# Info Interventions

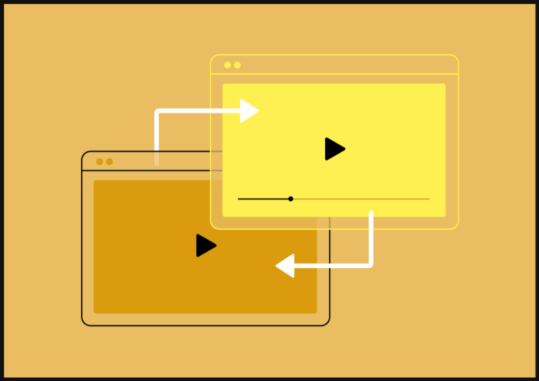
Info Interventions is a set of approaches, informed by behavioral science research and validated by digital experiments, to build resilience to online harms. Learn more at interventions.withgoogle.com.

Jigsaw is a unit within Google that explores threats to open societies, and builds technology that inspires scalable solutions.

## Accuracy Prompts

Accuracy Prompts ask individuals to consider the veracity of a bite-sized piece of content, priming them to remember their own commitment to sharing accurate information when it matters.

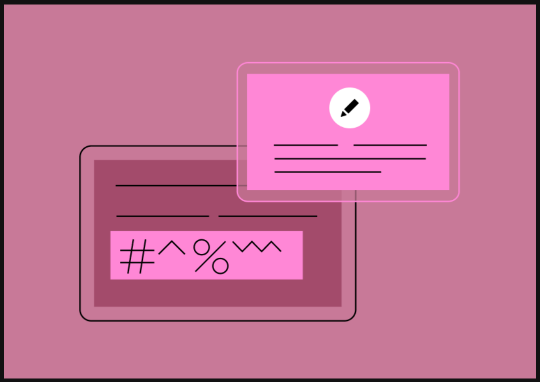
**Refocus User Attention Towards Accuracy**



## Redirect Method

The Redirect Method is a program aimed at reaching individuals who are vulnerable to recruitment by violent extremist groups. The pilot used ads to redirect users looking for extremist information to curated content that refutes ISIS's recruitment messaging.

**Interrupt Online Radicalization**



## Authorship Feedback

Authorship Feedback leverages Perspective API – a tool that uses artificial intelligence to detect toxic language – to provide real-time feedback to commenters who are writing posts by highlighting when their comments might be perceived as offensive.

**Promote Better Conversations**

## Prebunking

Prebunking is a technique to preempt manipulation attempts online. By forewarning individuals and equipping them to spot and refute misleading arguments, they gain resilience to being misled in the future.

**Increase Resistance to Manipulation**

# Accuracy Prompts

## FINDINGS

**50%** Those who received accuracy tips were 50% more discerning in sharing habits versus users who did not. *(Source:Jigsaw)*

**11%** Pre-roll videos on YouTube drove up to an 11% increase in confidence, three weeks after exposure. *(Source:Jigsaw)*

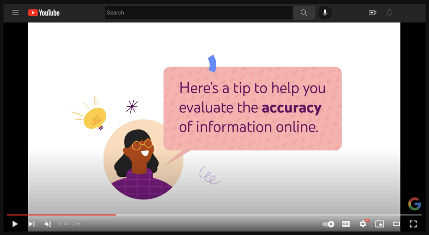
## HYPOTHESIS

Reminding individuals to think about accuracy when they might be about to engage with false information can boost users' pre-existing accuracy goals.

## EXAMPLES

We partnered with MIT and The University of Regina on a series of experiments to test whether accuracy prompts work cross-culturally, reduce sharing of false information, increase the sharing of true information online, and boost users confidence in their abilities to navigate information quality.

Online survey experiments were conducted across 16 countries with +30,000 participants in which they were asked to rate their intention to share true and false news headlines.



Here's a tip to help you evaluate the **accuracy** of information online.

An early prototype accuracy prompt asking users to reflect on the accuracy of a news headline before continuing to browse.

## IN PARTNERSHIP WITH

University of Regina | Hill levene School of Business

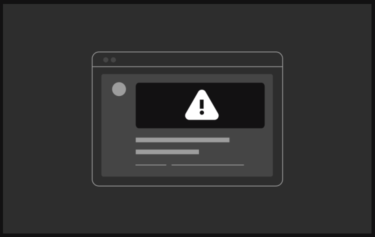
## RESOURCES

Published Study

Medium Blog

Related research

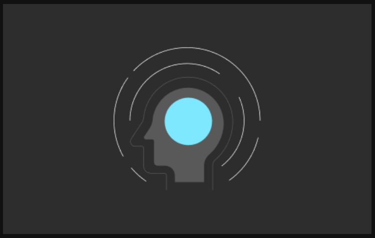
## HOW IT WORKS



1. The individual scrolls through their social feed and comes across content with potential misinformation.

2. An Accuracy Prompt is triggered and pops up over the content.

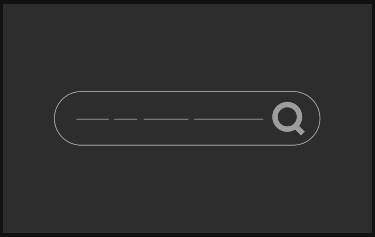
3. A bite-sized explanation on why they are seeing the reminder is served to the individual  and their attention is shifted to the accuracy of the content with information literacy tips.



4. The individual is now prompted to be more aware and may think twice when coming across similar content in their feed.

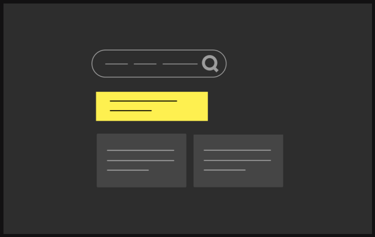
# Redirect Method

## HOW IT WORKS

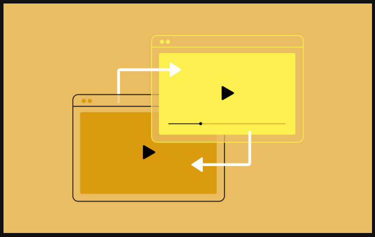


1. The individual completes an online search using keywords that indicate an interest in extremist propaganda.

2. The Redirect Method is initiated and picks up on the keyword to prompt an intervention.



3. An ad is presented to the individual featuring more information on their topic of interest.



4. Upon clicking the ad, the individual is redirected to content that counters false extremist narratives.

## RESOURCES

- Press article (Wired)
- Facebook Redirect Programme
- Canada Redirect Program

## FINDINGS

**320,000** Individuals reached over an 8 week pilot *(Source: Jigsaw & MoonshotCVE)*

**500,000** minutes of counter-narrative videos served. *(Source: Jigsaw & MoonshotCVE)*

## HYPOTHESIS

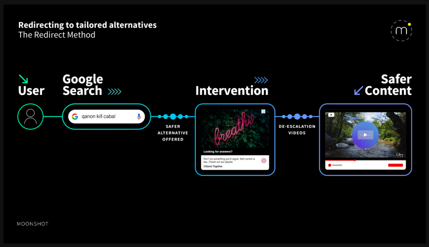
There is a window of opportunity during the radicalization process where individuals who are researching extremist ideologies can be persuaded by narratives refuting them.

## EXAMPLES

Jigsaw and Moonshot developed The Redirect Method's open-source methodology from interviewing ISIS defectors about the role of the Internet in their radicalization process. The insights informed the design of a pilot program using AdWords to reach people at risk of radicalization and used content to serve them with relevant counter-narratives.

The content was uploaded by users from all around the world to confront online radicalization, and selected by expert practitioners. Focusing on the slice of ISIS' audience most susceptible to its messaging, our methodology recognizes that even content not created for the purpose of counter-messaging can still undermine harmful narratives when curated, organized and targeted effectively.

Since 2016, Moonshot has partnered with an array of technology companies, including Facebook, to deploy advertising to those expressing an interest in other online harms, including white supremacy, violent misogyny, and conspiracy theories.



**Redirecting to tailored alternatives**
The Redirect Method

User

Google Search

qanon kill cabal

SAFER ALTERNATIVE OFFERED

Intervention

Looking for answers?

Don't do something you'll regret. Self-control is key. Check out our playlist.

Citizens Together

DE-ESCALATION VIDEOS

Safer Content

MOONSHOT

A campaign flow showing Moonshot using the Redirect Method to redirect individuals to safer content, in this case, services to exit white supremacist movements.

## IN PARTNERSHIP WITH

moonshot

# Authorship Feedback

## FINDINGS

34%

of users who received feedback powered by Perspective API chose to edit their comment. *(Source: Jigsaw)*

## HYPOTHESIS

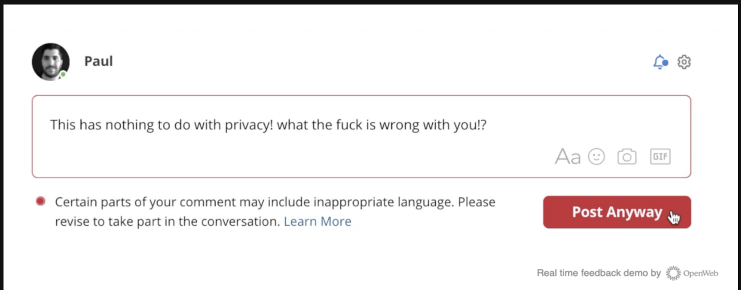
Giving users a moment to pause, reflect, and consider different ways of phrasing their comments can contribute to better conversations online.

## EXAMPLES

We partnered with several websites and developed a feature that directly integrated into comment publishing systems. As users typed comments, their text was checked through Perspective API before publishing the comment.

If the comment exceeded a predetermined threshold of toxic language measured by Perspective API, the user was offered a reminder and opportunity to rephrase their comment and try again.

Post-hoc analysis was conducted on the comments to determine edit rates and overall effect.



Paul

This has nothing to do with privacy! what the fuck is wrong with you!?

Certain parts of your comment may include inappropriate language. Please revise to take part in the conversation. Learn More

Authorship feedback message shown in red below a toxic comment on one of the websites supported by OpenWeb.

## IN PARTNERSHIP WITH

OpenWeb Coral BY VOX MEDIA

## RESOURCES

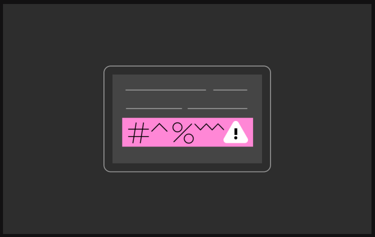
Medium blog

OpenWeb blog

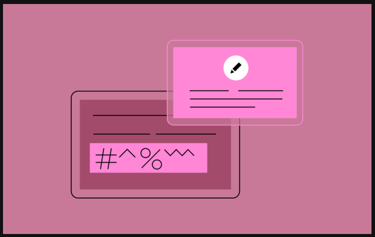
Related Research

## HOW IT WORKS

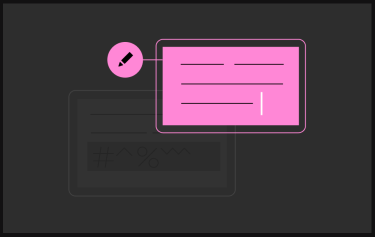
1. The individual writes a comment that is identified as "toxic" - a rude, disrespectful or unreasonable comment that is likely to make someone leave a discussion.



2. Perspective API picks up on the "toxic" comment using machine learning models that identify abusive language.



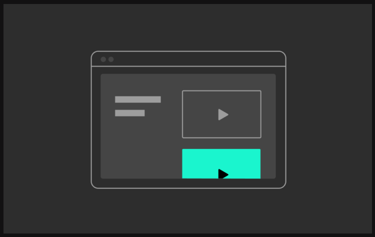
3. An authorship feedback message is shown, alerting the individual that their comment has been identified as risky/offensive or is misaligned with the publisher's Community guidelines.



4. The individual is encouraged to adjust the language before publishing their comment.

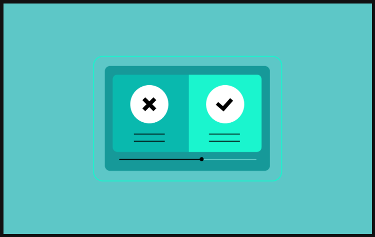
# Prebunking

## HOW IT WORKS

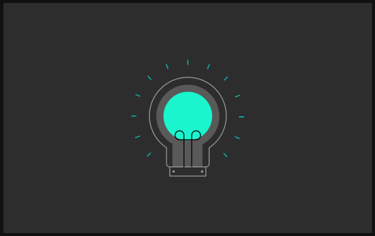


1. A prebunking video is served to a group of users as an ad in their social media feed.

2. Through short video messages the individual is informed of possible attempts to manipulate them online.



3. The individual is shown a relevant example of a manipulative technique or narrative and then given counter arguments to refute the claim.



4. By analyzing how well video viewers recall the techniques in a short survey relative to a control group, we can assess their likelihood to resist manipulative content in the future.

## RESULTS

73% of individuals who watched a prebunking video were more likely to consistently spot misinformation online *(Source: Science Advances)*

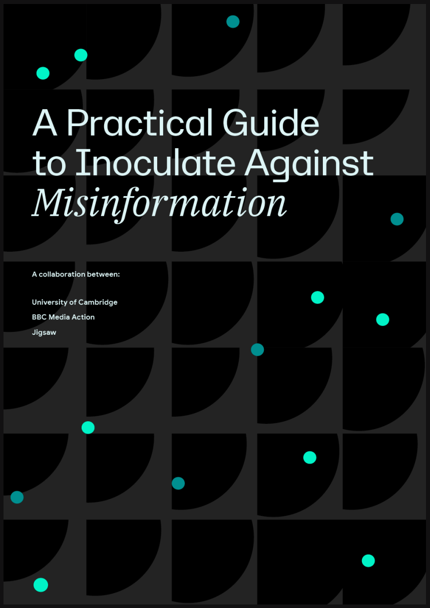
5% Prebunking videos as YouTube ads boosted recognition of manipulation techniques by 5%. *(Source: Science Advances)*

## HYPOTHESIS

Preemptive messages can help individuals identify manipulative narratives and strategies (e.g. the claims that “vaccines are unnatural” or that refugees steal jobs).

## Approach & Examples

Through lab testing and live experiments alongside academic partners at University of Cambridge, Harvard University, and American University, we \tested prebunking via short video messages to promote resistance to common rhetorical strategies and narratives that are used to perpetuate misinformation.



A Practical Guide to Inoculate Against *Misinformation*

A collaboration between:

University of Cambridge
BBC Media Action
Jigsaw

BBC Media Action, University of Cambridge, and Jigsaw developed a “how-to-prebunk” guide, to give practitioners guidelines and basic requirements to produce their own prebunking messages. Download the PDF

## IN PARTNERSHIP WITH

UNIVERSITY OF CAMBRIDGE

BBC MEDIA ACTION

# Prebunking

RESOURCES

Published study

Medium blog

Related research

Demagog, NASK and Jigsaw created a series of videos to counter anti-refugee narratives about Ukrainians living in Central & Eastern Europe. Watch all videos on YouTube.

IN PARTNERSHIP WITH

NASK DEMAGOG JSNS.CZ

In collaboration with the University of Cambridge and the University of Bristol we created 5 videos to prebunk particular manipulation techniques commonly encountered online. Watch all videos on YouTube.

IN PARTNERSHIP WITH

UNIVERSITY OF CAMBRIDGE University of BRISTOL

In partnership with scholars at Harvard T.H. Chan School of Public Health, American University, and trained medical professionals we preemptively corrected common misleading narratives about vaccine Watch all videos on YouTube.

IN PARTNERSHIP WITH

POLARIZATION & EXTREMISM RESEARCH & INNOVATION LAB

YouTube's global media literacy program Hit Pause helps teach viewers in markets around the world the skills to detect misinformation. The initial set of creative builds off Jigsaw's work to prebunk common manipulation techniques such as emotional language. Watch all videos on YouTube.

CREATED BY

YouTube

IN PARTNERSHIP WITH

NAMLE National Association for Media Literacy Education