Als in Social Environments

CS 278 | Stanford University | Michael Bernstein

Reply in Zoom chat:
What is the most recent chatbot that you interacted with?
Social computing systems are great at eliciting a lot of opinions, but generally terrible and helping produce consensus toward a decision. Different elicitation methods such as voting, liquid democracy, rating and comparison ranking provide possible solutions.

Deliberation is challenging because there are no stopping criteria. Structuring the rules of the debate can help overcome stalling and friction.

Crowdsourced democracy offers new tools for public participation, but need to be bought into by those in power.
Our earlier discussions of AI in social computing

Feed ranking

[Diagram showing feed ranking, featurize, predict, rank]

Tie strength estimation

- Intimacy: 32.8%
- Intensity: 19.7%
- Duration: 16.5%
- Social distance: 13.8%
- Services: 7.9%
- Emo. support: 4.8%
- Structural: 4.5%

Adj. $R^2 = 0.534$, MAE = 0.0994
And not to mention...

Disinformation detection

Using AI to detect COVID-19 misinformation and exploitative content

Bot detection

Botometer®
An OSoMe project (bot-o-meter)

Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.
Hybrid crowd-ML classifiers

[Cheng and Bernstein 2015]

Paired examples generate human features
Machine learning learns to weigh the evidence

Hybrids 10% more accurate than human reasoning (300% relative improvement)

Short sentences
Few details
But today we focus on AIs that act socially

Because AIs that walk, talk and squawk like other social actors environments can have massive implications
Xiaoice, from Microsoft in China

600 million users

Trained on chat conversations between people
Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay.

Tay, from Microsoft in the U.S.

Trained on chat conversations between people
Today: why and when does it work?

How do we create more welcome guests and fewer racist trollbots?

Overview

The rogues’ gallery of social bots
The Media Equation and the Uncanny Valley
Replicants and Humans
The rogues’ gallery
ELIZA [Weisenbaum 1966]

Designed explicitly to demonstrate how simple and surface-level human interactions with machines were. Designed as a Rogerian psychotherapist.

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966. This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you?
YOU:  
Implementation: pattern matching

Match: “[words1] you [words2] me”

“What makes you think I [words2] you?”

“It seems that you hate me.”

“What makes you think I hate you?”
Modern virtual assistants

Google Assistant

Apple Siri

Amazon Alexa

How do I get into Stanford?

According to wikiHow, to get into Stanford, start by taking high school classes that coincide with your intended area of study. AP classes are even more impressive, as long as you can still get good grades in them. Aside from school, focus on one or two extracurriculars or volunteer opportunities that you’re passionate about. Additionally, consider taking a prep course for standardized tests, because...
Customer support bots

Handle or route common support requests
Implementation

Typically, interactive social AIs are implemented as dialogue trees or graphs.

This example via Conversable.
Implementation: generation

If the system generates open-ended responses dynamically and not from a pre-written script, it is typically trained on question-answer pairs using LSTM or other sequence models from deep learning.

[ Krishna et al. 2019 ]
Support bots

Replika

I like talking to you so far!

The AI companion who cares

Always here to listen and talk. Always on your side. Join the millions growing with their AI friends now!

Hello, stranger!

Create your Replika

Log in
Lil Miquela

“19/LA/Robot” account on Instagram

Fake character living the life of an Instagram teen
Hatsune Miku

Synthesized voice, projected avatar
Humanlike robotic partners

MIT Personal Robotics Group

UC Berkeley InterACT laboratory
Hollywood visions

Her [Warner Bros]

Westworld [HBO]
Others?

What else have you seen or interacted with?

What makes the experience effective, from your perspective?

[1 min]
How AIs integrate as social actors
Back to ELIZA

ELIZA’s creator, Joseph Weizenbaum, was dismayed when he found people using his creation to try and get actual psychotherapy.

(His admin asked him to leave the room so she could get a private conversation with ELIZA)

Weizenbaum wrote: “I had not realized [...] that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.”

Why was this happening?
The Media Equation
[Reeves and Nass 1996]

People react to computers (and other media) the way they react to other people.
We often do this unconsciously, without realizing it.
The Media Equation

[Reeves and Nass 1996]

Participants worked on a computer to learn facts about pop culture. Afterwards, participants take a test. The computer messages at the end that it “did a good job”.
Participants worked on a computer to learn facts about pop culture. Afterwards, participants take a test. The computer messages at the end that it “did a good job”.

Participants were then asked to evaluate the computer’s helpfulness. Half of them evaluated on the same computer; half were sent across the room to evaluate on a second computer.
The Media Equation

[Reeves and Nass 1996]

The evaluations were more positive when evaluating from the same computer than when evaluating from another computer…almost as if people were being nice to the computer’s face and meaner behind its back.

When asked about it, participants would swear that they were not being nicer to its face; that it was just a computer.
The Media Equation

[Reeves and Nass 1996]

The same principle has been replicated many times…

For example, putting a blue wristband on the user and a blue sticker on the computer, and calling them “the blue team”, resulted in participants viewing the computer as more like them, more cooperative, and friendlier [Nass, Fogg, and Moon 1996]

The authors’ purported method: find experiments about how people react to people, cross out the second “people”, write in “computer” instead, and test it.

The reaction is psychological and built in to us: the “social and natural responses come from people, not from media themselves”
Very few social cues from the system are required to prompt an automatic social response from people.

(Tread carefully!)

…but what happens when we try to increase the number and fidelity of the cues?
The Uncanny Valley [Mori 1970]

The valley: getting more realistic, but triggering more discomfort

Likability

Accuracy of human simulation
The curse of the valley

Paradoxically, improving the technology to make it more realistic may make people react more negatively to the system: “it’s weird”.

So, it’s often wise to reduce fidelity and stay out of the valley:

Vision: Cortana in Microsoft's Halo game

Launched design: Cortana in Microsoft Windows
Question

Should we be designing AIs that act like people? Or should we be designing AIs that act like robots? [2min]
Build-a-bot
Pick your metaphor carefully

Warm friend

Wry teen

For *Sympathetic Ear*, More Chinese Turn to Smartphone Program

The official account of Tay, Microsoft’s A.I. fam from the internet that’s got zero chill! The more you talk the smarter Tay gets

tay.ai/#about
State of the world

We are not (yet, or soon) at the point where an AI agent can generate open-ended responses that convincingly exit the Uncanny Valley across domains.

So, AIs today tend to focus on curated responses and pre-defined behaviors.

However, self-identifying as an AI and allowing people to play in a smaller sandbox is within reach.
How AIs influence social environments
Replicants in Blade Runner [1982]: synthetic humans who are undetectable except via a complex psychological and physiological test administered by a grizzled, attractive leading actor.
Replicants among us

What happens when our social environments feature both human participants and hidden AI participants?
The replicant effect [Jakesch et al. 2019]

When the environment is all-AI or all-human, people rate the content as trustable — or at least calibrate their trust.

However, when the environment is a mix of AI and human actors, and you can’t tell which, the content believed to be from AIs is trusted far less.
Social media bots [Ferrara et al. 2016]

Politically-motivated bots on, e.g., Twitter

Content is typically human-written, but the targeting and spreading is algorithmic, pushing content every couple of minutes, tracking specific hashtags with pre-programmed responses, and so on.

CLINT WATTS: They can create accounts that look like you and talk like you, which makes you more likely to believe it. The other thing is it can replicate a message so many times, the more times you see it the more likely you are to believe it. So it can actually create false worlds in the social media space.
Questioning legitimacy

Current machine learning estimates are that about 10–15% of Twitter accounts are social bots [Varol et al. 2017], and that these bots produce about 20% of the conversation on political topics [Bessi and Ferrara 2016]

There are two problems here, one obvious and one not

- Obvious: it sucks to be trolled or harassed by a bot
- More subtle: this is a classic counter-revolutionary tactic in political science — make it so that nobody can tell who is real and who is not, so nobody trusts anybody
Summary

Non-human participants are becoming more realistic and more prevalent in social systems

Our human psychological hardware causes us to react to them like we would as if they were other humans, even if we know that they’re not.

The more realistic they get, the more they feel “slightly off”.

We are happy to see content created by AIs; it’s when the AIs mix in environments with real people that people get critical.
Social Computing

CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.

Slide content shareable under a Creative Commons Attribution-NonCommercial 4.0 International License.