CENTER FOR INFORMATION TECHNOLOGY POLICY

Adapting Security Warnings to Counter Online Disinformation

Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, Jonathan Mayer

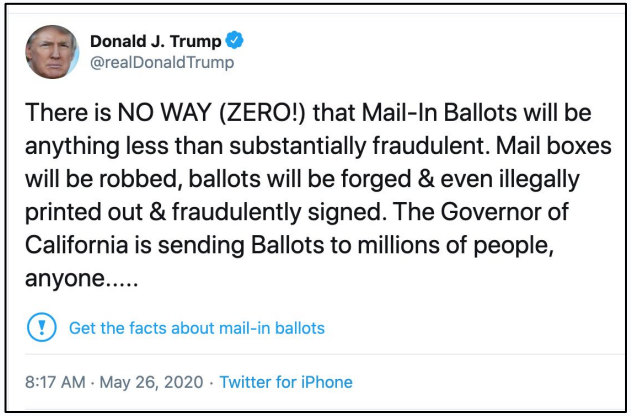
Princeton University

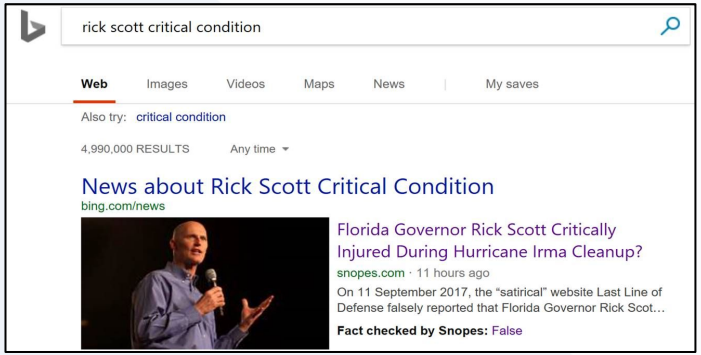
USENIX Security 2021

Outline

- Platform disinformation warnings: examples and evaluations
- Browser security warnings: a success story
- Our research: designing disinformation warnings that work
- Conclusions and recommendations

Platform Disinformation Warnings





Why use warnings?

- **Add context,** instead of restricting speech
- **Induce resistance** to misbeliefs[1] and **increase susceptibility to corrections**[2]

[1]Greene 1982 [2]Ecker 2010

Research on Modern Warnings

Methods

- In-laboratory survey experiments
- Self-reported assessments of:
  - Perceived accuracy
  - Likelihood of sharing
- Contextual warnings only
  - Primarily "disputed" warnings

Findings

- Warnings **modestly** decrease perceived accuracy[1,2,3]
- **Prior exposure** is more important than warnings[3]
- 3 separate studies found that warnings had **insignificant effects** on accuracy judgments[4,5,6]

[1]Pennycook 2017 [2]Clayton 2019 [3]Pennycook 2018 [4]Gao 2018 [5]Ross 2018 6[8]Seo 2019

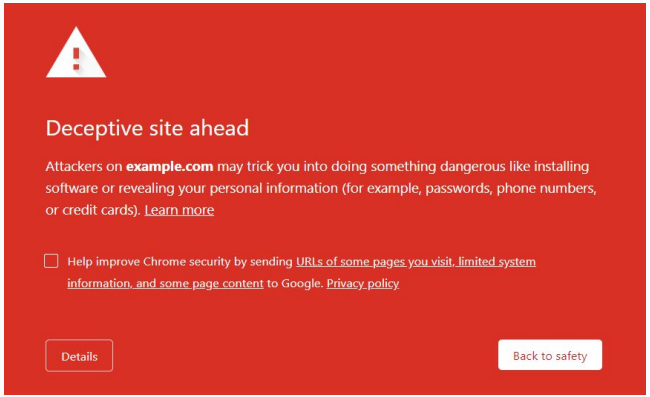
Browser Security Warnings

Goals

- Protect against **phishing**, **MITM**, **malware**, and other threats
- Retain **user choice**, which is important because of **false positives**

Research

- Clickthrough rate (CTR) is the key metric
- Early studies found **high CTR (~70%)**[1,2]
- Methods evolved from surveys to supervised tasks to field studies
- Modern warnings achieve **10-25% CTR**[3,4]



Google's interstitial warning for flagged sites [4].

Relevant findings

- Warnings must be **noticeable**, **credible**, and **motivating**
- Experimental tasks must be **realistic**
- **Interstitials** >> contextuals[1,5]

[1]Wu 2006 [2]Schecter 2007 [3]Akhawe 2013 [4]Reeder 2018 [5]Egelman 2008

Research Goals

Empirically evaluate **interstitial** and **contextual** disinformation warnings

- Will users **notice** and **understand** the warnings?
- Will users **change their behavior** after seeing the warnings?
- What **messaging strategies** are most effective at changing user behavior?

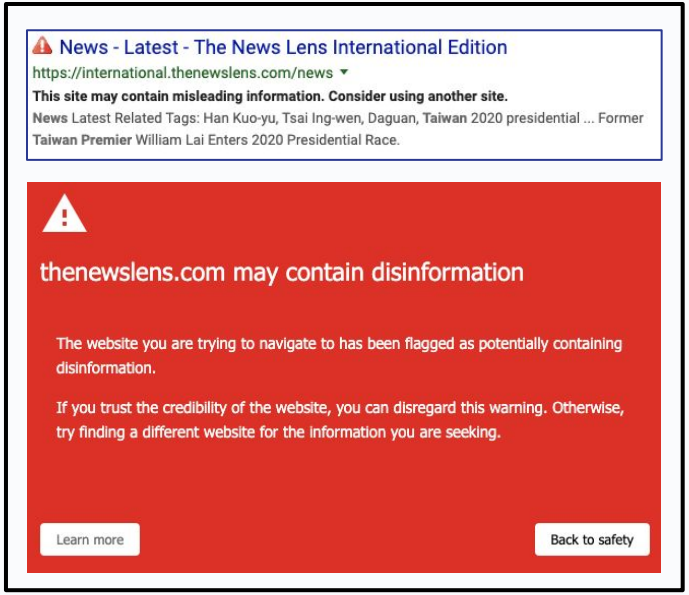
Qualitative Laboratory Study (*n* = 40)

Methods: think-aloud role-playing tasks & interviews

- 4 search tasks using Google Search & Chrome
- 2 control rounds & 2 treatment rounds, with either **contextual** or **interstitial** warnings
- **Primary** and **alternative** sources specified for each task

Data

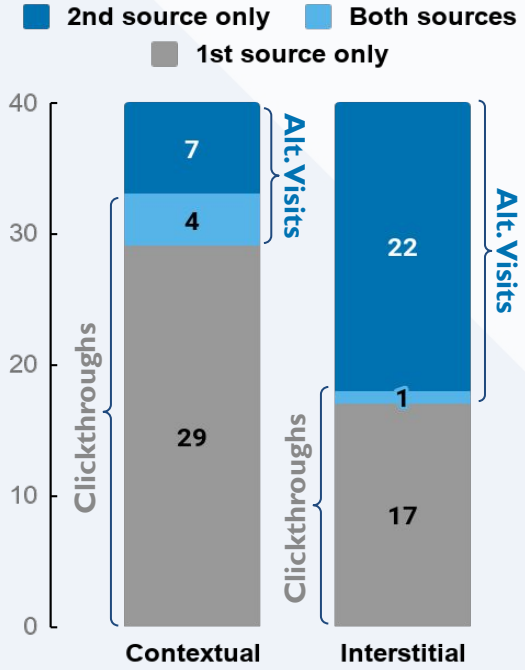
- Researchers' notes
- **Clickthrough rate** (CTR)
- A new metric: **alternative visit rate** (AVR)
- Follow-up interviews



The **contextual** warning (top) is adapted from the Google Search inline warning. The **interstitial** warning (bottom) is adapted from the Google Chrome SafeBrowsing warning.

Laboratory Results

Behaviors



Notice & Comprehension

Contextual (*n* = 20)

- 4 subjects did not notice the warning
- 9 more saw the icon but not the text

Interstitial (*n* = 20)

- 8 subjects did not realize the warning was about disinformation
- 7 of 8 still chose to go back

Takeaways

- The **interstitial** was **noticeable, comprehensible, effective**
- ~½ of AVs were subjects who comprehended the warning
- 3 **mechanisms of effect** emerged:
  - Informativeness
  - Fear of harm
  - Friction

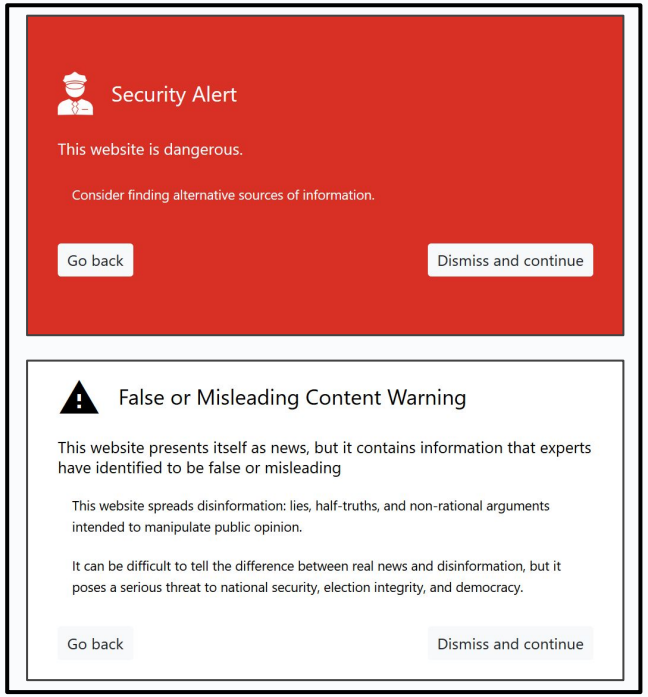
Quantitative Crowdworker Study

<u>Goals</u>

- Validate effect of **interstitial** warnings
- Identify **informative** & **threatening** designs and compare effect sizes
- Examine moderating effect of partisanship

<u>Methods</u>: search tasks & surveys (*n* = 238)

- 4 tasks using **simulated** search tool
- 8 warning designs; 4 for each theory of effect
- Treatments **adaptively** assigned
- **Surveys** after warning encounters
- **Bonus payments** for correct answers
- Track clicks to measure **CTR** & **AVR**



We adapted warnings from Google Chrome. For **harm** (top), we used the SafeBrowsing warning. For **informativeness** (bottom), we used the SSL warning.

Key Findings

- Participants were **significantly more likely** to visit alternative sources after seeing an interstitial warning
  - $z = 22.44$
  - $p < 0.001$
- Participants **reliably understood** our informative warnings
- Informativeness & harm scores had **no significant correlation** with AVR

	Liberal					Conservative				
	#	AVR	CTR	$\bar{i}$	$\bar{h}$	#	AVR	CTR	$\bar{i}$	$\bar{h}$
Control	318	20%	–	–	–	158	16%	–	–	–
Treatment	318	87%	16%	–	–	158	85%	17%	–	–
Selected treatments										
h1	120	85%	18%	−1.94 ±0.06	1.18 ±0.18	46	83%	17%	−1.91 ±0.11	–
h3	73	84%	18%	–	–	27	81%	22%	–	1.15 ±0.46
i3	39	87%	13%	1.41 ±0.43	–	10	90%	10%	–	–
i4	17	82%	12%	–	−0.76 ±0.69	25	76%	24%	0.88 ±0.69	−0.2 ±0.62

Informativeness ($\bar{i}$) and harm ($\bar{h}$) scores are on an interval scale [-2,2]

Conclusions & Future Work

Conclusions

- Contextual warnings are easy for users to **overlook**
- Interstitial warnings can **effectively communicate** to users and **change behavior**
- Behavioral effects may **not** result from informed decision making

Future Work

- More **behavioral** research on disinformation warning effects
- Large-scale field studies
- Redoubled efforts, transparency, and cooperation by platforms

Acknowledgements

Collaborators: Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer

USENIX organizers, reviewers and our shepherd

Contact: bkaiser@princeton.edu