



Prebunking Against Online Hate

Prebunking combines real-time narrative monitoring, rigorously tested interventions, and education into a single prevention pipeline that teaches people to recognize the manipulation techniques behind hateful narratives before those narratives become widely accepted.

THE CHALLENGE

Hateful narratives evolve fast. Antisemitic conspiracies, for example, often begin on the fringes of the internet, get repackaged as jokes or innocent-sounding questions, and drift into mainstream feeds, where they are normalized with every share. By the time a narrative spreads, it has done lasting harm: feeding harassment, deepening polarization, and in the worst cases, helping to justify violence.

To address this challenge, PERIL takes a proactive, evidence-based approach called “**prebunking**” (also known as attitudinal inoculation). Like a vaccine that prepares the body to fight an infectious disease before exposure, short inoculation messages teach people to recognize the manipulation techniques behind hateful narratives before they encounter them on their own. Because the approach targets tactics rather than beliefs, it is nonpartisan and builds a durable critical-thinking skill that protects against all forms of hate.

THE GUIDING QUESTION

Can we build population-level resilience to the manipulation techniques used in hateful narratives, and monitor, test, scale, and sustain it? Prebunking is designed to do exactly that.

HOW PREBUNKING WORKS

A short intervention (a video, post, exhibit, or game) does for a manipulative narrative what a vaccine does for a virus.

A VACCINE DOES THIS	PREBUNKING DOES THIS
EXPOSURE Introduces a weakened dose of a virus, too little to cause illness.	EXPOSURE Shows a weakened version of a harmful narrative, clearly explained as manipulation.
RECOGNITION The immune system learns the signature of the pathogen.	RECOGNITION The audience learns the tactic and how to identify when it is being used.
DEFENSE Antibodies are produced and held in reserve.	DEFENSE Emotional antibodies are developed, preparing the audience to reject similar attempts at manipulation.
PROTECTION The real pathogen is neutralized on contact.	PROTECTION The tactic loses its effectiveness the next time it appears in a real post.

Prebunking messages follow this tested structure every time. The approach rests on **inoculation theory** and a substantial body of rigorously evaluated research.

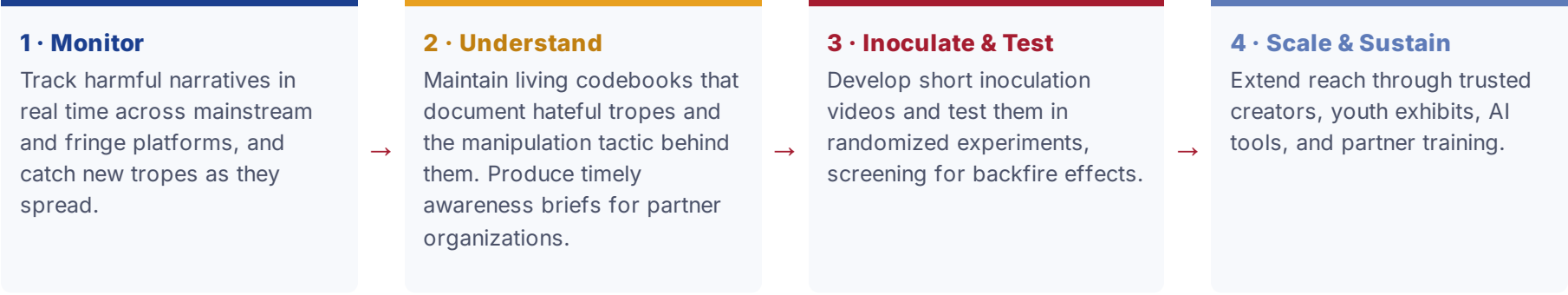
WHY PREBUNKING MATTERS

Two things set prebunking apart from the usual response to harmful content. First, prebunking targets manipulation techniques, not political opinions or ideological positions. It does not tell people what to think. It helps them notice when someone is trying to manipulate them, which makes it nonpartisan and effective across different audiences and belief systems.

Second, it works before exposure rather than after. A large body of research shows that arguing with people about a belief they already hold often backfires, leaving them more committed to hateful narratives than before. Reaching them first, and equipping them with the tools to recognize manipulation, carries no such risk. Prebunking focuses on the period before a narrative takes root, and on the people most likely to be swayed by it, rather than trying to undo the damage after the harmful content has spread.

AN INTEGRATED PREBUNKING PIPELINE

PERIL’s prebunking initiatives are not standalone projects. They form a connected pipeline where each stage supports the next: from spotting a narrative as it emerges on social media to inoculating the public against it and sustaining that protection over time.



THE EVIDENCE

Prebunking has a rigorously validated body of results dating back to the 1960s. PERIL’s own prebunking work has demonstrated the effectiveness of the approach:



WHAT SETS THIS WORK APART

Rigorous and Honest

Every prebunking intervention is tested in **randomized experiments** and screened for backfire effects before release. If an intervention doesn't work, it is redone. Evidence guides the work.

Built to Scale

AI presenters, trusted content creators, partner training, youth exhibits, and drafting tools extend reach far beyond what any single lab can produce, turning a proven method into **public prevention infrastructure**.

A Transferable Skill

By teaching people to spot manipulation tactics — not what to think — the work is **nonpartisan**, carries across narratives, and builds resilience to multiple forms of hate at the same time.