Moderation examples

From Daphne L: What happens when all the content moderators get sent home suddenly due to the onset of COVID?

From Julia R: In 2020-2021, Twitch struggled to build policies around “hot tub streamers”
Misinformation: going antiviral
CS 278 | Stanford University | Old Tom Holland
Announcements

You did it! Midterm and project milestone are in.

Final reflections for Assignment 3 are due Monday (we extended it — take a break). Check out the assignment website to see what happened to your crowdsourcing questions.

Final projects are due June 9 (the Friday that begins finals): a bit over two weeks from now.

No late days on the final project...because we don't get late days on reporting final grades for graduating students to the Registrar.

Last lecture: suggest topics you want me to discuss or riff on.
Last time

As Gillespie argues, moderation is the commodity of the platform: it sets apart what is allowed on the platform, and has downstream influences on descriptive norms.

Moderation works: it can change the community’s behavior

Moderation classification rules are fraught and challenging — they reify what many of us carry around as unreflective understandings.
Russian propaganda effort helped spread ‘fake news’ during election, experts say

Russian President Vladimir Putin, in an interview with RT in 2013, said that he wanted to “break the Anglo-Saxon monopoly on the global information streams.”

How Syria's White Helmets became victims of an online propaganda machine
Rumble Sends Viewers Tumbling Toward Misinformation

Research shows the emergent video platform can recommend conspiracy theories and other harmful content, often not.

By Todd Spangler

The New York Times

TikTok’s search engine repeatedly delivers misinformation to its majority-young user base, report says

By Emma Tucker, CNN

Published 7:11 PM EDT, Sun September 18, 2022
Social computing is in the thick of it

Social media didn’t create misinformation, but it did lower the friction associated with sharing and it made it easier for groups to find each other and share disinformation.

And even if we didn’t start the fire, we bear responsibility for what happens in our spaces.
The opportunity

“The digitization of information exchange, however, also makes the practices of disinformation detectable, the networks of influence discernable, and suspicious content characterizable.”

[Bliss et al. 2021]
Today’s scope

Misinformation combines a number of threads in technology, design, politics, communication, and many other disciplines.

I’ll be focusing on its intersection with the topics of this course: the design of social computing systems.

Outline

How does misinformation originate and propagate?
Can we detect misinformation?
Can design help?
Poll: which design will better reduce the spread of disinformation?
From Whence Disinformation?
Let’s talk terms

When will I be referring to misinformation and when to disinformation throughout this lecture?

Misinformation = anything false
  
  Might be a rumor, or something not necessarily intentionally false

Disinformation = the specific intent is to deceive
  
  Often built around a true or plausible core, wrapped up in a misleading way
Why now?

What does the internet actually change here, compared to offline interaction?

The effort required to connect groups together has lowered, making it possible for identity-based groups to connect that might otherwise have not

...where there might not otherwise have been a critical mass of members

...where it might otherwise have been too much effort
Why now?

This reduction in activation energy carries positives and negatives:

Positive: social movements that are forced underground, such as LGBTQ military service members, can connect with each other online [Sheng 2020]

Negative: hate groups can also connect with each other online
Why now?

When groups can convene and push their own narrative, it enables “common knowledge attacks on democracy” [Farrell and Schneier 2018].

In other words, it can destabilize democracy by flooding public debate and confusing our shared understandings and expectations, which are required for democracy to function.
Fingers pointed 👉

#1, “It’s trolls”: disinformation factories such as the Russian Internet Research Agency generate disinformation to harm us [Bail et al. 2020]

#2, “Post truth”: people default to motivated reasoning, which means that we are inclined to believe information that is consistent with our political views, and disinclined to believe information that contradicts our political views [Kahan 2017]. We are more loyal to political party than loyal to truth [Van Bavel and Pereira 2018]

While these explanations are not wrong, they are also not the explanations with the strongest evidence
state actors

Yes, state actors exist.

Twitter retweet network for Black Lives Matter in 2016: Russian IRA (orange) both posed as BLM activists on the left, and infiltrated anti-BLM communities on the right [Starbird, Arif, and Wilson 2019]
But is it just state actors?

Context is the Syrian Civil War and the White Helmets, a humanitarian response group. Anti-White Helmet accounts — pink — are dominant in volume, delegitimizing the White Helmets’ claims.

Not just bots and trolls: lots are journalists aligned with Syrian and Russian government interests, Syrian and Russian government members, and alt. media.

It looks more like activism than “just” disinformation.

[Starbird @ CS 547, 2019]
Disinformation campaigns often involve many **unwitting agents** who are unaware of their role and whose views and behaviors have been shaped by motivated actors [Bittman 1985, Starbird, Arif, and Wilson 2019].

**Cold War-era Soviet technique**: sell journalists on anonymous tips aligned with their beliefs. Once one journalist took the bait, others became interested.
Activist / motivated actor → Color the truth → Unwitting agents → Reinforcing existing beliefs

Orange arcs = blue accounts retweeting content from orange accounts = unwitting agents

Left-leaning

Right-leaning
Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

[Starbird 2021; Prochaska 2023]
Participatory Disinformation
The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.
Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

[Starbird 2021; Prochaska 2023]
Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.

"Grassroots" activists and social media influencers help amplify these stories up to the political elites.

Online "crowds" generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.
Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.

“Grassroots” activists and social media influencers help amplify these stories up to the political elites.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.
Participatory Disinformation
The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

- Political Elites
  - Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.
  - Online "crowds" generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.
  - "Grassroots" activists and social media influencers help amplify these stories up to the political elites.
  - Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.

- Audiences echo and reiterate a growing sense of grievance. Violent language increases.
Participatory Disinformation
The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

presented by
Kate Starbird, UW, HCDE
Center for an Informed Public

[Starbird 2021; Prochaska 2023]
Misinformation is a collaborative effort between producers and the audience [Starbird, Arif, and Wilson 2019].

Political Elites

Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.


"Grassroots" activists and social media groups help amplify the frame, up to the political elites.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

Political elites echo, false/misleading stories reinforce the frame. Mobilize and organize the audiences.

Protests

#StopTheSteal

January 6 Attack on Capitol

[Starbird 2021; Prochaska 2023]
While we are more likely to believe news that is concordant with our beliefs, the larger effect is whether we engage with higher-level reasoning instead of automatic reasoning [Pennycook and Rand 2021].

Better discernment of true vs. false news

Across a bunch of studies

Engaged in higher-level reasoning (= bigger effect)

VS.

Saw belief-concordant news (= smaller effect)
More like post-attention…

[Pennycook et al. 2021]

People rate accuracy as the single most important factor when deciding whether to share.

However, whether a headline is politically concordant has a much larger effect on sharing intention than the headline’s accuracy.

So what gives?
More like post-attention...

[Penneycook et al. 2021; Penneycook and Rand 2022]

Theory: we don’t pay attention to accuracy, and are more focused on pleasing followers or signaling group membership

Evidence: focusing participants’ attention on accuracy before seeing a headline reduces sharing of false headlines by half

<table>
<thead>
<tr>
<th>% of people likely to share false headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
</tr>
<tr>
<td>30%</td>
</tr>
<tr>
<td>primed to think about accuracy</td>
</tr>
<tr>
<td>10%</td>
</tr>
<tr>
<td>20%</td>
</tr>
<tr>
<td>30%</td>
</tr>
</tbody>
</table>
What about deepfakes?

Generative AI models can create potentially plausible fakes

But even very simple transformations can convince easily
How much do we consume?

Most people rarely see misinformation.

National sample of mobile, desktop, and TV consumption: misinformation is 0.15% of Americans’ media diet [Allen et al. 2020]

The average US adult saw ~1 misinformation story in the 2016 election [Hunt and Gentzkow 2017]

Exposure to misinformation is highly concentrated [Guess, Nyhan, and Reifler 2020]: 1% of people account for 80% of exposures to misinformation [Grinberg et al. 2019]

This exposure is typically pro-attitudinal [Guess, Nyhan and Reifler 2020]
How much do we share?

It’s rare: most never share disinformation.

In the 2016 election, >65 year olds were 7x more likely than teenagers to share disinformation, and conservatives were more likely to share than liberals [Guess, Nagler, and Tucker 2019].
Mainstream media amplify the message

Analysis of mail-in voter fraud disinformation suggested that social media played a smaller role than mainstream media in 2020:

1) Trump tweets

2) Mainstream media, trying to be neutral and avoid claims of anti-conservative bias, cover Trump’s claims and thereby spread them

[Benkler et al. 2020]
So, from whence disinformation?

Finger #1: “It’s trolls.”

Actually: “It’s motivated actors, who activate unwitting agents.”

Finger #2: “Post-truth”

Actually: “People care about being accurate, but generally aren’t paying attention to accuracy when they share.”

“It’s everywhere”

Actually: “Exposure and sharing is rare, but very concentrated.”
Classification
Machine learning

Some categories of misinformation (e.g., near copies of flagged articles) can be flagged automatically

ML APPLICATIONS

Using AI to detect COVID-19 misinformation and exploitative content

May 12, 2020

The COVID-19 pandemic is an incredibly complex and rapidly evolving global public health emergency. Facebook is committed to preventing the spread of false and misleading information on our platforms. Misinformation about the disease can evolve as rapidly as the headlines in the news and can be hard to distinguish from legitimate reporting. The same piece of misinformation can appear in slightly different forms, such as an image modified with a few pixels cropped or augmented with a filter. And these variations can be

Factcheckers

Twitter and Facebook have started relying on third party fact checkers to decide whether an article is misinformation.
Factcheckers

Twitter and Facebook have started relying on third party fact checkers to decide whether an article is misinformation.

However, this does not cover the long tail: Facebook’s partners comprise 26 fact checkers who collectively review 200 articles per month [Rodrigo 2020].

Fact checkers can also take days to do the research, by which time the article or video has spread widely.
Are there alternatives?

Representative samples of “crowd jurors” can be as accurate as fact checkers and much faster [Allen et al. 2021]

But, allowing anyone to participate can make the flagging another lever for politically motivated groups.
Why is classification so hard?

For one, fact checkers are thorough.

More centrally, much disinformation isn’t entirely made up. It’s often a slant on a true story, walking the line between truth and falsehood.

So what do we do, if most stories don’t have a clear “true”/“false” line? What do we remove?
Twitter’s pre-Elon criteria

<table>
<thead>
<tr>
<th>Misleading Information</th>
<th>Label</th>
<th>Removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disputed Claim</td>
<td>Label</td>
<td>Warning</td>
</tr>
<tr>
<td>Unverified Claim</td>
<td>No action</td>
<td>No action*</td>
</tr>
</tbody>
</table>

| Propensity for Harm | Moderate | Severe |

See [Atreja, Hemphill, and Resnick 2023] for more
Disinformation campaigns

[Starbird, Arif, and Wilson 2019]

Instead of classifying individual pieces of content, we can study and classify disinformation campaigns — a collection of information actions

1) Is this campaign pushing a false narrative?

Then, classify:

2) Is this article a part of this disinformation campaign?

This way, the decision can be made on the aggregation of actions, and a classifier only needs to assign the story to the campaign.
Interventions
Reduce feed ranking

Platforms can (temporarily) reduce the feed ranking of links that might be disinformation, slowing their spread while fact checkers review it.

Ex: Article is lower in your Facebook feed, video is recommended less often on YouTube.

Pros: walks a line between removal and unconstrained spread.

Cons: opaque, unclear when it’s happening, likely too late once other media start reporting on it.
Correction: “Well actually”

Early research suggested that fact corrections could backfire and reinforce incorrect beliefs [Nyhan and Reifler 2010]

But, recent work fails to replicate the backfire effect and suggests that people do revise beliefs after a correction [Wood and Porter 2019; Guess and Coppock 2018]

There does seem to be one negative of correction, though: getting corrected leads people to decrease the quality and increase the political slant and toxicity of future content [Mosleh et al. 2021]
Better: provide information before narratives begin
An intervention gallery

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi.

Learn how voting by mail is safe and secure

9:20 AM · 10/09/20 · Twitter Web Client
An intervention gallery
An intervention gallery
An intervention gallery

U.S. elections
Robust safeguards help ensure the integrity of election results.

CHRISTINA BOBB: Trump won. MSM hopes you don’t believe your eyes.
Implied truth effect

Labeling some stories as false leads people to believe that everything not explicitly labeled as false...is true. [Pennycook et al. 2020]

This is problematic when fact checkers can only check a tiny percentage of all content on the site.
Source credibility signals

Technique: flag source trust level (e.g., NewsGuard)

Unfortunately, these signals do not change news consumption behavior [Aslett et al. 2022]
**Priming accuracy**

Bringing attention to the accuracy of information shared on Twitter **improves the quality of news shared later** [Pennycook et al. 2021]

Why? Recall: we’re not in a post-truth world, where people don’t care about accuracy. We instead tend to be more focused on other motivators, like pleasing our followers.
Priming accuracy in practice

We designed an intervention that reduced shares of flagged content on TikTok by 24% via a large scale RCT, thread ↓ 1/7

We put a short prompt on videos that reminded people to think about the accuracy of the content they were watching. And then - when people went to share the video - we reminded them again that the video was flagged & asked them if they were sure they wanted to share. 3/7

In addition to successfully reducing shares by 24%, our intervention also reduced likes by 7%, and views by 5%. 6/7

[https://twitter.com/IrrationalLabs/status/1357033901311451140]
Back to our question: which design will better reduce the spread of disinformation?
Facebook's arc

fact check → related articles → fact check
No politically-neutral option

There exists vastly more conservative-leaning disinformation than liberal-leaning disinformation [Hunt and Gentzkow 2017]

This difference persists even if the links are evaluated by balanced groups, or by groups of only conservatives [Mosleh et al. 2023]

So the issue is hot-button political, in addition to intersecting questions of freedom of expression

What do you think the platforms should do? [2min]
Summary

misinformation != disinformation

Disinformation is often created and amplified collectively by motivated actors and their audience

People share misinformation when they are not paying enough attention to accuracy cues

Fact checkers are often slow; but, it might be possible to classify campaigns rather than information items


Aslett, Kevin, et al. "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions." Science advances 8.18 (2022): eabl3844.


References


References


References


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.