October 2020

Inauthentic Behavior Report

FACEBOOK

Our team works to find and stop various forms of inauthentic behavior (IB) on our platform. Our work against Coordinated Inauthentic Behavior (CIB), the most egregious form of IB, is among our most known and public efforts. Today, we're beginning to report on our efforts against other forms of IB to paint a more complete picture of the actions we take against a broad range of adversarial behaviors.

PURPOSE OF THIS REPORT

Our goal with this new reporting series is to share trends and tactics we see in IB. We'll discuss examples of enforcement actions we've taken to highlight notable deceptive tactics and how we evolve our responses. By publicizing our findings, we aim to advance the public's understanding of this evolving space, including the gray areas where harm and deception aren't as clear cut. For example, a local candidate may set up Pages to post campaign content without disclosing who's behind them or a social media agency may generate "astroturf" posts that look like they are made by individuals to drive more people to their clients' websites. These areas call for a broader societal discussion to set boundaries for what is and isn't acceptable behavior online.

In future reports, we will share more examples of these gray area behaviors and how we tackle them. This first report lays out foundational concepts that are core to our IB enforcement.

WHAT IS INAUTHENTIC BEHAVIOR?

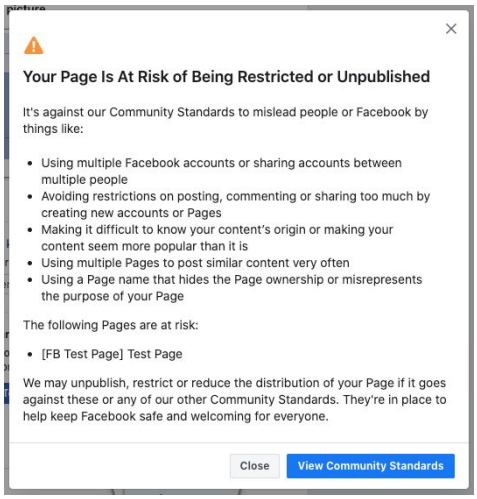
Inauthentic behavior, as detailed in our Community Standards, is an effort to mislead people or Facebook about the popularity of content, the purpose of a community (i.e. Groups, Pages, Events), or the identity of the people behind it. It also includes behaviors designed to mislead Facebook and evade the controls and limits we place on the use of our platforms. As we discussed before, every IB enforcement is based on behavior, rather than the content posted.

Where CIB networks require the central use of fake accounts in an identity-based deception, other IB activity is primarily centered around amplifying and increasing the distribution of content. IB can sometimes involve the use of fake accounts or other

inauthentic assets, but we typically see little attempt to obfuscate their identity from Facebook and only the most superficial attempts to construct a false identity.

ENFORCING AGAINST INAUTHENTIC BEHAVIOR

In countering IB, our primary goal is to mitigate immediate harm to people on our platforms. When the harm is less apparent, such as with authentic users engaging in IB, our first goal is to educate people of our policies so they can correct the violating behavior. For example, if a business engages an outside marketing manager who uses inauthentic behavior tactics to boost the business' reach, we may take steps to warn that business while reducing their distribution to offset any artificial gains. If an organization creates an array of Pages that appear independent without context on who controls them, we may require the organization behind them to disclose their affiliation with the Pages in lieu of any punitive actions.



In the majority of cases, our IB enforcements begin with a warning that gives the user an opportunity to remediate. We've found that it does in fact often lead to people adjusting their behavior to comply with our policies. If they fail to comply, we impose limitations on their ability to use our platform. In the most egregious, harmful or adversarial scenarios, we remove accounts, Pages and Groups for failing to mitigate IB violations.

Through our study of deceptive behaviors and borderline tactics, it's become clear that both legitimate, highly active users and deceptive actors regularly develop new techniques that test the boundaries of our policies. This often raises gray area questions about permissible behavior on our platform: can a political campaign rely on spammy tactics to amplify their messages? What techniques should be allowed to gain likes and followers on social media? Where should we draw the line between creative Page branding and a misleading organization?

Because we know these behaviors will keep evolving, our policies and enforcement responses must continue to evolve as well. We ensure consistency in our response by enforcing against various inauthentic behaviors using specific internal protocols under the IB policy umbrella through scaled automation and manual investigations. We will not

share particular violation thresholds to avoid tipping off deceptive actors on how to game our detection systems.

TACTICS AND TRENDS

Across the thousands of IB enforcements to date, there are a few common themes that guide our enforcement protocols.

Inauthentic distribution tends to be the majority of IB enforcements where inauthenticity is used to cheat our distribution systems and mislead people about the popularity of a piece of content. For example, we've seen social media agencies create inauthentic accounts, Groups and Pages to post at high frequencies to drive people to their clients' Pages or off-platform domains. These behaviors can look very similar to how authentic media, businesses and activists engage with their audiences so we carefully investigate edge cases to avoid over-enforcing. When we find inauthentic distribution, we remove inauthentic assets and restrict the distribution of the violating entity's account or Page until they remediate their behavior. If they fail to do so, we may remove Pages and Groups for repeated or egregious violations.

Abusive audience building is used to dupe people into joining a particular audience or community. Most often we see this used to evade our controls on spam and inauthentic distribution. For example, a Page may switch from politics to sports content, repeatedly change its name to the latest trending topics or share viral clickbait in order to build an audience. Abusive audience building can also include misleading claims about the poster's identity or the purpose behind a Page or Group. When we find abusive audience building behavior, we apply restrictions on the responsible actors' access to the platform and may remove the Pages or Groups built using these tactics.

Image: Example of abusive audience building by a Pakistan-based spam actor that repeatedly changed its Page name and content

Financial motivation tends to be the primary driver behind the vast majority of IB. Much of the activity we've seen is focused on driving people to off-platform websites filled with ads or merchandise, including under false pretenses such as suggesting that the seller is supporting a cause or is part of the same community as their target audience. To enable monetization, these Pages or Groups engage in a number of deceptive or border-line deceptive tactics to entice people to follow them. They often use hot-button issues and other spammy lures. They are well-attuned to their target audiences and will quickly pivot to post about the latest viral content or news to deceive people into clicking links to their site.

Given the current political climate in the US, politics has also become a common lure. These activities can be mistaken for politically-motivated influence operations at first glance, when in fact they are using political themes globally as another form of clickbait, similarly to celeb-bait or puppy memes.

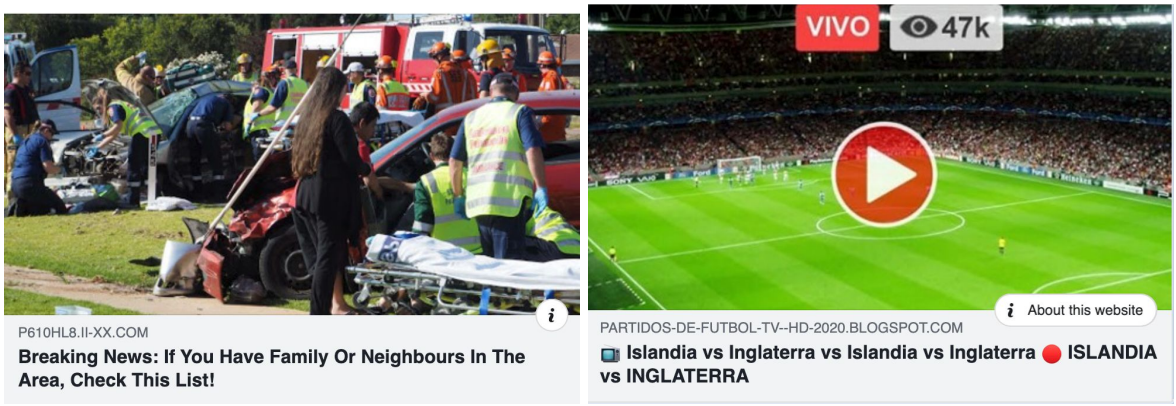


Image: Examples of ad farm links tied to inauthentic behavior. These were automatically detected and removed.

The geographic distribution of IB, like with many other internet deceptions, follows a distinct global trend with much of the activity originating in countries with cottage industries specialized in propagating deceptive schemes to exploit internet platforms. Many of these actors sit at the edge of cybercrime and trade the components involved in these financially-motivated operations across the internet. That includes creating a supply chain of advertising accounts, web domains, compromised accounts and other malicious elements they can use to engage in IB and other violations. These actors are persistent and demonstrate adversarial intent which is why we keep evolving our policies and detection systems to take action against them.

Examples of posts by IB actors from Vietnam (left) and Morocco (right) that we removed for abusive audience building.

IB ENFORCEMENT EXAMPLES

Inauthentic behavior is often detected using automation and through proactive case review at scale, but we know some of these actors and novel tactics can at times evade our detection. To identify these cases, we rely on our investigative teams and external researchers to identify new IB techniques so we can adapt our defenses. In this report, we're sharing several recent examples of our IB enforcements that reflect some of the more complex IB cases we've seen that involve multiple adversarial techniques. This report does not include examples of entities that complied with our requirement for remediation to avoid imposing public harm on people and organizations that are now following the rules.

In these examples, the actors behind them were able to build an audience for a short period of time. Because their content does not otherwise violate our policies, these show the importance of behavior-based enforcement.

Natural News

In May and June 2020, we removed 15 Pages and blocked links to at least 850 domains associated with Natural News. The people behind this activity engaged in repeated and egregious violations of our inauthentic behavior policy. The US business behind these Pages relied on content farms in Macedonia and the Philippines, misled people about the origin and popularity of its content, inauthentically amplified its posts with fake accounts and engaged in deceptive tactics to evade our IB enforcement.

These Pages were removed for violating our spam and inauthentic behavior policies. The network targeted people in the US with political and health-related posts – all to drive traffic to its off-platform store and sites.

We first investigated and removed violating assets of Natural News from our platform in June 2019. Natural News' initial presence on Facebook was built without engaging in violating behavior before the network turned to inauthentic behavior to amplify its content. After the removal, rather than remediating, this network tried to come back using different brands. The people behind this site attempted to disguise some of their content and hide their affiliation with Natural News, while continuing to engage in violating behavior.

As a result of these repeated violations, Natural News and its CEO are banned from Facebook and our enforcement systems continue to monitor for their attempts to come back.

Examples of Natural News content driving people to off-platform domains and webstore

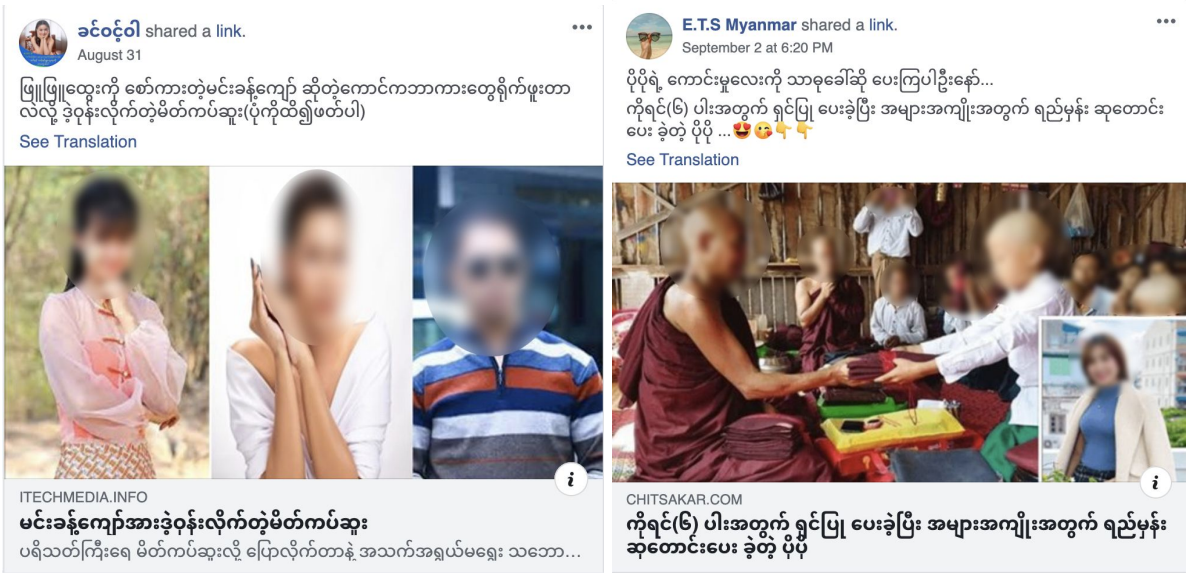
Spam Networks and IB Activity in Myanmar

In August and September 2020, we removed 655 Pages and 12 Groups tied to a number of separate spam networks that distributed clickbait to drive traffic to ad-heavy domains in Myanmar. These networks misled people about the purpose of their Pages and used fake accounts to evade our limits on the frequency of posting. The people behind this behavior created Pages and Groups that were made to look independent of each other and posted at high rates to drive people to connected ad-heavy websites.

In an attempt to build audiences, they posted content ranging from celebrity gossip to local news. A minority of posts from some of these networks and their ad-heavy websites focused on politics in Myanmar, including support for the military and references to ethnic tensions. We did not see evidence of these networks being politically motivated.

Instead, they focused on current events most likely to drive clicks and redirect traffic to off-platform domains.

These networks were removed for violating our spam and IB policy.



Examples of content spread through inauthentic distribution that drove users to off-platform domains

Albanian Ad Farm Network

Over the last several months, we removed 8 Facebook accounts, 14 Groups, 1 Page and 8 Instagram accounts. They were managed by financially-motivated actors in Albania to drive traffic to ad-heavydomains where they posted articles about US politics. These groups were removed within days of their creation. They grew their membership quickly by tricking people in the US into sharing the Groups with their friends on Facebook.

These actors used what appeared to be compromised email accounts to take over Facebook accounts. They then used these accounts to make this network's content appear more popular than it was. This network was removed for violating our IB policy. The Albania-based individuals behind this activity are now banned from our platform.

Image: Example of US-themed political clickbait linking to an ad farm operated from Albania

Foreign Spammers Leveraging US Protests

In May and June 2020, as part of our monitoring for potential IB activity during the US civic protests, we identified and removed a range of inauthentic behavior actors attempting to build audiences by posting viral content from these events. For example, we took down 4 Pages and 13 Groups that were created by several unconnected foreign spam groups from Botswana, Bangladesh, Cambodia and Vietnam targeting people in the US for monetization purposes.

We continue to see deceptive actors try to exploit moments of crisis, tragedy and tensions around the world. IB actors often seek to leverage current events and hot-button issues like the Black Lives Matter movement or COVID-19, as well as celebrities and new TV shows to drive clicks to ad farms or merchandise sites. We're using all the tools and detection systems in our arsenal to identify and stop them.

In these particular cases, we saw spam actors quickly pivot to leverage topics including racial and social injustice and police brutality in the US to trick people into joining their Groups and following their Pages to then direct them to ad farms or merchandise stores. Our Page transparency tools exposed to the public that the people behind these Pages and their content originated from outside the US.

We removed the people behind this activity for violating our IB policy.

Images: Page profile photos used by spammers from Vietnam (left) and Cambodia (right)