Warning Users about Online Disinformation

Ben Kaiser
INTERFACE 2/24/2020
Outline

- Background
- Ways of countering disinformation
- Labels, fact checks, and warnings
- Security-style warnings
- Ethical considerations
Online Disinformation

- Disinformation preys on cognitive shortcuts we take when evaluating info
  - Preference for familiarity
  - Reliance on endorsements
  - Self-confirmation + motivated reasoning

- The internet has made it easy, cheap, and effective to spread disinformation
  - Traditional markers of credibility / legitimacy are gone
  - Popularity + repetition can be manufactured
  - Algorithmic feeds encourage selective exposure

- Political operatives and profiteers are taking advantage
  - Foreign influence
  - Domestic political influence
  - Clickbait news and ad fraud
Types and Terminology

- **Disinformation**: sometimes a catchall; also specifically refers to false information distributed *with the intent to harm*
- **Misinformation**: false information distributed *without harmful intent*
- **Malinformation**: true information distributed *with harmful intent*
- **Fake news**: sometimes a catchall; politically loaded; let’s avoid this
- **Junk news**: information that presents itself like news but does not follow journalistic norms like transparency, objectivity, and veracity
- **Clickbait**: sensationalized, insubstantial information
7 COMMON FORMS OF INFORMATION DISORDER

1. SATIRE OR PARODY
   - No intention to cause harm but has potential to fool

2. MISLEADING CONTENT
   - Misleading use of information to frame an issue or individual

3. IMPOSTER CONTENT
   - When genuine sources are impersonated

4. FABRICATED CONTENT
   - New content is 100% false, designed to deceive and do harm

5. FALSE CONNECTION
   - When headlines, visuals or captions don’t support the content

6. FALSE CONTEXT
   - When genuine content is shared with false contextual information

7. MANIPULATED CONTENT
   - When genuine information or imagery is manipulated to deceive
Countering Disinformation

1. Automated detection
   a. Claim verification
   b. Credibility assessment
2. Deplatforming + cutting off revenue
3. Improving recommendation and ranking algorithms
   a. Easy: deprioritizing known disinformation
   b. Hard: creating platforms that encourage thoughtfulness
4. Fact-checking + real journalism
5. Warnings + labels
Labels and Warnings - Fact Checks

News about Rick Scott Critical Condition

Florida Governor Rick Scott Critically Injured During Hurricane Irma Cleanup?

snopes.com - 11 hours ago
On 11 September 2017, the "satirical" website Last Line of Defense falsely reported that Florida Governor Rick Sco...
Labels and Warnings - Other Types

Related Information

Source Information (stance or credibility)
Security-style Warnings

- In infosec, security warnings are a huge field of research
- After dozens of studies and field tests, they have become very effective
  - The first study in 2003 found that about 70% of people ignored warnings
  - The latest studies show that only 10-25% click through
  - **Interstitial** warnings are much better than **passive contextual** warnings
- They have clean design, clear messages, and actionable choices
Our Study

- We adapted these warnings for disinformation and tested them in two studies.
- Subjects searched for answers to questions on Google, and we tested if warnings made them more likely to double check information.
- They are extremely effective: 86% double checked in treatment rounds versus 19% in control rounds.
Ethical Questions

● Is it OK to adapt security warnings for disinformation?
  ○ Malware and disinformation are very different threats. Is it coercive to make people think they’re in danger?
  ○ Could we accidentally habituate people and therefore make security warnings less effective?

● Should a tool that can make people disbelieve information be built?
  ○ Could it be abused for censorship?
  ○ Even if it’s not misused, is it inherently restricting freedom of choice?

● How strong of a warning is too strong? How do we decide where the line is?

  How much control should platforms exercise over what information users see?