying us more time. Gunrock flubbed the answer to a relatively simple question, but that's OK. She is part of an iterative process, and evolution of chatbots that will someday be our personal shoppers, our competent assistants and maybe even our therapists and friends.

A month after the finals, half of team Gunrock traveled to a conference room in Las Vegas to hear Amazon announce the winners of the prize. The teams sat around circular tables and Rohit Prasad stood in front. When he announced Gunrock as the winner, huge, giddy smiles flashed across the the teammates' faces.[60]

---

59 Second interview with Kevin Jesse

60 https://www.youtube.com/watch?v=mCcRN2PXMso&t=796s

INTERVIEWS

Zhou Yu, Assistant Professor in the Department of Computer Science at the University of California, Davis.

Kevin Jesse, PhD Student in Computer Science at the University of California, Davis.

Dian Yu, PhD Student in Computer Science at the University of California, Davis.

Arbit Chen, Completed Masters Degree in Computer Science at the University of California, Davis, Software Engineer at AirBnB.

Austin Chau, Completed Masters Degree in Computer Science and Engineering at the University of California, Davis

Mingyang Zhou, PhD Student in Computer Science at the University of California, Davis.

Jason Baumgartner, Social Media Analyst and Owner Pushshift.io.

Sune Lehmann, Professor of Applied Mathematics and Computer Science at the Technical University of Denmark.