Published September 2023

Democracy By Design

A Content-Agnostic Election Integrity Framework for Online Platforms

+accountable tech
CAP
cdt
Center for Humane Technology
Common Cause
epic.org ELECTRONIC PRIVACY INFORMATION CENTER
ISD Institute for Strategic Dialogue
issue one FIX DEMOCRACY FIRST
PUBLIC KNOWLEDGE
Ranking Digital Rights

Overview

With over 50 countries set to hold elections amidst tectonic geopolitical and technological shifts, 2024 will present unprecedented challenges for democracy and the information gatekeepers who support it.

Those challenges are compounded by increasing volatility of the content policy landscape, as partisan debates over where platforms should draw the lines grow more bitter, and repressive regimes engage in draconian censorship under the guise of countering disinformation.

But there is a clear path forward to avoid these pitfalls, build consensus, and protect both freedom of expression and free and fair elections – a path to build *Democracy By Design*.

Below, we outline a set of high-impact, content-agnostic election integrity recommendations for online platforms – readily actionable interventions rooted in their own product design and policy toolkits, and in many cases, empirically backed by their own research.

The *Democracy By Design* framework includes three major planks, under which there are specific recommendations and background to support them:

- Bolstering Resilience, which focuses on 'soft interventions' that introduce targeted friction and context to mitigate harm;

- Countering Election Manipulation, which outlines bulwarks against evolving threats posed by malign actors and automated systems; and

- Paper Trails, which highlights key transparency measures needed to assess systemic threats, evaluate the efficacy of interventions, and foster trust.

The result is a framework meant to avoid political landmines and broadly resonate across nations with distinct laws and cultures, and platforms with incongruous architecture and resources – a consensus roadmap to enhance systemic resilience against election threats.

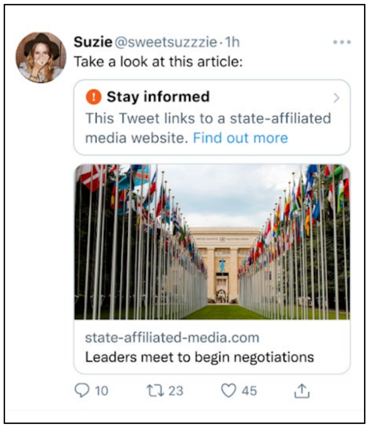
Bolstering System Resilience | *Targeted Friction and Context to Mitigate Threats*

Election integrity vulnerabilities on social platforms often stem from their own architecture: The features designed to make platforms frictionless and maximize engagement – from recommendation algorithms to reshare buttons – can serve to warp discourse and undermine democracy.

Extensive research from experts and tech companies themselves indicates that content-agnostic soft interventions that introduce targeted friction and context – bolstering the resilience of their systems and better informing users – can significantly mitigate threats.

1. **Employ well-designed interstitials when users engage with things like old articles or state-controlled media.** Election authorities provide voters with official information about casting their ballots; polling places have check-in stations and safeguards against foul play; voting machines prompt voters to double-check their ballots before submitting them. Platforms should similarly work to protect and inform voters by injecting targeted friction and context into their systems via well-designed interstitials (click-through screens) and labeling. Examples of such tools – some of which have already been implemented across various platforms – include:

   - Pop-ups asking users if they want to read an article before sharing, or alerting users if articles are old
   - Clearly labeling accounts of government officials and using interstitials to alert users who try to read or share content from state-run media
   - Placing posts that contain misleading or unverified election information behind click-through warning labels that include clear context and facts
   - Appending distinct verification badges to the official accounts of local, state, and national election authorities



Suzie @sweetsuzzzie · 1h

Take a look at this article:

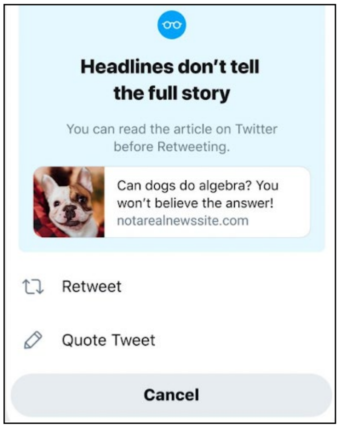
Stay informed

This Tweet links to a state-affiliated media website. Find out more

state-affiliated-media.com

Leaders meet to begin negotiations

Twitter state-affiliated media label



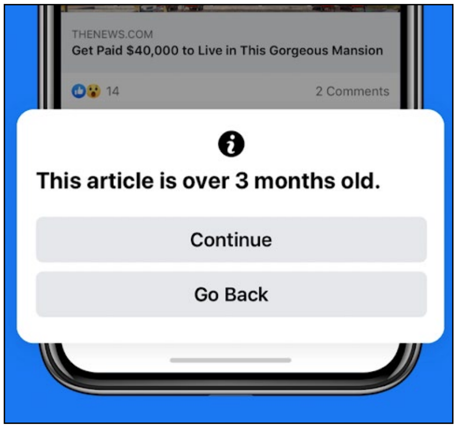
Headlines don't tell the full story

You can read the article on Twitter before Retweeting.

Can dogs do algebra? You won't believe the answer!

notarealnewssite.com

Twitter prompt to read before RT



THENEWS.COM

Get Paid $40,000 to Live in This Gorgeous Mansion

This article is over 3 months old.

Facebook interstitial when sharing old article

2. **Utilize virality circuit breakers to automatically flag fast-spreading posts and trigger a brief halt on algorithmic amplification.** Not all high-reach content comes from users with huge followings – especially now, as platforms increasingly rely almost entirely on algorithms in determining what to surface with little friction. Inspired by automatic triggers used in financial markets to prevent panic selling at moments of high volatility, experts have proposed that platforms introduce circuit breakers to automatically flag posts that are beginning to gain virality, temporarily pause algorithmic boosting, and present users an interstitial about the post's virality and status upon click-through or reshare. It's a proposal that's been lent credence by tech companies themselves, with Meta reportedly testing the concept and Snap going a step further, as an executive testified that all posts on their Spotlight platform were speed-bumped and checked before reaching 25 unique viewers. Proper thresholds and processes will of course vary by platform and content type, so companies should outline their policies and publish pertinent aggregate data in transparency reports.

3. **Restrict rampant resharing during election season by removing simple share buttons on posts after multiple levels of sharing.** Frictionless resharing is a staple of social platforms – and a key driver of toxicity. Internal Meta research showed users are *4x* more likely to encounter falsehoods in a reshare of a reshare than in the News Feed in general, and concluded aggressively limiting these 'deep reshares' would be "an effective, content-agnostic approach to mitigate the harms." Meta also placed limits that proved effective on how many times WhatsApp messages could be forwarded after mass-sharing exacerbated unrest in India and Brazil. Platforms should remove share buttons on posts after multiple levels of reshare, and/or test other mechanisms that enhance reshare friction in a targeted manner during election season, with careful consideration of the impact on legitimate advocacy campaigns.

4. **Implement clear strike systems to deter repeat offenses, curtail the outsized impact of malign actors, and better inform users.** This is not meant to dictate specific content policies; strike systems should be based on violations of platforms' own standards. Most platforms already utilize some form of strike system to levy sanctions on repeat offenders in recognition of the disproportionate harm they drive, but both the policies and their application are typically vague and/or obscured from users – often under the argument that it's impossible to clarify such rules without helping bad actors game the system. And even approaches that have more explicitly addressed this threat – like Twitch's Harmful Misinformation Actor policy, and Twitter's bygone 5-strike Civic Integrity policy – have focused chiefly on when to suspend the worst actors. Platforms should develop and implement transparent strike systems that clearly outline escalating 'soft interventions' to limit the impact of repeat offenders, such as restricting resharing, curtailing algorithmic amplification, and placing posts behind click-through warning labels with context. This approach steers clear of the false choice between censorship and inaction, and would demystify enforcement decisions, deter habitual rule-breaking, and defang the malign actors who pose the greatest election integrity threats. A detailed example of what such a strike system might look like for a given platform, developed by Accountable Tech, can be found here.

Countering Election Manipulation | *Safeguards Against Malign Actors and Automated Systems*

Malign actors have weaponized social platforms to meddle in elections, attack democracy, and erode our shared reality. Now their capacity for manipulation has been turbocharged by new technology, including powerful algorithms and generative AI tools tailormade for high-impact, low-cost influence operations.

Platforms must do everything in their power to thwart unlawful efforts to interfere with elections or individuals' free exercise of their right to vote – including efforts to intimidate voters or mislead them on how to participate – and to counter manipulation more broadly.

1. **Prohibit – in policy and practice – the use of generative AI or manipulated media to:**

   a. ***Falsely depict election irregularities.*** In recent elections, we've seen users intentionally and unintentionally share images and videos that have been taken out of context in ways that appear to show election irregularities (i.e. burning ballots), undermining faith in the legitimacy of the democratic process. Using tools like reverse-image search in previous election cycles, researchers were often able to trace the provenance of content relatively quickly and neutralize false narrative – but if new AI-generated hoaxes of this nature flood the zone, without the same capacity to identify them, the harms to democracy could be severe. Platforms should establish clear policies prohibiting the false depiction of election irregularities and mechanisms for meaningful enforcement and deterrence.

   b. ***Fraudulently misrepresent the speech or actions of public figures in video, audio, or images.*** The deepfake threat has loomed for years, but generative AI has drastically lowered the barrier to entry for anyone – be it political opponents or malign actors – to convincingly put words into public figures' mouths and fool voters or hijack the discourse. While many platforms already have baseline policies addressing synthetic or manipulated media, they are often overly vague. Platforms should expound upon them, clearly detailing – including example cases – the scope of these policies, and should clarify that any AI-generated or -manipulated content likely to harmfully deceive voters about public officials' speech or actions is strictly prohibited. (Such a policy might narrowly exempt, for example, clearly labeled parody or satire.)

   c. ***Create personalized political or issue ads that microtarget voters with distinct content generated by using their personal data.*** Even before the recent explosion in generative AI, concerns have been ballooning over the manipulative nature of hyper-targeted personalized advertising – particularly in the political sphere. The EU is currently debating strict new limitations on how political ads can be targeted. Google has similarly restricted political ad targeting categories to age, gender, and general location. The threats are even more harrowing when considering the potential for generative AI ad tools to be deployed to serve bespoke ads to exploit individual voters based on their behavior, identity, or even inferences about their current mood. With the dominant players all already touting new generative AI ad products, it's critical that platforms strictly prohibit the use of such tools to create personalized political or issue ads that microtarget voters with distinct content generated by using their personal data. Additionally, they should maintain robust ad transparency libraries that shed light on ad targeting and variations, including cataloging any ads removed for violations.

2. **Implement strong provenance standards and require clear disclosures within any political or issue ad that features AI-generated images, video, or audio.** Given the breakneck speed at which generative AI and synthetic media tools are being incorporated into every phase of content creation, all participants in the online information ecosystem have an urgent responsibility to empower users by providing as much transparency as possible about the provenance and authenticity of content, disclosing when and how video, images, or audio has been generated or manipulated. Leading technology and media companies are already working to establish global, interoperable protocols to do just that, but the efficacy of such efforts will depend on widespread adoption; platforms should commit to working collaboratively and implementing strong provenance standards across the digital sphere, and require clear disclosures to be included within any political ad utilizing AI-generated or -manipulated media.

3. **Prompt users as election season[1] starts, outlining the main parameters of core algorithmic systems and giving users autonomy to adjust ranking criteria so their recommendations are not based on profiling or optimized for engagement.** One of the greatest threat vectors for large-scale manipulation and harm, as shown by platforms' own internal research, is the black box recommendation and curation systems that have increasingly usurped users' autonomy in determining what content they see. This threat has come into sharp focus as nations grapple with fears that the Chinese Communist Party could effectively weaponize the powerful TikTok algorithm to reshape public discourse and undermine democracy in critical moments without detection. Echoing bipartisan US legislation and proposals from European leaders, platforms should prompt users as election season starts, outlining the main parameters of core algorithmic systems and giving users autonomy to adjust ranking criteria so their recommendations are not based on profiling or optimized for engagement, along with easily turning off non-essential discovery tools. For example, users might opt to revert to a reverse-chronological feed of accounts they follow, or to prioritize content from authoritative news sources or diverse viewpoints. All user choices should be respected for the duration of election season, at a minimum.

Paper Trails | *Fostering Trust through Meaningful Transparency*

Just as machine voting systems are backed by paper trails to ensure our elections are trustworthy and secure, platforms that shape the information ecosystem must finally open up the black box and start showing their work. The status quo has yielded distrust from all sides, with partisans suspecting foul play, neutral observers unable to make informed evaluations, and non-English speakers left behind.

Platforms must begin disclosing meaningful data to voters, independent researchers, and election officials regularly in order to engender trust and substantiate integrity measures.

[1] For the purposes of this framework, "election season" refers to the period beginning 60 days before Election Day and ending after the peaceful transfer of power, and should be considered a baseline, as elevated risk may extend further.

1. **Clearly and accessibly detail in one place all election-related (or applicable platform-wide) policies and approaches.** Platforms have developed complex and constantly evolving approaches to countering election-related threats. But it's hard to keep track of all the policies, which can be rolled out via disparate press releases or posts from spokespeople. And enforcement can be uneven and opaque, sowing distrust and confusion among affected users. Platforms should ensure that all pertinent policies are distilled into one easily accessible and intelligible election policy hub. And users should be notified if they are subject to any enforcement action in clear language explaining the rule that is being applied and its implications, and given the opportunity to appeal.

2. **Release regular transparency reports during election season – broken down across widely spoken languages – detailing high-performing and violative content, enforcement, and resource allocation.** While many platforms now release transparency reports, they tend to be quarterly or biannual English-language-centric products with company-selected categories of data and metrics – not standardized, frequent, or granular enough to illuminate the state of the information ecosystem around elections, nor the efficacy of interventions to mitigate them. Platforms should commit to releasing reports as frequently as possible during election season that include the following information:

    - A snapshot of the week's highest-performing election-related content – the posts, advertisements, accounts, URLs, and groups that generated the most engagement

    - Aggregate data about enforcement of all election-related policies, including specifics about the measures detailed in previous sections and their effectiveness

    - Details about the number of full-time staff and contractors working on election integrity, broken down by department role and primary languages spoken

3. **Provide independent researchers with direct access to platform data to inform studies, threat analysis, and systemic impact assessments.** The widespread distrust of tech companies' impact on our elections, our discourse, and our democracy will persist so long as platforms continue to be black boxes, with the public kept in the dark on everything from the basic inputs of their recommendation algorithms to the enforcement of their content policies. While bipartisan legislation in Congress to remedy this has not yet passed, the EU's Digital Services Act is set to finally give researchers in Europe access to some of that data so they can study the systemic impacts of the products, services, and features at the center of our information ecosystem. Companies must grant direct and meaningful access to platform data – in a responsible and privacy-preserving manner – to independent researchers in the US and globally such that they can analyze and help deter election threats in real-time and summarize their findings for the broader public.

Democracy By Design | *Summary*

BOLSTERING SYSTEM RESILIENCE

Targeted Friction & Context to Mitigate Threats

Election integrity vulnerabilities on social platforms often stem from their own architecture: The features designed to make platforms frictionless and maximize engagement – from recommendation algorithms to reshare buttons – can also serve to warp discourse and undermine democracy.

Extensive research from experts and tech companies themselves indicates that content-agnostic soft interventions that introduce targeted friction and context – bolstering the resilience of their systems and better informing users – can significantly mitigate threats. *Platforms should:*

- **Use well-designed interstitials** when users engage with things like old articles or state-controlled media
- **Utilize virality circuit breakers** to automatically flag fast-spreading posts and trigger a brief halt on algorithmic amplification
- **Restrict rampant resharing** during election season by removing simple share buttons on posts after multiple levels of sharing
- **Implement clear strike systems** to deter repeat offenses, curtail the outsized impact of malign actors, and better inform users

COUNTERING ELECTION MANIPULATION

Safeguards Against Malign Actors & Automated Systems

Malign actors have weaponized social platforms to meddle in elections, attack democracy, and erode our shared reality. Now their capacity for manipulation has been turbocharged by new technology, including powerful algorithms and generative AI tools tailormade for high-impact, low-cost influence operations.

Platforms must do everything in their power to thwart unlawful efforts to interfere with elections or individuals' free exercise of their right to vote – including efforts to intimidate voters or mislead them on how to participate – and to counter manipulation more broadly. *Platforms should:*

- **Prohibit, in policy and practice, the use of generative AI or manipulated media to:**
  - Falsely depict election irregularities
  - Fraudulently misrepresent the speech or actions of public figures in video, audio, or images
  - Create personalized political or issue ads that microtarget voters with distinct content generated by using their personal data
- **Implement strong provenance standards** and require clear disclosures within any political or issue ad that features AI-generated images, video, or audio
- **Clearly present users with the main parameters of algorithmic systems** as election season starts, and prompt them to opt in

PAPER TRAILS

Fostering Trust Through Meaningful Transparency

Just as machine voting systems are backed by paper trails to ensure our elections are trustworthy and secure, platforms that shape the information ecosystem must finally open up the black box and start showing their work. The status quo has yielded distrust from all sides, with partisans suspecting foul play, neutral observers unable to make informed evaluations, and non-English speakers left behind.

Platforms must begin disclosing meaningful data to voters, independent researchers, and election officials regularly in order to engender trust and substantiate integrity measures. *Platforms should:*

- **Clearly and accessibly detail in one place** all election-related (or applicable platform-wide) policies and approaches
- **Release regular transparency reports during election season** – broken down across widely spoken languages – detailing high-performing and violative content, enforcement, and resource allocation
- **Provide independent researchers with direct access to platform data** to inform studies, threat analysis, and systemic impact assessments