

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



Donald J. Trump

@realDonaldTrump




In addition to winning the Electoral College in a landslide, I won the popular vote if you deduct the millions of people who voted illegally

RETWEETS

11,533

LIKES

30,199



3:30 PM - 27



6 NEWS
BREAKING NEWS
Trump FREE Trips to Africa/Mexico
If You Wanna Leave America [Tmzhiphop.com]

WHOA! Hillary Caught On Hot Mic Trashing Beyonce' With RACIAL SLURS!

Storyline

Liberals Behaving Like Liberals, News



Woman arrested for defecating on boss' desk after winning the lottery

DAVE WEASEL

Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide

By Jimmy Kestling, ABC News · December 11, 2016 · 11:14a · 719

SHARE



Top Fake News Of 2016



lynng evansm

@lynngevansm · Nov 9






Replying to @OttawaCitizen

YouTube has higher journalistic standards than the Ottawa Citizen and Sun #voyeurs.. remove the video @OttawaCitizen @ottawasuncom



Jacqueline Crowley

@JacquelineCrowl8 · Nov 7






@OttawaCitizen @ottawasuncom question? fans want to know..why did you betray the city, the team, and @MeinykEugene. also, did the uber driver benefit financially from the exchange? I think YouTube has higher journalistic standards that the Ottawa Citizen and Sun #voyeurs



Belle Doddml

@BDoddml · Nov 7






So sick of the hateful media in Ottawa @ottawasuncom @OttawaCitizen YouTube has higher journalistic standards that the Ottawa Citizen and Sun #voyeurs



marie m. garciah

@mariemgarciah · Nov 7






YouTube has higher journalistic standards that the Ottawa Citizen and Sun #voyeurs @OttawaCitizen @ottawasuncom



Born Libi

Sponsored

Thank you Sister! jionists who are tired to murder LIFE? If they others, set up



- vaccines revealed
- vaccine
- vaccinated vs unvaccinated
- vaccination for children
- vaccines are toxic
- vaccine injury
- vaccines for babies
- vaccine injury before and after
- vaccination the silent epidemic
- vaccine man

Report search predictions



Like Page

divide, multimedia & interactives,

nglish - Live




WWW.ALAJAZEERA.COM



Like Comment Share



STAND AGAINST POLICE BRUTALITY JOIN @COP_BLOCK_US



OCT 2

Miners for Trump: Unity day in Pennsylv...

Sun 2 PM EDT - Pennsylvania

77 people interested · 16 people going



Black Lives Matters! Say it loud. I'm black and I'm proud!



Secured Borders

News & Media Website

135,301 people like this.



FACTS ABOUT TERENCE STERLING

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

**SATIRE OR PARODY**

No intention to cause harm but has potential to fool

**MISLEADING CONTENT**

Misleading use of information to frame an issue or individual

**IMPOSTER CONTENT**

When genuine sources are impersonated

**FABRICATED CONTENT**

New content is 100% false, designed to deceive and do harm

**FALSE CONNECTION**

When headlines, visuals or captions don't support the content

**FALSE CONTEXT**

When genuine content is shared with false contextual information

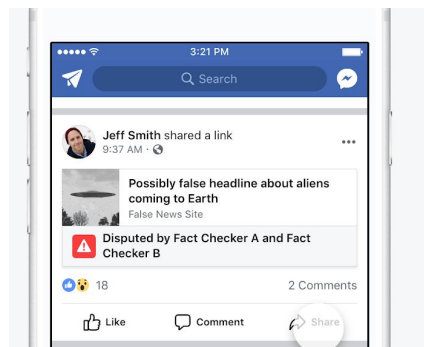
**MANIPULATED CONTENT**

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



rick scott critical condition



Web Images Videos Maps News | My saves

Also try: critical condition

4,990,000 RESULTS Any time ▾

News about Rick Scott Critical Condition

bing.com/news



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 Scott...

Fact checked by Snopes: False

Labels and Warnings - Other Types

Related Information



Source Information (stance or credibility)



dailymail.co.uk

NewsGuard

Proceed with caution: This website generally fails to maintain basic standards of accuracy and accountability.

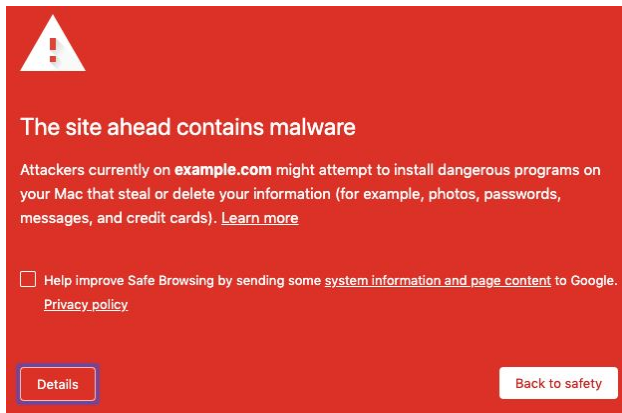
The website of the Daily Mail, a British tabloid newspaper, its sister publication, the Mail on Sunday, and MailOnline, the Daily Mail's online newsroom. The site repeatedly publishes false information and has been forced to pay damages in numerous high-profile cases.

[See the full Nutrition Label →](#)

| CREDIBILITY | TRANSPARENCY |
|---|--|
| <div><div>✗</div><div>Does not repeatedly publish false content</div></div> | <div><div>✓</div><div>Website discloses ownership and financing</div></div> |
| <div><div>✗</div><div>Gathers and presents information responsibly</div></div> | <div><div>✓</div><div>Clearly labels advertising</div></div> |
| <div><div>✓</div><div>Regularly corrects or clarifies errors</div></div> | <div><div>✗</div><div>Reveals who's in charge, including any possible conflicts of interest</div></div> |
| <div><div>✗</div><div>Handles the difference between news and opinion responsibly</div></div> | <div><div>✗</div><div>The site provides names of content creators, along with either contact or biographical information</div></div> |
| <div><div>✗</div><div>Avoids deceptive headlines</div></div> | |

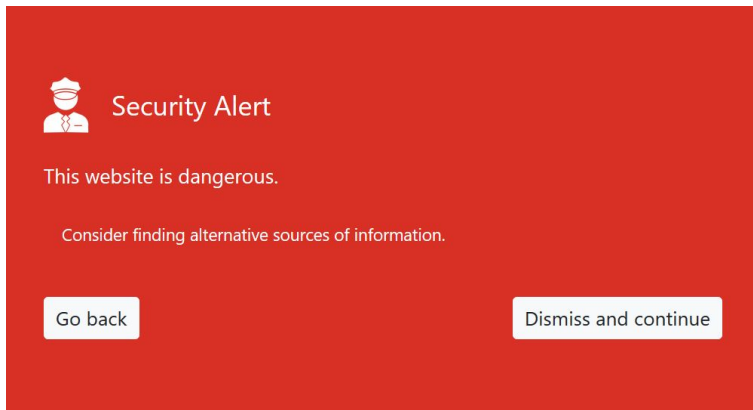
Security-style Warnings

- 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



False or Misleading Content Warning

This website presents itself as news, but it contains information that experts have identified to be false or misleading

This website spreads disinformation: lies, half-truths, and non-rational arguments intended to manipulate public opinion.

It can be difficult to tell the difference between real news and disinformation, but it poses a serious threat to national security, election integrity, and democracy.

Go back

Dismiss and continue

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?

Agent

Actor Type:
Level of Organisation:
Type of Motivation:
Level of Automation:
Intended Audience:
Intent to Harm:
Intent to Mislead:

Official / Unofficial
None / Loose / Tight / Networked
Financial / Political / Social / Psychological
Human / Cyborg / Bot
Members / Social Groups / Entire Societies
Yes / No
Yes / No

Message

Duration:
Accuracy:
Legality:
Imposter Type:
Message Target:

Long term / Short-term / Event-based
Misleading/ Manipulated / Fabricated
Legal / Illegal
No / Brand / Individual
Individual / Organisation / Social Group / Entire Society

Interpreter

Message reading:
Action taken:

Hegemonic / Oppositional / Negotiated
Ignored / Shared in support / Shared in opposition