Elections integrity best practices
Defining and achieving success

Election Integrity Series Part 2
2023-10-04
About the Integrity Institute

The Integrity Institute is a non-profit working to advance the theory and practice of protecting the social internet, powered by our community of integrity professionals.

With years of experience mitigating harms to people and communities within more than 40 online platform companies, we bring seasoned, insider knowledge to leaders theorizing, building and governing online platforms and help them put integrity front and center.

Here’s how we do it:

- We build and empower a community of integrity professionals in tech, giving them the tools and research they need to make online platforms safer and healthier for people and societies.
- We advise online platforms, policymakers and academics to put integrity at the heart of company governance, compliance and tech regulation.
- We educate the public about what an integrity-first future looks like for the social internet.
Introduction

- This deck is part 2 of the Integrity Institute's elections integrity best practices series.
- Part 1 gave an overview on responsibly supporting elections on online platforms.
- Part 2 expands with more depth on:
  - Setting goals, defining and measuring success, and readiness
  - A typology and examples of product design considerations and interventions
  - Working with external parties, for various purposes and in various contexts
  - Discussions of critical topics, e.g. generative AI’s impact on elections and types of abusive actors
- We hope this guide will provide ideas, frameworks and knowledge that inform how platforms can address election integrity risks.
Contents

- Preparation and assessment
  - Setting goals
  - Defining and measuring success
  - Success and monitoring metrics
- Interventions and product levers
  - Interventions
  - Product levers
- Working with external stakeholders
  - Research collaboration
  - Interacting with elections authorities globally and in the US
  - Working with politicians and governments
  - Escalation pathways
- Critical topics
  - Generative AI
  - Abusive actor taxonomy
  - Classifiers
- Conclusion: Impossible trade-offs
Preparation and assessment

Setting goals, defining and measuring success, and monitoring trends
Setting goals

Every platform should explore the role they play and the experiences they want people to have — and not have in elections contexts.

Key factors in each election context (see part 1 for more) include:

- How legitimate the election is, including the country’s electoral infrastructure
- The strength of social institutions and risks of violence
- How the platform is used in the country
- How strong the organization’s capabilities and network are in the country

From this, platforms must determine what they can and should accomplish, considering risks, constraints and trade-offs:

- Maybe this is focusing on a readiness assessment and being prepared for what’s to come;
- Maybe a platform is in position to set goals to mitigate specific harms (e.g. disenfranchisement);
- Mitigations may be in order for risks to users; to society; and/or to the platform itself (e.g. reputation or regulatory);
- And maybe platforms can set goals that maximize positive outcomes (e.g. voter engagement).
Creating measurable and achievable goals

- Clear, actionable goals help organizations identify risks, align priorities and strategies, and then hold themselves accountable.
- Goals should be SMART (specific, measurable, achievable, relevant and time-bound).
- Measurable goals can be both qualitative and quantitative, for example:
  - Preparedness goals are often more qualitative - e.g. checklists or traffic light charts. They might draw on questions around internal organization and ownership; systems and processes; red team reports and risk assessments; local expertise and language support; policy readiness; infrastructure (classifiers, measurement, operations).
  - Goals around mitigating harms can be more quantitative, e.g. making sure prevalence (and perceived prevalence) of violating content is minimized and acceptable.
  - Goals around maximizing good can be both qualitative and quantitative - e.g. informing voters and increasing participation; being viewed as responsible and playing a positive role; increasing engagement.
- Weigh up competing priorities and be realistic about what can and can’t be achieved with the available resource. And maybe this exercise will show the need and ROI for increasing those resources!
Defining success informs what to measure

- If success goes beyond minimizing harm, measurement should go beyond counting violating or harmful content.
- Know that harms will always be present at some level and prepare for that reality.
- Also measure the experiences you want people to have, and the role your platform plays in the election.
- You may need to measure the overall state of the information ecosystem, e.g. the amount of civic content:
  - Which of it is explicitly political (discussing candidates or the election itself)
  - Which discusses political issues, including issues that aren’t inherently political but have elevated importance
- It’s useful to define what you consider high quality content and content producers, which might include things like:
  - Whether the account is transparent about who they are and what they’re trying to achieve
  - Whether the content is from a trusted, authoritative source (you may want external guidance on this)
  - Algorithmic measures of user-identified quality - the content’s importance in the ecosystem (e.g. PageRank)
  - Whether the content is violating or harmful, or goes against the company mission or values
  - Whether it’s clickbait, from an anonymous source or unoriginal content
As discussed in part 1 of our guide, elections are pivotal moments that test the robustness of existing practices and processes - very much including metrics.

Platforms should evaluate how well they already follow best practices for measuring success, in particular around mission and values:
- Measuring long term success, rather than short term success
- Measuring how people get utility and value from the platform
- Measuring and value mission-aligned engagement and growth, not just raw engagement

When election success goals and metrics are in tension with how the organization define and measure success overall, there will be internal conflict.

Consider, discuss and iterate as needed.

Vigorously communicate internally how the selected election goals and metrics align with and support overall success - but only once alignment is reached.
Defining success metrics for harmful content

- Look to our framework for measuring the lifecycle of harmful content.
- For each type of violating or near-violating content, measure:
  - The number of exposures
  - The number of people exposed
  - Which surfaces the exposure happened on
  - The mechanism of exposure (e.g. algorithmic recommendation, group message)
  - How long was the content on the platform before being moderated
  - Virality factor - i.e. how much exposure violating content got relative to baseline expectations
  - Appeal rate and appeal success rate
Creating user-centered metrics and measures

- Platforms should offer their users ways to flag and report bad experiences. This can help measure the scale of the problem, and give rich insights on users’ preferences and behaviors - for example:
  - How many total flags users are providing
  - How many flags are on civic content
  - How much of the most-flagged content actually violates policies

- We also recommend direct measures (e.g. surveys) of user safety and experience:
  - Try asking about exposure ("have you seen...?") and sentiment ("do you feel that...?").
  - Compare these measures with other metrics (e.g. prevalence of harmful content) to identify possible issues in measurement and definitions - perception and prevalence should generally align.
Inspiration: Example success metrics

Here are a few examples:

▷ Prevalence and frequency of civic misinformation compared to a rolling 28-day average:
  ▷ Both absolute counts and rates of violating content tend to increase during times when politics is especially salient, so setting quantitative targets can be fraught; there’s no magic formula to balance uncertainties and impacts.

▷ Primary/secondary click-through rate (CTR) on civic literacy features (e.g. informational blocks or prompts):
  ▷ The platform might measure how CTR compares with that of advertising content or user-shared links.

▷ Qualitative examples - the platform might consider success as having:
  ▷ Risk and readiness assessments for the specific country
  ▷ Election-related policies in place for the country in local languages
  ▷ Content moderation operations in place

▷ Quantitative examples - the platform might measure:
  ▷ What % of impressions are on civic content
  ▷ What % of impressions on civic content are violating
  ▷ Turnaround time for actioning user-flagged and classifier-flagged content
  ▷ What % of impressions are on high quality content
Monitoring and trend detection

- There’ll always be unknown unknowns; effective monitoring is a platform’s continuous internal alert system, and a way to finding out about harms before they become critical.
- Effective, actionable monitoring requires:
  - A comprehensive view of content on the platform
  - The ability to spot new trends early
  - A process for escalating new trends to the right decision-makers for response
- The comprehensive view of content comes from combining four classes of metrics:
  - How much attention the content gets (e.g. engagement, views)
  - How much the content over-performs against baseline engagement for the creator or space
  - How likely it is to be about a sensitive topic or problematic (content and actor-based classifiers)
  - How likely it is to be about specific individuals or communities
Inspiration: example ways to detect trends

Platforms might measure and monitor:

- **How much attention the content gets (e.g. engagement, views):**
  - This can be done through engagement metrics like views, like, favorites, reactions, replies, shares etc.

- **How much the content over-performs against baseline engagement for the content creator or space:**
  - Understand historical baseline engagement for content from the creator or content in the space - e.g. how much engagement the user’s posts got in the last week or month, or how much engagement posts in a particular group got over the last week or month.
  - Estimate the creator or content space’s importance off-platform - e.g. external traffic (from search engines, other sources, their own domains).
  - Understand the user’s PageRank (major news organizations and people with large platforms tend to have high PageRank).
  - Divide current engagement metrics by estimates - e.g. current engagement divided by baseline engagement, and current engagement divided by external traffic and PageRank.

- **How likely is the content to be about a sensitive topic or problematic:**
  - Content based classifiers: look at whether the content scores high on civic, election or political classifiers, as well as broader classifiers (e.g. hate speech, calls for violence, misinformation).
  - Actor based classifiers: look at whether the content comes from accounts with a history of talking about civic, election or political issues, as well as history of posting violating content.
Finding trends and taking action

- These monitoring metrics are fundamental to detecting and actioning new on-platform trends. Some teams (e.g. trust and safety, product, operations) should be watching them carefully.
- There’s also outside-in monitoring - using tools to monitor external data sources and media about elections and your platform:
  - Someone inside each company probably already has a Google news alert on the platform name
  - Other teams are regularly interacting with external stakeholders (journalists, governments, etc)
  - Distilling election risk planning and goals into a 1-pager and sharing background on the election details (candidates, issues, electoral process) helps teams spot issues they should escalate
- Build an escalation process that directs trends to decision-makers and allows you to take action:
  - This process can uncover policy gaps - e.g. an unforeseen type of harmful content for which the platform has not specified a policy, and is therefore unregulated.
  - It can also uncover process gaps, such as content evolving faster than corresponding detection classifiers - or going viral faster than operations teams can action it.
- Comprehensive monitoring and escalation paths minimize the risk of unaddressed harm – and being caught off-guard by external researchers or the press finding problems before a platform understands and is equipped to respond.
Interventions and product levers

Actions that can be taken when election monitoring surfaces negative or harmful trends on platforms
Platform strategies for elections integrity

 Platforms have built and deployed diverse interventions and product levers to enforce community standards and manage integrity risks across many contexts:

 By interventions, this guide refers to reactive steps taken upon specific examples of platform content or users that violate guidelines, such as content moderation.

 By product levers, this guide refers to proactive choices about product shape that are intended to reduce risks or harms, not sanctioning specific violating behaviors.

 Due to the specific risks and stakes around elections, each intervention or product lever – even ones you already use in other contexts – should be weighed against election-specific considerations.

 Particulars will vary with platforms’ structures, usage and goals; this guide can’t be exhaustive but provides some categories and specific examples. Take what you can, for your needs – either to implement or to spur further thinking.
Introduction to interventions

- Interventions are actions a platform takes to enforce their policies or community standards; generally these actions are viewed by users as sanctions or penalties.
- Interventions generally require three things: a policy, a means to identify violations (e.g. directly by the platform or via a third party), and a process to apply the intervention.
- Interventions can:
  - Target content, actions, or entities (e.g. user accounts, hashtags, forums servers)
  - Be strong (e.g. content removal, bans) or soft (e.g. warning labels or algorithmic demotion)
  - Be applied 1:1 on violations (e.g. a warning label on a post that violates misinformation policy) or based a pattern of violations (e.g. blocking a hashtag based on a high rate of hate speech within it)
- It’s best to disclose interventions to users; some platforms offer opportunities to appeal.
- Interventions will always generate some number of false negatives, false positives and borderline cases where users will disagree with the platform's actions.
Using interventions for elections integrity

Policies, identification methods and sanctions may each be adapted for elections:

- **Policies** may need to be modified or new policies created, for example:
  - Whether certain references to an opposition party would be considered hate speech
  - Whether candidates running for office are exempt from fact-checking
  - New policies might address political advertising regulations, or disenfranchisement

- **Identification**: fast-moving, politicized elections raise many challenges ([see deep dive](#)).

- **Sanctioning** content and users may also be election-specific or modified, e.g.:
  - Interventions specific to elections timelines, e.g. rounds of voting, the period during when votes are tallied, transition of power post-election
  - Legal and internal approvals processes that work at pace to mitigate false positive and negatives
  - New sanction thresholds that reflect heightened social risks
  - Heightened transparency, appeals, and other aspects of **procedural justice** (aka due process)

- Following are a couple of intervention approaches, with specific examples.
Approach: Labeling violating content

- Rather than removing violating content, platforms can apply labels to it.
- Advantages include:
  - Labels are a way to alert users to danger without removing the content, thus preserving things that are newsworthy while allowing users to get extra context.
- Disadvantages include:
  - The harmful content remains on platform the label may mitigate but not fully ameliorate the harm.
  - Labels can sometimes end up increasing visibility of the authors seek to generate outrage (Zannettou and Sanderson et al have written on this).
- This application of labels as an intervention on policy-violating content is conceptually and practically distinct from in-product labels that do NOT rely on identifying violating content; discussion and examples of the latter can be found under product levers.
Example: Meta's third-party fact-checking

Meta works with independent, accredited organizations to fact-check content on Facebook and Instagram.

When fact-checkers rate content, one action (of many) that Meta may take is to attach a label to content containing false information in order to provide additional context.

These labels, if sufficiently prominent to gain users' attention, can arm users to approach content more cautiously.

Meta may also downrank content, **explicitly resulting** in a significant reduction in engagement for repeat violators.

Information Facebook flagged as partly false.
Approach: Limiting actors

- If a platform has a reasonable expectation that someone will violate policy and create harm in the future, they can restrict that person’s ability to keep using the platform.
- This can take many forms, such as time-outs, feature restrictions (e.g. blocking their inability to livestream; to monetize; or to advertise) and permanent deplatforming.
- Advantage:
  - Limiting actors can pre-empt future harm, rather than being reactive to each user action or piece of content.
- Disadvantages:
  - It can generate backlash; without clear policy and explanation, limits can be perceived as biased.
  - When taking action on individual users, especially prominent actors, it is best practice to consider procedural justice (aka ‘due process’) - e.g. avoid punitively downranking without disclosure.
    - Certain highly adversarial situations, such as blocking bulk shill account creation, may be exceptions where features like appeals would not add value.
- Note that this type of intervention against specific actors based on policy violations is distinct from product levers such as ranking algorithms (e.g. PageRank) that differentially reward valuable users, or that ‘gate’ sensitive feature access with requirements such as identify verification or time-on-platform.
Example: deplatforming candidates

One of the most difficult decisions platforms can face is when and how to enforce platform guidelines on prominent users – such as actively-campaigning candidates.

In 2020, Donald Trump transgressed the community guidelines of several prominent platforms, forcing exceedingly difficult trade-off decisions.

Some platforms let prominent political candidates’ posts stay up because of their newsworthiness, even if they violated a policy. Others treated candidates like any other user and didn’t make exceptions.

Research suggests that deplatforming results in migration to other online spaces. Deplatformed users tend to move to fringe platforms, and maintain the levels of overall activity albeit to a reduced audience.

A list of social media suspensions, Ballotpedia.

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Platform</th>
<th>Date</th>
<th>Suspension length</th>
<th>Current status</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim Banks (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>October 23, 2021</td>
<td>Suspended for 14 days</td>
<td>Suspension ended</td>
<td>For violating the platform’s policy  prohibiting “targeted misgendering or degendering of transgender individuals”</td>
</tr>
<tr>
<td>Briscoe Cain (R)</td>
<td>Texas House of Representatives</td>
<td>Twitter</td>
<td>September 13, 2019</td>
<td>Suspended for 141 days</td>
<td>Suspension ended</td>
<td>For tweeting an alleged threat against Beto O’Rourke (D)</td>
</tr>
<tr>
<td>Steve Daines (R)</td>
<td>U.S. Senate</td>
<td>Twitter</td>
<td>February 7, 2023</td>
<td>Suspended for less than a day</td>
<td>Suspension ended</td>
<td>For violating the platform’s Sensitive Media Policy</td>
</tr>
<tr>
<td>Marjorie Taylor Greene (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>August 10, 2021</td>
<td>Suspended for 7 days</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Marjorie Taylor Greene (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>July 19, 2021</td>
<td>Suspended for 12 hours</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Marjorie Taylor Greene (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>January 17, 2021</td>
<td>Suspended for 12 hours</td>
<td>Suspension ended</td>
<td>For tweeting about alleged 2020 election fraud in violation of Twitter’s civic integrity policy</td>
</tr>
<tr>
<td>Marjorie Taylor Greene (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>January 2, 2022</td>
<td>Permanent suspension</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Marjorie Taylor Greene (R)</td>
<td>U.S. House of Representatives</td>
<td>Facebook</td>
<td>January 3, 2022</td>
<td>Suspended for 24 hours</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Ron Johnson (R)</td>
<td>U.S. Senate</td>
<td>YouTube</td>
<td>June 11, 2021</td>
<td>Suspended for 7 days</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Barry Moore (R)</td>
<td>U.S. House of Representatives</td>
<td>Twitter</td>
<td>January 9, 2021</td>
<td>Temporarily suspended</td>
<td>Suspension ended</td>
<td>For tweeting about alleged 2020 election fraud and the attack on the U.S. Capitol</td>
</tr>
<tr>
<td>Rand Paul (R)</td>
<td>U.S. Senate</td>
<td>YouTube</td>
<td>August 10, 2021</td>
<td>Suspended for 7 days</td>
<td>Suspension ended</td>
<td>For violating the platform’s COVID-19 misinformation policy</td>
</tr>
<tr>
<td>Donald Trump (R)</td>
<td>President of the United States</td>
<td>Facebook</td>
<td>January 6, 2021</td>
<td>Suspended for 24 hours</td>
<td>Suspension ended</td>
<td>For posting about alleged 2020 election fraud and attack on U.S. Capitol</td>
</tr>
<tr>
<td>Donald Trump (R)</td>
<td>President of the United States</td>
<td>Facebook</td>
<td>January 7, 2021</td>
<td>Suspended through January 2023</td>
<td>Suspension ended</td>
<td>For potentially inciting violence</td>
</tr>
<tr>
<td>Donald Trump (R)</td>
<td>President of the United States</td>
<td>Twitter</td>
<td>January 6, 2021</td>
<td>Suspended for 12 hours</td>
<td>Suspension ended</td>
<td>For tweeting about alleged 2020 election fraud and attack on U.S. Capitol</td>
</tr>
</tbody>
</table>
Introduction to product levers

- In contrast with interventions, product levers aren’t sanctions.
- They change the user experience to reduce risks and improve outcomes for the platform, users and society – but without being specifically tied to policies, detection, or operational actions.
- This tends to mean less time spent working through edge cases with lawyers (and users!) – although they should still be fair and publicly defensible.
- Product levers might:
  - Prompt or nudge people to avoid violating, or help improve their on-platform contribution (e.g. overlays that ask “are you sure you want to post this?”)
  - Work at the algorithmic level to improve user experience and safety (e.g. they might boost content for being informative as well as being engaging)
  - Work at the structural level by providing or preferencing safer paths or patterns (e.g. Mastodon's decision to not provide "quote tweets")
- Product interventions rarely scale with invested effort, which is both a strength (no need to staff up to detect more bad behavior) and a weakness (you can’t pull in extra staff to fight flare-ups)
- For further reading, we recommend the Prosocial Design Network and New Public.
Using product levers for elections integrity

- Platforms can ask themselves how their product will function in elections contexts:
  - How legitimate use of the product without malicious intent might have bad effects
  - How bad actors might use the product in ways that cause harm

- For example, a post sharing information about violence at a polling site could save lives if true - or disenfranchise voters, if false. These types of post tend to get widely shared, and levers can help:
  - Providing authoritative information in product, or ways for people to get authoritative information
  - Pre-bunking common hoaxes, e.g. vote-by-phone scams
  - Encouraging people to read critically during key periods
  - Reducing virality by discouraging people from resharing emotionally charged material

- Product levers provide different tradeoffs, in contrast to interventions:
  - Procedural justice is less relevant as product levers don’t apply sanction.
  - However, some users may be, or feel, disproportionately affected; even a consistent rule may fall harder on some people than others.
  - Well-designed levers should not need surge staffing to review and enforce, as many enforcement interventions do.
  - Poorly-designed levers can be hard to fix or remove if things don’t go as expected – test before you deploy.
Approach: adding authoritative information

Platforms can add authoritative content around elections in dedicated spaces or embedded in core experiences (e.g. feeds, search) to ensure users have access to accurate information.

Advantages include:
- This doesn’t rely on detection and enforcement of false or misleading content, and can help pre-bunk faster and more widely than fact-checking or other intervention strategies.
- This positions the platform as pro-democracy (and usually non-partisan), and can be an opportunity to engage productively with election authorities in some contexts.
- It pairs well with efforts around civic engagement, voter registration and digital literacy.

Disadvantages include:
- Authoritative information can be difficult to build and maintain, and requires respected partners (e.g. journalists, elections authorities) to build trust. Sometimes platforms invite parties or candidates to add content, though this comes with its own risks.
- Consider how pushy to be and how often this content will interrupt users’ experience. There’s a balance between offering help and being pedantic.
Example: public service announcements

Twitter launched a dedicated *Explore* tab to serve as the centralized hub for real-time political conversation, resources and breaking news on top-tier and priority elections.

This included voter education public service announcements, created using information from nonpartisan government and voting advocacy organizations.

Left example taken from Twitter’s *Explore* tab related to Brazil; right is from US mid-terms (2022).
Approach: "soft" UX measures to discourage anti-social behavior

- These measures are nudges and other user experience changes that guide users toward healthier and less dangerous outcomes – for example:
  - Contextual friction, e.g. “are you sure you want to reshare this link before you open and read it?”
  - AI-generated suggestions to tweak language, to reduce the chances a post will offend readers
  - Removing or redesigning features that are disproportionately subject to abuse (e.g. quote tweets)
- **Advantages include:**
  - These measures are less "heavy" than removal or bans and do not carry the stigma of "shadow-banning," since users can work through or around suggestions if they want to.
  - They allow users to maintain a sense of control while prompting them towards healthier behaviors.
- **Disadvantages include:**
  - They won’t stop users who feel their content or actions are justified and may aggravate those users.
  - Users can become blind to UX interventions over time, so effectiveness can fade with novelty.
  - If your interventions and product levers use the same classifiers, you might be training adversarial users on how to avoid detection for sanctions.
Example: share confirmation

Facebook tested a number of these soft measures, such as this one when a user shares a link without opening it first.

This feature nudges users to read information before sharing it, which reduces high-velocity sharing of low-quality information.
Approach: reducing virality

- Virality reduction measures aim to reduce the speed at which problematic content can accrue reach.
- This is distinct from measures that reduce the reach of particular pieces of violating content. Product levers for virality reduction are systemic changes either:
  - Uniformly applied to all users or content in the system, or
  - Selectively applied to a subset of users or content which, while not proven violating, have qualities that suggest disproportionate likelihood of harm.
- Advantages include:
  - Reducing virality can reduce harms even when you’re unable to identify users, content or actions as violating.
  - Virality reduction is often invoked as a break-glass lever to quickly reduce risk.
- Disadvantages include:
  - This lever doesn’t prevent or remove harmful content, it just reduces its exposure.
  - Uniform reductions may run the risk of making the platform less compelling.
  - Selective reductions can introduce bias or be perceived as shadow-banning.
Examples: US 2020 post-voting violence concerns

Meta employed multiple virality reduction measures, both uniform and selective, to reduce the potential for physical election violence after voting in the US 2020 election.

One uniform lever Facebook applied was to reduce how many people per day could be invited to a group by a particular person. They were (rightly) concerned that Facebook groups were being used to organize violence around the election results. This lever slowed the growth of all new groups, giving more time to review and respond to (intervene) violating groups.

One selective lever Facebook applied was to selectively demote near-violating (or ‘borderline’) content likely to incite violence. This content was not policy-violating at a detection certainty level that Facebook used for existing content moderation, but likely contained some content that would be violating if they had time to do deeper detection, and could be a discovery mechanism for users to connect with violating content.
Approach 4: bridging algorithms

- Bridging algorithms are "algorithms designed to elevate content that resonates with diverse audiences".
- In contrast to how engagement-based ranking tends to promote polarizing content, bridging algorithms overweight content that is broadly socially acceptable, in some ways replicating the agenda-setting of traditional media.
- Bridging can be applied to feed ranking, recommendations and comment ranking, at a content or actor-level.
- Advantages include:
  - Can help identify higher-quality (more informative, less biased) content.
  - Can reduce polarization by advancing content that more people can at least engage with, if not agree with.
  - Improves the user experience and information quality without the overhead of fact-checking partnerships.
- Disadvantages include:
  - Generally it’s hard to explain to users (although perhaps no more so than engagement-based ranking).
  - Can be complex to implement, both algorithmically (what content is bridging) and in how to inject those algorithmic outcomes back into the user experience (how to promote the bridging content).
  - Bridging tends to privilege the center (ie what’s most broadly accepted), which can perpetuate widely-held beliefs (accurate or otherwise), power structures and biases.
Example: Twitter Birdwatch

Twitter says: "Birdwatch works by using bridging-based ranking. Rather than majority rules, or popularity, Birdwatch shows notes that are found helpful by people who tend to disagree (in their Birdwatch ratings). This increases the odds that context added to Tweets is helpful to wide audiences."

Unfortunately, it’s hard to achieve cross-ideology agreement in the political sphere. Less than 10% of Community Notes (as Birdwatch is now called) are made public.
Working with external stakeholders

Creating safe and effective engagement between your organization and external parties
Introduction to working with external stakeholders

- **Part 1** offers considerations for operational practices around interactions with external parties, plus a compendium of resources for engaging with civil society groups.
- Here, we dive deeper into:
  - Research collaborations and productive engagement with academic experts
  - Engaging with elections authorities
  - Other government actors and politicians
  - Escalations, especially external issues raised by governments and civil society
  - External communications and transparency
- Across all these engagements:
  - Teams or individuals experienced with public policy, operations, specific regions or markets, risk management, and legal should be involved.
  - Platforms should strive to work through cultural differences, invest time to build strong relationships and clarify shared goals and responsibilities.
## Models of research collaboration

Academics, nonprofits, and civil society organizations often want to collaborate to identify risks and harms around elections. This can take the form of:

- **Direct employment of researchers embedded on product and policy teams**
- **Expert consultation**, where researchers advise product and policy teams on specific topics
- **Workshops** for rapid iteration of product or research ideas
- **Targeted** or **open funding models**, where grants or gifts are provided (with or without data sharing) for researchers to pursue research outside of platform participation
- **Open datasets and tools** for academic community investigation of platform data and use

**Literature reviews**
Evidence and summaries of what the research community knows about a given topic - e.g. how effective political misinformation interventions are.

**Qualitative fieldwork**
Interviews and observation in election-relevant regions to understand risks and platform uses - e.g. how political groups coordinate information campaigns online/offline.

**Content analysis**
Content-level (human or ML) labeling to understand and measure political content trends (e.g. coding hateful speech trends in text, images, and video during election periods).

**Surveys**
Scalable quantitative measures of attitudes, knowledge, and needs. Can be associated with platform usage (e.g. whether political attitudes relate to online community membership).

**Experiments**
Product changes; can be combined with surveys to measure non-behavioral causal changes (e.g. whether a voting reminder increased people’s expressed intent to vote).
Research collaboration: criteria, risks and examples

Platforms should consider these things when planning research collaborations, especially with academic researchers:

- Whether to bring researchers on staff as domain experts or to maintain external relationships
- How to fund: through direct payments, targeted grants or open gifts (minimized university overhead)
- Access to data and user privacy: how the researcher can access the data (e.g. direct access to data, public open data sharing, analyze only)
- Ethics: working with ethics consultancy organizations or utilizing academic Institutional Review Boards
- Publications: whether to go with blog posts, white papers or peer-reviewed academic journals
- Researcher freedom: whether the company should have a say over what gets published

Here are some examples of elections-related platform research:

- Social Science One
- Meta Election Research US 2020 Project
- Meta Elections and Social Issues Ads
- Meta Integrity targeted grants program
- Journal of Online Trust and Safety
- NDI’s Open Data Initiative
Things to know about elections authorities globally

- Most countries outside the US have centralized (federal level) commissions or management bodies (EMB) to manage elections.
- EMBs frequently hold mandates to:
  - Oversee voter and candidate registration
  - Conduct voter education
  - Procure and distribute electoral materials
  - Count and verify the vote
  - Administer out-of-country voting (if applicable)
- EMBs may also hold statutory powers to formulate regulation or jurisdiction to resolve some election-related disputes. Other entities (e.g. ministry of interior, ministry of foreign affairs, civil registry, security services) may share responsibility.
- National EMBs have sub-national branches accountable to the central entity. They’re usually led by a multi-member leadership and decision-making body, and supported by a professional secretariat in charge of executing the work of the EMB.
- EMBs may be permanent or temporary bodies and can vary in size from 10 or so people to hundreds or thousands of staff.
- Poll workers are often recruited temporarily among civil servants such as teachers and health officials.
- Election commissions can help provide official candidate lists, although these typically are made public close to the election date. In some countries, especially those with parliamentary systems like India, candidate lists can run in the thousands.
- Look out for these common job titles if you want to speak with EMB leadership:
  - Election Commissioner
  - Chairperson (ie head commissioner)
  - Election Officer or Chief Election Officer (in Latin America)
  - Justice, Magistrate or Councilor (if Electoral Tribunals have both administrative and judicial functions)
  - Chief Executive, Executive Director, Secretary, Chair or Secretary General (often the most senior member of the secretariat)

This guidance was developed in part with the assistance of the International Foundation for Electoral Systems (IFES)
How to interact with these authorities

- Election commissions become increasingly busy (and harder to get in touch with) closer to the election.
- Establish contact via the office of the chairperson or a designee at least 3 months before election day.
- Platforms that reach out for the first time with less than 6 weeks to election day may not get a response.
- EMBs’ contact info will be on their websites, but going through a trusted intermediary such as an international organization may increase the likelihood that platforms will be able to engage with the EMB productively.
- Platforms may be operating in a low-trust environment, as EMBs have likely had previous experiences working with social media and technology companies that have shaped their perception of how productive these interactions are likely to be (for better or worse).
- Almost all countries have a national EMB; communicating directly with them is the best way to reach election officials across the country, as they have existing ways to disseminate information to sub-national entities.
- To reach election officials across countries, consider exploring opportunities to engage with regional or global networks of election management bodies - e.g. Africa has 1 continental and 4-subnational EMB networks.

This guidance was developed in part with the assistance of the International Foundation for Electoral Systems (IFES)
What’s most useful for platforms to share

- If the EMB uses the platform, products or physical infrastructure, share a contact they can reach in case the account or services provided are compromised - or if they need to report content or accounts that threaten the physical security of voters, candidates or election workers (by violating policy or contravening local law).
- Share opportunities to amplify civic and voter content on your platform, including election-specific features or products. Share any offered training on how civic actors can best use your platform.
- Share community standards in the local language, and emphasize elements that touch on voter suppression, threats to election officials, political ads or election-related incitement or misinformation. Be aware that EMBs may define misinformation and disinformation more broadly.
- Share any policies around political ads and any transparency features.
- Find time and cover for officials to ask clarifying questions, and share info about products and services. Don’t assume that EMBs know how the platform works or who the user base is.
- EMBs may be eager to understand how other election officials have coordinated with the platform. In addition, some EMBs organize tech fairs where they invite a number of companies to showcase their solutions.
- EMBs in the global majority may not have the institutional capacity to share or receive information about potential misinformation; consider focusing the partnership on providing helpful in-product information about when, where and how to vote in the local languages.
Increasing legitimacy of platform efforts

- Not all EMBs are independent and democratically legitimate, and it isn’t always obvious when this is the case.
- Before engaging with a new election commission, platforms can consult with:
  - Peers at other technology companies that have engaged with the EMB in question
  - International organizations that specialize in building relationships with election authorities (e.g. the International Foundation for Electoral Systems, International IDEA, NDI and IRI)
- Some regions have survey data offering insights into popular or expert opinion about the quality of elections and level of trust in an EMB. See election observer reports, Varieties of Democracy (V-Dem) datasets and regional public opinion polls like Afrobarometer or LAPOP.
- Work with international organizations, including election observers. There’s a ready body of civil society, democratic activists, election observers, international election organizations and democratic actors within governments – all with appetite and capacity to translate your transparency tools into accountability that helps keep democracies strong.
- Platforms that verify accounts should consider their long-term strategy around verifying candidates. The threshold for candidacy in many countries is quite low, and this could mean verifying thousands of accounts. Establish focused criteria and keep verification separate from notability indicators.

This guidance was developed in part with the assistance of the International Foundation for Electoral Systems (IFES)
Interacting with elections authorities in the US

- US elections are atypical in several ways (see part 1). Two particular differences to highlight:
  - Outside the US, countries usually have a centralized election commission; US elections are administered by state and local election officials.
  - Policies and practices around voting vary across and within states, making truth about how to vote a very local phenomenon.
- External communications must engage a large and diverse set of stakeholders in the US because of how the system is set up:
  - State officials usually hold the titles of election commissioner or administrator, often under the auspices of the secretary of state’s office.
  - Regional points of contact – state and even local level – may be needed to get accurate information about voting policies and issues, to engage on account verification, reporting and removal of spoof accounts, and to monitor misinformation, disinformation and malinformation.
- We recommend a dedicated internal person or team manages these external partnerships; assess if they need more support during key moments in the election cycle.
- Establish relationships as far in advance of an election as possible - share contact info with the National Association of State Election Directors, the Election Center and the state’s association of county/municipal clerks or election officials (if relevant).
- Share trends and analysis around election misinformation with election offices so they can tailor their communication efforts
- Share resources and tools to strengthen offices’ cybersecurity, particularly on your platform.
- Coordinate with election officials before sharing information about how to register or vote to a broad audience; officials can help verify accuracy and prepare them for increased traffic.

This guidance was developed in part with the assistance of the Bipartisan Policy Center.
Working with politicians and governments

At some point, politicians and governments will want to use platforms to get information out and engage with users. Platforms will need to determine their approaches around things like:

- **Political and issue ads:** whether to accept them or provide transparency.
- **Fundraising:** whether politicians can raise money on the platform, plus customization and other things necessary to comply with relevant rules.
- **Newsworthiness:** whether to allow newsworthy posts even if they violate policies.
- **Content moderation:** whether to have a special process for moderation, including if a politician is under attack.
- **Fact-checking:** whether to allow politicians’ posts to be fact-checked and labeled.
- **Feature access:** how to roll out new features and whether you ensure some candidates don’t get access before others.
- **Verification:** their approach to building lists of candidates and politicians globally (this can be challenging).
- **Amplification:** whether political and government figures get boosts or demotions in feed.
- **Account security:** whether to have protections to prevent account takeovers (e.g. 2FA reminders)
- **Harassment:** whether to have extra monitoring for accounts that may face disproportionate harassment (e.g. female candidates).
Working with politicians and governments

Even if politicians and governments don’t use a particular platform, there are other ways that platform may need to engage. This includes:

- Lobbying and policy-making: engaging on proposed legislation that affects the platform.
- Government requests: governments will ask for data information on users and may also ask the platform to take down content that violates local law.
- Customer support: each platform needs to decide whether to provide support to politicians and governments in a similar way to how they support celebrities and other high-profile users.
Escalations from civil society and governments

- Many external actors and organizations have strong opinions about platform enforcement actions.

- Our [first guide](#) provided guidance for operational best practices when handling external requests, especially from candidates and politicians; here we cover additional aspects of handling requests from two key groups of stakeholders:

  - **Civil society actors** (organizations, activities and concerned citizens) tend to report ongoing issues to platforms, e.g. content that’s potentially violating or otherwise harmful to society.

  - **Governments** also identify problematic content (privately and publicly), sometimes with the implied threat of future legislative or regulatory action if platforms don’t take action. However, they may also directly allege legal violations by platform users or the platforms themselves.
How and why civil society escalates to platforms

- Civil society actors focus on matters of public interest or potential harms in specific communities – e.g. misinformation, voting rights, disenfranchisement, press freedom, political violence.
- They tend to identify, document and define their issues, which may involve detecting and monitoring issues, and identifying specific pieces of violating content (or behaviors, actors, etc) on platforms.
- There are some common reasons why a platform may not have sanctioned the content they reference:
  - Platform policy doesn’t cover the issue
  - The platform doesn’t believe the content violates the policy
  - There’s a policy exception (e.g. Meta’s ‘newsworthy’ policy)
  - The content violates the policy, but it wasn’t detected
- Thus, escalations from civil society actors might be helpful in patching detection misses, or to challenge a platform's conception and handling of an issue more broadly.
Handling requests from civil society

- It can be helpful to monitor and address these requests, but it comes with risks:
  - Public dialog over policy differences (civil society organizations tend to be effective at outreach)
  - Allegations of bias from other actors, who may accuse the platform of favoritism
- For platforms that invite civil society actors to submit reports:
  - Define membership criteria and set clear expectations – this will encourage structure and make reports more actionable than one-off emails to executives or PR departments.
  - Consider whether participation in any such program public would benefit, or put at elevated risk, the platform and partner organizations.
- When platforms receive unsolicited reports outside a defined program they should:
  - Follow best practices, with a structured and documented review process that "firewalls" your decision-making from undue external influence (see part 1).
  - Engage PR and partnerships teams in the process to determine whether you should communicate outcomes or other responses back to the requestor.
Handling requests from government and law enforcement

- Government or law enforcement agencies report content without reference to legal violations or direct sanctions of the platform can be treated similarly to civil society requests.

- However, also engage internal government relations expertise/teams, as careful attention to communication around these informal requests may avert formal legal notices, as well as unfavorable future legislation or regulation.

- When governments do make allegations of legal violations – e.g. through a takedown request (TDR) or data disclosure request (DDR) - platforms should follow a distinct process:
  - Many platforms have a dedicated Law Enforcement Response Team (LERT), often first established to deal with Child Sexual Abuse Material (CSAM). This team (or a similar model) is a good way to handle these types of request.
  - It’s crucial to understand the laws in question. Platforms may be legally obligated to take specific actions in response to formal requests, and to report a response back through a specific mechanism.
Problematic requests from government and law enforcement

- Sometimes it’s unclear whether there is a legal issue, or whether the government simply believes something is harmful. If this happens, platforms should proceed with extreme caution – seek clarity from whoever made the request and get legal advice.
- Legal requests may also conflict with platforms' values, forcing agonizing decisions about how to respond.
- Governments have very significant coercive levers, including blocking specific platforms (or internet access) within their borders and arresting in-country platform users and employees. Evaluate the legal risks early and determine:
  - How the platform intends to function in the country (including whether or when to exit the country)
  - Criteria by which the LERT or other intake team can distinguish routine from problematic requests or legal notices
  - Internal process to escalate problematic requests and notices for (likely executive) decision-making
- Options for response may include:
  - Compliance
  - Legally contesting requests
  - Non-compliance, which can be either:
    i. General, where a platform refuses all requests
    ii. Specific, where a platform accepts some requests but refuses others as egregious
  - Platforms may face pressure to publicly defend their responses to government requests, including specific compliance and non-compliance decisions – and this public discussion may also be forbidden by the requesting government.
External communications and transparency

- Greater transparency and engagement is an important way for companies and platforms to improve perception among key press outlets and policy makers.

- Clear and stable communication avoids impacting campaign dynamics by perceived (or real) rule changes around on-platform political communications too close to an election. Candidates, elected officials and political parties should be able to plan their campaigns based on established guidelines.

- Establish key spokespeople to utilize with media and lawmakers. Along with key executives, this should include subject matter experts who can go deep on issues to help with credibility.

- There’ll always be a balance of what can be shared publicly and what must remain confidential to avoid aiding any bad actors or for other sensitive reasons – but consider sharing more rather than less when possible.

- Strategic communication and transparency can also play a key role to help thwart the idea that something bad is happening when it isn’t, or it’s not to the degree portrayed (‘perception hacking’).

- Consider a proactive effort to tell the story of implementing these systems and structures. Integrity investments take resources, both in capital and in people, so companies should get credit where due.
Critical topics

Important topics for building your elections integrity practice: generative AI, a taxonomy of bad actors, and guidance for using classifiers
Malicious use of AI in politics and elections

- Among its other capabilities, AI, particularly Generative AI (a.k.a., GenAI) can be used to produce a wide range of synthetic and modified text, audio, images, and video at a far greater speed, scale, and sophistication than previous methods.
  - Given those factors, coupled with the increasing accessibility of AI tools and systems such as Midjourney, Dall-e, Stable Diffusion, and OpenAI’s GPT-4, AI is a potentially disruptive and destabilizing force in politics and elections.

- AI enables the production of deceptive or manipulative media by those with modest technical skills, funds, or — especially useful for foreign influence operations — local language proficiency.

- Because AI remains an emerging technology, much of the threat modeling is somewhat speculative and reactive. Consequently, vigilance is especially critical.
AI abuse: core risks and threat actor tactics

- Deepfakes (synthetic media created using AI deep learning techniques) can be used to manipulate or falsify images, audio, or video for nefarious purposes – e.g., defaming a candidate or sparking civil unrest to disrupt elections.
- Media manipulation isn’t new, but deepfakes are easier to produce by orders of magnitude and can be harder or even impossible to distinguish from authentic content.
- In 2023, the use of deepfakes in politics began moving from novelty to more widespread adoption globally:
  - Deepfake videos defaming candidates have been reported in the Slovakian parliamentary election, the Taiwanese presidential election, elections in Africa, and elsewhere. In all cases, the deepfakes were distributed using social media and messaging platforms.
  - In the U.S., deepfakes were used in a presidential campaign and an RNC ad.
  - Deepfakes are not limited to national elections. On the eve of the Chicago mayoral election, a deepfake portraying a candidate making statements favoring police brutality was posted to a social media platform.
- Actors may use AI-powered botnets and other tactics to engage in information barrage (aka ‘flooding the zone’) based on the belief that high-volume information pollution will overload independent fact-checkers and prove too much for social platforms to detect and remove.
Generative AI: mitigations and considerations

- Platform policies – clarify your rules for synthetic/manipulated media, for example:
  - Whether to prohibit all content that would mislead typical users
  - Whether to distinguish between deepfakes intended to mislead and deepfakes intended for parody or satire
  - Whether to label all doctored or fabricated content or focus on political ads

- Platform detection:
  - Manage your classification models with the recognition that language facility and uniqueness are less reliable indicators of authentic content than they used to be – e.g., for engagement measures, content ranking, and recommendations.
  - Establish communication channels with other tech platforms so that you can quickly share information to contain the spread of deepfakes intended to incite violence.

- Work closely with other platforms to rapidly share and respond to threat information, especially on the eve of elections.
  - Plan for explosive scenarios such as 11th hour deepfakes intended to disrupt the voting process in targeted regions by sparking racial unrest and rioting.

- Regulatory landscape:
  - Expect AI use to be an active area of regulation globally; monitor for emerging changes accordingly.
  - From August 2023, the EU AI Act (draft) designates certain AI use in elections as “high risk,” – and this requires additional controls and scrutiny, plus clear and conspicuous disclosure for deepfakes. The act is expected to be approved by the end of 2023, with a 2-year window before enforcement begins.
  - As of August 2023, the US Federal Elections Commission is starting to evaluate potential regulations for the use of AI deepfakes in campaign ads.

- GenAI vendor controls:
  - Prominent AI providers have pledged to develop watermarking capabilities. However, open source models are, by definition, modifiable by users - meaning that watermarks or other integrity actions can be removed and circumvented.
  - Some AI providers prohibit certain types of targeted and harmful content, e.g., content intended to target particular voting demographics. Policies have proven hard to enforce; ‘jailbreaking’ (also known as ‘adversarial prompting’) has been repeatedly shown to bypass safeguards.
Actor taxonomy: overview

We often break down the actors through the kinds of risks they pose, based on the constraints placed around their activities - especially during civic events like elections.

Understanding the tactics they employ and the goals they set out to achieve can help you better understand how to limit their ability to misuse and abuse your platform.

This non-exhaustive grouping discusses some of the key actors engaged in the recurring spread of mis-, dis-, and malinformation operations on social platforms (and beyond them), using these platforms as a pillar within a multi-pronged attack to promote their agenda.

We use anecdotal evidence from recurring attacks on social media to identify their tactics and goals, so that platforms can develop a better understanding of the landscape and imminent risks.

Legally constrained
Constrained by election law and platform policies

The state and state controlled actors
Extremist groups
Political candidates and operatives

Loosely unregulated
Generally not subject to most laws around election interference

Purely commercial actors
Influencers
Ideologically motivated media
Unwitting propagandists

Extrajudicial
Cannot be constrained by local law, as a practical matter

Hacktivists
Foreign states and sponsored actors
Common tactics and goals across actor types

These goals tend to be shared by multiple types of actors:

- Promoting ideologically-aligned content over other content available on the platform
- Achieving higher reach for manipulated content
- Increasing the potency of disinformation campaigns in influencing public opinion
- Shifting public voting behavior in their favor
- Attracting aligned supporters for fringe communities and improving the communities’ financial outcomes

And these tactics are common across actors:

- Using AI to generate falsehoods, including creating deepfakes and generating false news articles
- Spreading disinformation about opposing candidates to mislead supporters
- Promoting misleading narratives about a particular candidate to gain support
- Exploiting social media for propaganda, including creating and sharing misleading content to question electoral processes
- Creating and promoting divisive content to sow discord and increase polarization
- Hacking and leaking sensitive information to discredit opposition or manipulate public opinion
- Running disinformation campaigns on various platforms, including mainstream social media, local social media and streaming platforms
- Exploiting existing societal fractures to divert attention from sensitive issues
- Microtargeting specific voter groups based on their demographics and political leanings
Tactics employed by particular actor types

**Foreign and domestic state actors**
- State-controlled media in authoritarian regimes may promote the state’s agenda.
- These actors may curtail dissent through internet shutdowns (see Uganda example).

**Commercial actors**
- These actors may use social media for targeted ads, often spreading misleading or false information (see Russia and COVID-19 examples).
- They may also manipulate online discourse for brands.

**Hacker groups**
- These groups are usually anonymous and tend to do things like leak people’s personal details and spy on dissidents’ social media activity.

**Extremist groups**
- Activists may use social media to organize protests and direct actions, e.g. pizzagate.

**News media**
- Media outlets may promote partisan views and spread disinformation, e.g. Fox vs. Dominion.
Classifiers defined and how they can be used

- Classifiers are mechanisms that attempt to categorize things, generally making probabilistic determinations of binary (is/is not) or multi-bucket (e.g. topical or geographic) classifications.
  - E.g. this piece of content **is not** hate speech; that user **is** in Ohio; a discussion group **is** about politics
- Classifiers have several common uses:
  - Monitoring: to identify content and actors important for understanding the nature of civic, political and election discussions
  - Moderation: to identify violating content (or users, groups, etc) to be actioned (e.g. removed or enqueued for review)
  - Product levers: to classify or score in order to determine eligibility to “turn on” certain product features, such as targeted boosts.
  - Emergency interventions: prepare for, qualify or gate application of, and target “break the glass” measures
- In support of these goals it is common to classify classifiers into groups:
  - Violating classification: hate speech, election misinformation, etc
    - Ideally, you have relevant classifiers for all policy violating content
    - Covering text, images and video as applicable on the platform
  - Topic level classification: civic, political, about a specific country, etc
  - Quality classification: Low quality content, high quality content
Implementing classifiers

- Classifiers can be implemented in many ways:
  - Quick-and-dirty detection pipelines can be built with keywords or regular expressions, or even heuristics (e.g. "is reported as violating by >X% of viewers").
  - Machine Learning (ML) classifiers are generally trained on a body of labeled content from your platform.
  - Rich media often benefits from image identification and/or text extraction.
  - Classification can leverage internal (e.g. the text of the content) and external (e.g. user reports of the content as violating) signals.

- These techniques can be combined:
  - For example, a keyword, regular expression or heuristic filter that you would not trust by itself could be incorporated as one signal used by an ML classifier, alongside other signals.
  - ML classifiers can also be chained together, with one classifier output being a signal input by another classifier; e.g. a classifier for "account is a bot" could be one signal consumed by an "is high quality content" classifier.

- If this is daunting: not everything has to be built and done in-house. You can make use of third party services for content classification or open source tools for classification.
Classification challenges around elections

- Elections will be a stress-test of how effective your classifier frameworks are overall:
  - New policies and potentially untested classifiers will become more relevant as the discussion turns more towards politics and the election.
  - Discussions can move swiftly in response to evolving situations (e.g. while polls are open or in response to news).

- ML classification is the gold standard, but elections are particularly challenging for model training:
  - You may not have a relevant body of labeled content for election-specific policies, or for specific political environments.
  - Even if you have an existing corpus it may not be relevant to current events and will lag as issues shift.

- Bright lines between are effectively impossible, e.g. between ‘political’ and ‘non-political’:
  - Non-political issues can become political as issues are dynamic and often politically charged.
  - Be aware of potential bias in application of binary interventions (removal, demonetization).

- Fast-moving adversarial behavior will impact classifier performance:
  - Users will seek to still be identifiable to audiences while avoiding keyword filters.
    - For example: “Anti-vaccine” → “Anti-vaxx” → “jab-free” etc
Be especially mindful where platforms have global reach

- Building classifiers for elections around the globe is challenging; all elections are unique.
- Every election is geographically and temporally specific – the candidates, political issues, timing, rules and regulations.
- Understand the potential for differential outcomes (ie bias) across languages and speech communities (people sharing a language or dialect, who may also share local speech norms), especially where languages correlate with political affiliation.
- These challenges may force you to do things like:
  - Fall back on or supplement ML classification with election-specific keyword or regex-based detection that’s easier to update on the fly
  - Use specialized techniques for rapid retraining of ML pipelines (with attendant costs and risks)
- Overall, you should expect and prepare for classifier performance to degrade around elections. Understand the impact of false positives and false negatives, and whose voices are being silenced as a result.
Conclusion
Impossible trade-offs: resources and realities

- Elections surface many impossible trade-offs, starting from the moment you acknowledge their importance and decide to wade into this critical field:
  - How to support as many elections as possible and whether to go deeper on one at the expense of another
  - When to start and stop supporting a specific election
  - How to prioritize which risks to focus on for an election (e.g. misinformation, account takeover, impersonation, violence, harassment, harmful group activity, information operations)
  - Whether to focus on risks related to the process of carrying out an election, or to include risks that could impact who or what someone votes for
  - Understanding the imbalances built into products and systems that could result in biased enforcement (e.g. lack of classifier support for languages spoken by smaller subset of a country’s population that is a marginalized minority in the country) – and deciding how to mitigate those risks

- As covered in the first section of this deck, setting goals and selecting metrics is essential to facing and answering these questions. Work forward from what you feel you must do and what you can do well. Acknowledge the wins you seek and the limitations you must accept, and own your choices.
Impossible trade-offs: principles in conflict

- Whatever goals you set will inevitably come in conflict with competing principles and concerns:
  - How to balance one goal (e.g. protecting freedom of expression or the press) against another (e.g. intervening on user harms such as misinformation, voter suppression, or harassment of civic actors)
  - Deciding when expressing outrage becomes harassment or hate speech, tipping the scales towards enforcement

- Every action (including no action) involves risks and impacts to your platform, its users and society more broadly:
  - If you fact-check politicians you may be perceived as biased or injecting yourself into the election, or you may lose users or be sanctioned by governments.
  - If you exempt politicians from fact-checks on a newsworthiness basis, it’ll bring its own risks too.
  - Platform, user and stakeholder goals are very likely to be misaligned in less free elections, where governments can present platforms with no-win situations.

- Detection of elections issues can never be flawless; you’ll inevitably sometimes intervene in error, while missing other content that clearly violates your policies. Some users, content, or actions will exist right on the borderlines, and you’ll have internal disagreements. And some set of external actors will be upset with any decision you make.
But the highest stakes demand our best efforts

- Integrity efforts are not binary – platforms cannot simply be labeled as ‘safe’ or ‘not safe’. And they’re also not simply ‘free’ or ‘not free’; each represents a complex set of factors, best placed along a continuum.
- It’s no surprise that elections integrity is and will always be fraught; a continuous struggle and drive for improvement.
- Elections create winners and losers. Platforms are critical to amplifying campaign messaging and fundraising efforts, and can play a big role in outcomes. The increased attention on online elections integrity makes these impossible trade-offs all the more difficult to navigate. There’s rarely a single right answer for anything, much less across all internet platforms and global elections.
- But these stakes – global democracy and freedoms, and the selection of the leaders and policies under which we all live – demand that we assess and tackle the risks; navigate the impossible choices and take action.
- Map out the risks and challenges explicitly and early. Make choices with the information you have, write down the thinking that went into your decision, be ready to pivot if new information arises and conduct retrospectives to understand how to improve in the future.
- Draw on the experience, precedents and expertise of others such to inform your decision-making (e.g. Oversight Board, Integrity Institute (see our resources in part 1)).
- Your work can help us all better protect elections, democracy, freedoms and safety around the world.
Thank you

The Integrity Institute would like to thank the John S. and James L. Knight Foundation for their support of this project.

This resource was prepared by our Elections Best Practices Working Group, led by Glenn Ellingson (lead author), Katie Harbath (lead author) and Abagail Lawson. The deck was co-authored by Jeff Allen, Eric Davis, Alex Leavitt, Swapneel Mehta and Edgard Oliveira.

The authors are extremely grateful for the contributions from experts listed below, in particular the members of the Elections Best Practices Working Group. Their inclusion here doesn’t mean they endorse or support any of the recommendations or statements in this deck.

Working group: Kathryn Campbell, Diane Chang, Lucia Gamboa, Jenn Louie, Kay Spencer, Rebecca Thein, Sam Toizer, Amanda Wall

The following experts and organizations also provided commentary and advice in the development of the deck:

Bipartisan Policy Center
International Foundation for Electoral Systems

Dean Jackson
Natalie Shaw
Questions? Comments? Corrections?

Get in touch if you have feedback or ideas for future elections resources.
Integrity Institute