What is shadowbanning? How do I know if it has happened to me, and what can I do about it?

November 3 2022, by Marten Risius and Kevin Marc Blasiak

Credit: AI-generated image (disclaimer)

Tech platforms use recommender algorithms to control society's key resource: attention. With these algorithms they can quietly demote or hide certain content instead of just blocking or deleting it. This opaque practice is called "shadowbanning."

While platforms will often deny they engage in shadowbanning, there's plenty of evidence it's well and truly present. And it's a problematic form of content moderation that desperately needs oversight.

What is shadowbanning?

Simply put, shadowbanning is when a platform reduces the visibility of content without alerting the user. The content may still be potentially accessed, but with conditions on how it circulates.

It may no longer appear as a recommendation, in a search result, in a news feed, or in other users' content queues. One example would be burying a comment underneath many others.

The term "shadowbanning" first appeared in 2001, when it referred to making posts invisible to everyone except the poster in an online forum. Today's version of it (where content is demoted through algorithms) is much more nuanced.

Shadowbans are distinct from other moderation approaches in a number of ways. They are:

- usually algorithmically enforced
- informal, in that they are not explicitly communicated
- ambiguous, since they don't decisively punish users who violate platform policies.

Which platforms shadowban content?

Platforms such as Instagram, Facebook and Twitter generally deny performing shadowbans, but typically do so by referring to the original 2001 understanding of it.

When shadowbanning has been reported, platforms have explained this away by citing technical glitches, users' failure to create engaging content, or as a matter of chance through black-box algorithms.

That said, most platforms will admit to visibility reduction or "demotion" of content. And that's still shadowbanning as the term is now used.

In 2018, Facebook and Instagram became the first major platforms to admit they algorithmically reduced user engagement with "borderline" content—which in Meta CEO Mark Zuckerberg's words included "sensationalist and provocative content."

YouTube, Twitter, LinkedIn and TikTok have since announced similar strategies to deal with sensitive content.

In one survey of 1,006 social media users, 9.2% reported they had been shadowbanned. Of these 8.1% were on Facebook, 4.1% on Twitter, 3.8% on Instagram, 3.2% on TikTok, 1.3% on Discord, 1% on Tumblr and less than 1% on YouTube, Twitch, Reddit, NextDoor, Pinterest, Snapchat and LinkedIn.

Further evidence for shadowbanning comes from surveys, interviews, internal whistle-blowers, information leaks, investigative journalism and empirical analyses by researchers.

Why do platforms shadowban?

Experts think shadowbanning by platforms likely increased in response to criticism of big tech's inadequate handling of misinformation. Over time moderation has become an increasingly politicized issue, and shadowbanning offers an easy way out.

The goal is to mitigate content that's "lawful but awful." This content trades under different names across platforms, whether it's dubbed "borderline," "sensitive," "harmful," "undesirable" or "objectionable."

Through shadowbanning, platforms can dodge accountability and avoid outcries over "censorship." At the same time, they still benefit financially from shadowbanned content that's perpetually sought out.

Who gets shadowbanned?

Recent studies have found between 3% and 6.2% of sampled Twitter accounts had been shadowbanned at least once.

The research identified specific characteristics that increased the likelihood of posts or accounts being shadowbanned:

- new accounts (less than two weeks old) with fewer followers (below 200)
- uncivil language being used, such as negative or offensive terms
- pictures being posted without text

- accounts displaying bot-like behavior.

On Twitter, having a verified account (a blue checkmark) reduced the chances of being shadowbanned.

Of particular concern is evidence that shadowbanning disproportionately targets people in marginalized groups. In 2020 TikTok had to apologize for marginalizing the black community through its "Black Lives Matter" filter. In 2021, TikTok users reported that using the word "Black" in their bio page would lead to their content being flagged as "inappropriate". And in February 2022, keywords related to the LGBTQ+ movement were found to be shadowbanned.

Overall, Black, LQBTQ+ and Republican users report more frequent and harsher content moderation across Facebook, Twitter, Instagram and TikTok.

How can you know if you've been shadowbanned?

Detecting shadowbanning is difficult. However, there are some ways you can try to figure out if it has happened to you:

- rank the performance of the content in question against your "normal" engagement levels—if a certain post has greatly under-performed for no obvious reason, it may have been shadowbanned
- ask others to use their accounts to search for your content—but keep in mind if they're a "friend" or "follower" they may still be able to see your shadowbanned content, whereas other users may not
- benchmark your content's reach against content from others who have comparable engagement—for instance, a black content creator can compare their TikTok views to those of a white creator with a similar following
- refer to shadowban detection tools available for different platforms such as Reddit (r/CommentRemovalChecker) or Twitter (hisubway).

What can users do about shadowbanning?

Shadowbans last for varying amounts of time depending on the demoted content and platform. On TikTok, they're said to last about two weeks. If

your account or content is shadowbanned, there aren't many options to immediately reverse this.

But some strategies can help reduce the chance of it happening, as researchers have found. One is to self-censor. For instance, users may avoid ethnic identification labels such as "AsianWomen."

Users can also experiment with external tools that estimate the likelihood of content being flagged, and then manipulate the content so it's less likely to be picked up by algorithms. If certain terms are likely to be flagged, they'll use phonetically similar alternatives, like "S-E-G-G-S" instead of "sex."

Shadowbanning impairs the free exchange of ideas and excludes minorities. It can be exploited by trolls falsely flagging content. It can cause financial harm to users trying to monetise content. It can even trigger emotional distress through isolation.

As a first step, we need to demand transparency from platforms on their shadowbanning policies and enforcement. This practice has potentially severe ramifications for individuals and society. To fix it, we'll need to scrutinize it with the thoroughness it deserves.

Provided by The Conversation