REMODE POLICY PAPER

Best Practices for Deletion of Harmful Content on Social Media

Choosing between hard and soft deletion policies

Authors
Christian Djeffal
Hannah Tilsch
Lisa Mette
Chithra Madhusudhanan

CONSTITUTIONAL DESIGN LAB
Technical University of Munich
TUM THINK TANK
Reboot Social Media lab

Abstract

Social media platforms must rethink their approach to content moderation, moving beyond the binary question of whether to remove harmful content and instead focusing on how content is removed. One design choice in this area is between hard and soft delete. The current norm of hard deletion, where offending posts are entirely erased without any indication they ever existed, causes conversations to lose important context. We propose that platforms shift to a policy of soft deletion as the default. With soft deletion, when a post is removed for violating content guidelines or laws, a notice is put in its place indicating that it was deleted and why. This preserves the flow and coherence of discussions while still removing the harmful content itself. However, we believe impacted users should be given a choice. Platforms should allow those affected by harmful posts to opt out of soft deletion in favor of hard deletion on a case-by-case basis. The key is providing agency to those most directly impacted. When implementing soft deletion notices, platforms must be thoughtful about what information to include. At a minimum, notices should indicate that a post was removed, specify which rule was violated, and ideally provide a link to the relevant content policy. Notices could also include the username of the poster and the date of the original post. This additional context promotes transparency and accountability.

Bibliographical Information

© 2024 by the Professorship of Law, Science, and Technology at Technical University of Munich of Prof. Dr. Christian Djeffal, Arcisstr. 21, 80333 Munich, Germany, Law, Science and Technology, This is version 1.0 DOI: 10.13140/RG.2.2.12892.19842/3

Table of Contents

1. Introduction

Regulatory developments in the last few years have highlighted the importance of content moderation on social media platforms. Most commonly, content moderation involves the removal of content that violates community rules, platform terms and conditions, or even criminal laws. However, when it comes to content removal, the primary focus has been on whether content should be removed or not, whereas less attention has been paid to how content should be removed and what happens when content is deleted.

Within the project Reinnovating Content Moderation (REMODE), we discovered that social media users repeatedly express dissatisfaction with the removal practices of platforms. Users specifically affected by malicious content provided feedback that they were unsatisfied with hard deletion practices, as hard deletion makes it impossible to contextualize the remaining content following a deletion. Conversations in which certain posts were deleted entirely, i.e., without leaving any trace of the deleted content, no longer made sense. Given the varying approaches platforms take to delete harmful content, we decided to explore design options for content deletion practices. This report, thus, presents design options and discusses their relevance concerning context retention, transparency, and solidarity with affected users.[1]

The REMODE project developed a participatory risk governance method and toolbox for social media platforms to enable user autonomy, stimulate good governance, and enforce law and ethics by design.[2][3] The EU's Digital Services Act (DSA), which entered into force for all online platforms in the EU on February 17, 2024, imposes legal requirements on large social media platforms to assess risks and design risk mitigation. REMODE is designed as a progressive way to conduct participatory risk management processes with users. It pushes for tangible and far-reaching ideas to improve social media while involving the most affected groups as much as possible. The Project was funded by the TUM THINK TANK's Reboot Social Media Lab from July 2022 to September 2023.

[1] We extend our gratitude to the following colleagues for their valuable suggestions and insights: Urs Gasser, Jürgen Pfeffer, Yannis Theocharis, Markus Siewert, Desantila Hysa, Craig Speck, Katharina Köhler, Akanksha Bisoyi. Any remaining errors are solely our responsibility.

[2] Djeffal, C., Hysa, D., Herpers, D., Mette, L. and Kearney, C. (2023) REMODE: Re-Designing Content Moderation, TUM Think Tank. doi:10.13140/RG.2.2.27262.56646.

[3] Hochschule für Politik München an der Technischen Universität München (no date) *REMODE. Re-Booting Content Moderation*. Available at: REMODE Re-Booting Content Moderation (Accessed: 21 November 2023).

2. Background

As the Secretary-General of the United Nations, António Guterres articulates: "The proliferation of hate and lies in the digital space is causing grave global harm – now!"[4] The misuse of social media platforms, once intended to foster connections, poses many challenges, especially for the management of offensive content online.[5] Exposure to insults, hate speech, and other harmful content on forums and in comment sections negatively impacts the user experience and can cause serious harm to involved individuals.[6] Users who have had to contend with harmful content online describe the experience as "traumatic" and "dehumanizing."[7] Moreover, individuals exposed to violent or abusive content have been reported to self-censor or leave social media platforms entirely, raising concerns about their participation online and, by extension, in the offline public sphere.[8] Thus, fostering a safe and constructive online space requires platforms to moderate harmful content that violates either laws or platform terms, thereby protecting user rights, freedoms, and interests and making online interactions more comfortable and fruitful.

Currently, much of the discourse on harmful content on social media, such as harassment and hate speech, centers on whether or not to delete specific offensive posts.[9] When moderating harmful content, however, it is crucial to consider the mechanisms and processes underlying content moderation and their significance to users. How the resulting content moderation decisions are enforced and communicated to users is decisive when designing content deletion practices. In other words, content deletion on social media is not only about *what* should be

[4] Guterres, A. (2023) *Secretary-general's opening remarks at Press Briefing on policy brief on information integrity on Digital platforms secretary-general, United Nations*. Available at: Secretary-General's opening remarks at a press briefing on Policy Brief on Information Integrity on Digital Platforms (Accessed: 21 November 2023).

[5] Gillespie, T. (2019) Custodians of the internet [Preprint]. doi:10.12987/9780300235029.

[6] Saha, K., Chandrasekharan, E. and De Choudhury, M. (2019) *Prevalence and psychological effects of hateful speech in online college communities*, *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference*. Available at: Prevalence and Psychological Effects of Hateful Speech in Online College Communities - PMC (Accessed: 21 November 2023).

[7] Madrigal, D.H. and Thakur, D. (2022) *An unrepresentative democracy: How disinformation and online abuse hinder women of color political candidates in the United States*, *Center for Democracy and Technology*. Available at: An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States (Accessed: 21 November 2023).

[8] Amnesty International (2018) *#Toxictwitter: Violence and abuse against women online*, *Amnesty International*. Available at: #Toxictwitter: Violence and abuse against women online - Amnesty International (Accessed: 21 November 2023).

[9] Risch, J. and Krestel, R. (2018) *Delete or not delete? semi-automatic comment moderation for the Newsroom*, *ACL Anthology*. Available at: Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom - ACL Anthology (Accessed: 21 November 2023).166; Haimson, O.L. *et al.* (2021) 'Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), pp. 1–35. doi:10.1145/3479610.; Stockinger, A., Schäfer, S. and Lecheler, S. (2023) 'Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) Technological Tools,' *New Media &amp; Society* [Preprint]. doi:10.1177/14614448231190901.

deleted but also *how* it should be deleted. In particular, we advocate that content deletion should promote democracy, the rule of law, and human rights. To this end, the involvement of users impacted by harmful content should be considered when designing the deletion process.

As the REMODE method is strongly participatory, involving groups most affected by specific harms, we conducted interviews with impacted social media users on a rolling basis throughout the project. One interview with an affected individual highlighted an aspect of content moderation that has yet to receive much attention: How should harmful content deleted by the platform be displayed to other users on the user interface? This raises the question of *hard or soft deletion*. Hard deletion erases all traces of deleted content on the user interface, while soft deletion retains some evidence of the deleted content in the thread, e.g., via a notice of deletion. In this policy paper, we review content deletion practices of social media platforms, outline alternative approaches, and recommend best practices for the deletion of harmful content.

3. Analysis

How platforms approach the removal of content that is illegal or in violation of community rules or terms and conditions has significant implications for affected users. Platforms can either delete a post containing harmful content entirely from the user interface or they can retain some information in the harmful content's place. In this context, a "post" refers to the entire artifact, with username, content, date/time, and other metadata, while "content" refers specifically to the text, image, and/or video within the post.

In this section, we describe the consequences of hard deletion from the perspective of impacted platform users, outline the implications of pursuing soft deletion as an alternative to hard deletion, and offer an overview of how platforms currently go about deleting harmful and offensive content.

3.1. Consequences of Hard Deletion

In an interview with an individual targeted by harmful content online, we gained insight into the shortcomings and consequences of current platform deletion practices, which generally classify as *hard deletion*. The impacted user described an instance in which a platform hard deleted another user's entire post reported for offensive content, with no transparent communication indicating the reason for removal. They recounted how the deletion made it impossible to contextualize the remaining posts in the discussion thread, as the offender's comments had disappeared from the user interface without a trace. The affected individual also raised concerns that other users engaging on the forum are not informed about what transpired before the removal, nor is the hater dissuaded from perpetrating further attacks. In these situations, the impacted user emphasized feeling neglected and powerless when confronted with haters. They stressed that they were missing a feeling of solidarity and support from other users as well as social media platforms in countering and preventing hate online.

> "I just would have wanted it to be visible that it was removed because it was disinformation [...] and that others could see who posted it. [...] **It's simply deleted as if nothing had happened,** [...] but the damage has already been done. [...] I feel like it lacks [...] that there are consequences, so that an individual can't do this over and over again as many times as they want. [...] That they realize, I won't keep getting away with this unrecognized." – Affected User (translated from German)

These comments are supported by a 2021 study,[10] which explored how users who interacted with a post on social media, e.g., reshared or commented, are impacted if the person who originally posted the content retrospectively modifies it. Whereas the focus of this paper is the deletion of harmful content in violation of laws or terms and policies by the platform, the study focused on the modification or deletion of any content by its poster. Still, one of the study's findings echoes the comments of our interviewee. Participants in this study emphasized the importance of context, especially on platforms where users' posts build on each other and enable them to engage in conversation. Specifically, participants raised concerns about "...only deleting the one side of the story" and how this might remove essential context in a discussion.[11] In cases where posts are deleted or modified beyond minor spelling fixes, many study participants desired some marker or notification to inform about the change and preserve context.

3.2. Implications of a Soft Deletion Design

The alternative to the practice of hard deletion is *soft deletion*, which generally describes instances of content removal where evidence of a post's prior existence remains visible. Whereas hard deletion results in a loss of context, soft deletion should retain context and create greater transparency in platform removal practices for harmful content.

[10] Yılmaz, G.S. *et al.* (2021) *Perceptions of retrospective edits, changes, and deletion on social media*, *Proceedings of the International AAAI Conference on Web and Social Media*. doi: 10.1609/icwsm.v15i1.18108
[11] Ibid.

Calls for greater transparency on social media platforms can be seen throughout the EU's DSA for, among other things, enabling scrutiny, accountability, and redress options for content moderation decisions. Articles 17 and 24 (5) of the DSA, e.g., require statements of reasons to both be provided to users whose content has been moderated and added to the publicly accessible Transparency Database. These statements of reasons provide transparency regarding the type of restriction enacted, the grounds for the restriction, and the facts around the situation in which the content moderation decision was made.[12] Similarly, soft deletion can serve to make platform communications regarding content deletion practices and the evaluation and classification of malicious content more transparent. This transparency would not just be delivered to users who posted the content that is being moderated but to all users on the platform, including those directly affected or targeted by the content. This could be done, for instance, by replacing the deleted content with a deletion notice that references the relevant content guidelines or laws that were violated.

However, when pursuing greater transparency through soft deletions to safeguard values such as personality rights and democracy, it is necessary to balance these interests with privacy and data protection, especially those of the user whose content is being removed. When content is deleted and replaced by a deletion notice stating the grounds for the decision to delete, a violation of guidelines or laws may be brought into connection with the deleted content's author's username or name. Even if the platform removes the username when replacing allegedly harmful content with a deletion notice, other users may have screenshotted or may know who posted the now-deleted content. Thus, user reputations may be negatively impacted by deletion notices. This can be problematic in cases where deletion notices claim a specific violation that has not yet been legally substantiated or is outright erroneous. Even in cases where the content has been accurately assessed, removed, and labeled by the platform, users may rely on their right to be forgotten and ask for such content to be removed after some time.[13]

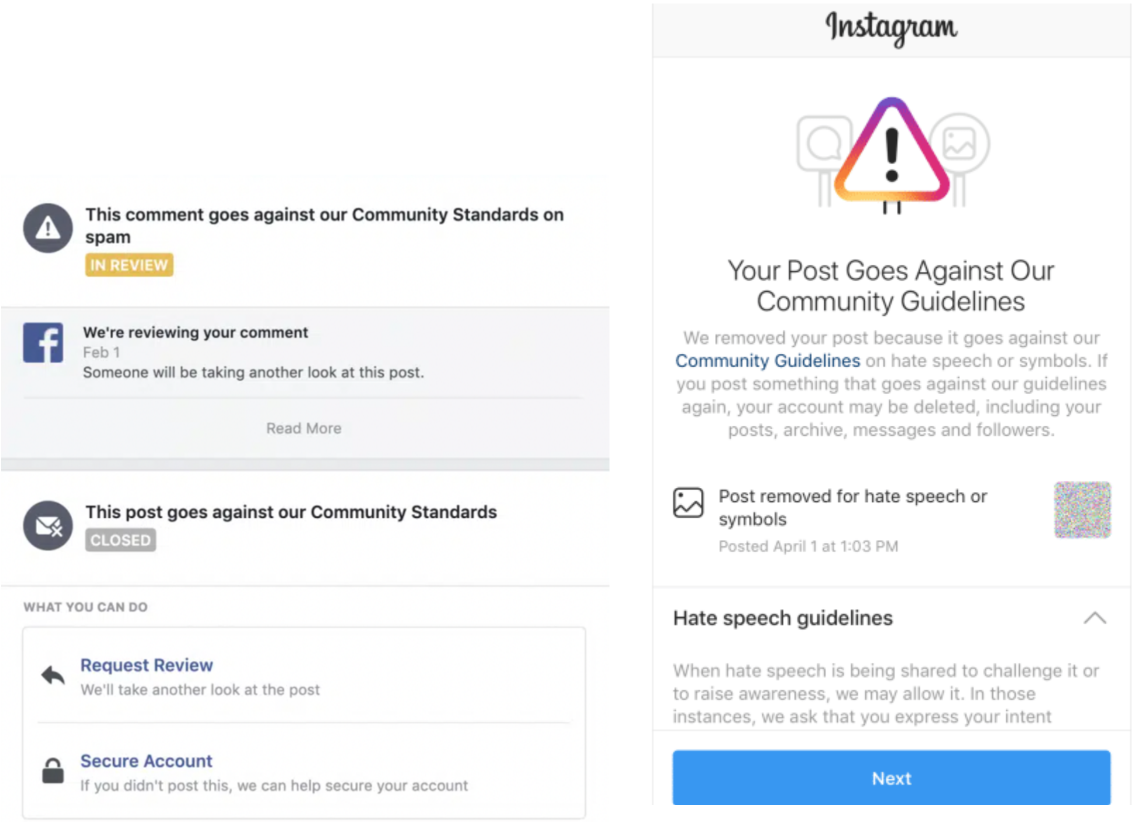
3.3. Common Platform Practices

While the terms and policies of social media platforms indicate their commitment to removing content that violates their content rules and guidelines, they generally do not provide information on how the deletion will be implemented on the user interface or the backend. Instagram's

[12] European Commission (no date) *DSA Transparency Database FAQ*. Available at: DSA Transparency Database FAQ (Accessed: 3 April 2024).

[13] See Art. 16 GDPR and Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González (C-131/12) EU:C:2014:317, [2014] QB 1022.

Community Guidelines, e.g., do not specify how deleted content will appear to others on the platform.[14] The terms and conditions of most social media platforms do, however, state that when platforms delete a post, the poster will be informed, usually with an explanation and information about available decision review options.[15] Under the DSA, this is now obligatory: Platforms must provide a statement of reasons to the user who posted the content when they remove or otherwise restrict the content's visibility. In some cases, deletion notices provided to the post's author may hint that the respective post is now only visible to the original poster and no longer visible to other users. In other cases, other users who reported a post for removal may be informed about its deletion following platform review. However, these notices are generally not visible to all users interacting with the deleted post or the thread/forum it was embedded in.



This comment goes against our Community Standards on spam

IN REVIEW

We're reviewing your comment

Feb 1

Someone will be taking another look at this post.

Read More

This post goes against our Community Standards

CLOSED

WHAT YOU CAN DO

Request Review

We'll take another look at the post

Secure Account

If you didn't post this, we can help secure your account

Instagram

Your Post Goes Against Our Community Guidelines

We removed your post because it goes against our Community Guidelines on hate speech or symbols. If you post something that goes against our guidelines again, your account may be deleted, including your posts, archive, messages and followers.

Post removed for hate speech or symbols

Posted April 1 at 1:03 PM

Hate speech guidelines

When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent

Next

Figure 1: Deletion notifications (to content poster) on Facebook and Instagram.[16]

Evidence of hard deletion by social media platforms is hard to find, as by definition, there is no trace or indication of the deletion that has occurred. Without more insight into how posts that

[14] *Community Guidelines* (no date) *Help center*. Available at: Community Guidelines | Instagram Help Center (Accessed: 21 November 2023).

[15] Meta (no date) *Taking down violating content*, *Transparency Center*. Available at: Taking down violating content | Transparency Center (Accessed: 21 November 2023).

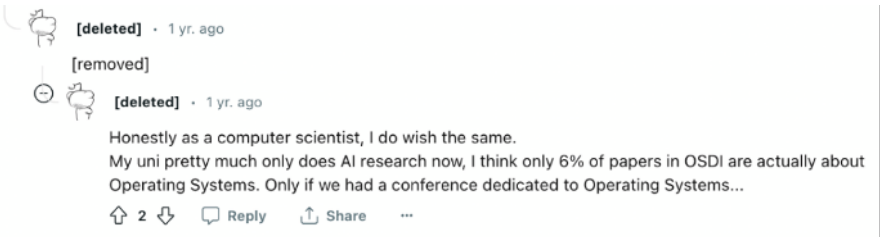
[16] Felicitas, A. (no date) *Instagram's Warning Notification Gives At-Risk Accounts a Second Chance*, *AdvertiseMint*. Available at: Instagram's Warning Notification Gives At-Risk Accounts a Second Chance (Accessed: 21 November 2023).

have been deleted by the platform appear to other users, we must rely on user screenshots and discussions about deletion activity. Thus, the following overview of current deletion practices is based on information available through desktop research and user reports.

Existing platform practices for content deletion range from a complete, hard deletion to a softer deletion, where evidence of the post's prior existence remains visible and possibly a general reference to a violation of content guidelines is made. Specifically, posts can be:

1. Deleted without notice or any other trace: A harmful post in a thread is deleted to be no longer visible to other users. The deleted post is missing entirely, and there is no indication of its deletion. This is considered hard deletion. Evidence of the post's prior existence is lost, causing the surrounding posts to stand alone and lack context. In consequence, the thread becomes challenging to comprehend.

2. Deleted with a "deleted" or "removed" label: A harmful post in a thread is deleted so that the post's content is no longer visible to other users. A deletion label (in Figure 2, "removed") replaces the deleted post in the thread, indicating that a post previously existed. The label only states that a post was deleted; it provides no further reason for the removal, e.g., that the content violated platform rules or laws. Still, a trace of the previously existing post remains, preserving the original flow of the thread.



[deleted] · 1 yr. ago

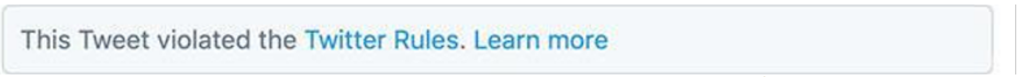
[removed]

[deleted] · 1 yr. ago

Honestly as a computer scientist, I do wish the same.
My uni pretty much only does AI research now, I think only 6% of papers in OSDI are actually about Operating Systems. Only if we had a conference dedicated to Operating Systems...

Figure 2: Deletion label on Reddit.[17]

4. **Deleted with a notice referring to the reason for deletion:** A harmful post in a thread is deleted so that the post's content is no longer visible to other users. A deletion notice, which includes a reference to a violation of content rules or any other reason for deletion, replaces the deleted post. The notice remains within the thread, showing that a post previously existed and provides a limited indication of why the content has been deleted.



This Tweet violated the Twitter Rules. Learn more

Figure 3: Deletion notice on Twitter.[18]

[17] r/technology. (2022) *Scientists increasingly can't explain how AI Works*, *Reddit*. Available at: Scientists Increasingly Can't Explain How AI Works (Accessed: 21 November 2023).
[18] *X account notices and what they mean - suspensions and more* (no date) *Twitter*. Available at: X account notices and what they mean - suspensions and more (Accessed: 21 November 2023).

5. Policy Options

This section outlines three key design options for shaping how platforms carry out content deletion. We first explore the options of hard versus soft deletion, then discuss platform control as opposed to user control, and conclude with a survey of the various options platforms have when deciding how to design soft deletion notices.

5.1. Hard or Soft Deletion?

Platforms may choose between implementing a hard or soft deletion when removing reported content that violates the platform's terms and policies. Hard deletion refers to the absolute removal of the respective post from the platform's user interface without leaving any notice or other trace of the post's previous existence. This practice is common among social media platforms today. However, it raises concerns regarding the transparency of content deletions and the traceability of online discussions when entire posts are deleted from a thread. In contrast, soft deletion provides increased transparency about the deleted posts and the deletion decision. To this end, platforms may include deletion labels or notices in place of the original post. In general, this additional information serves to preserve the flow of the thread and provides information about the deleted post and the reasons for its deletion.

5.2. Platform Control or User Control?

One crucial aspect in deciding between hard and soft deletion is considering whether to adopt either one as a norm or provide both as an option. Currently, platforms control how content deletions appear to other users of the platform and generally choose to pursue hard deletions. As an alternative, platforms could facilitate user control by providing the user specifically affected by the harmful content with a choice between a soft and hard deletion on the user interface.

5.3. Options for Designing Soft Deletion Notices

When adding a **soft deletion notice** in place of a harmful post, platforms may consider including the following information:

1. **Username of the author:** Platforms may keep the name of the user who posted the harmful post visible, although the content has been deleted. Alternatively, the platform may remove the username by striking it out, replacing it with a generic username, e.g., "former user", or leaving it out altogether.

2. Reason for deletion: Platforms may provide a reason for the deletion, which justifies their decision and provides transparency for other users interacting with the deleted post. When providing a reason, platforms may offer a generic reference to the grounds for removing the content, such as, “This post violates our content rules”, or they may offer a more specific reference to grounds for the decision, such as, “This post violates our content rules on Bias and Discrimination. [Learn more]”, for instance including a link to the relevant content policy and illustrative examples. These references to grounds relied upon may specify applicable laws for illegal content or the platform’s content rules and community guidelines. Platforms may also choose to explain how the content violated or was found to be incompatible with the grounds for deletion. This expands on the grounds provided by offering a more personalized explanation for the content removal. Links to more information about the moderation action, including specificities and timelines, can be provided.

3. Enforcing entity: Platforms may choose to indicate who deleted the harmful post to make clear that the post was deleted either by the platform commonly following human review of flagged or reported content or the author of the content after being intimated by the platform of its inappropriate nature. Alternatively, they may state that the content was deleted without mentioning the responsible entity.

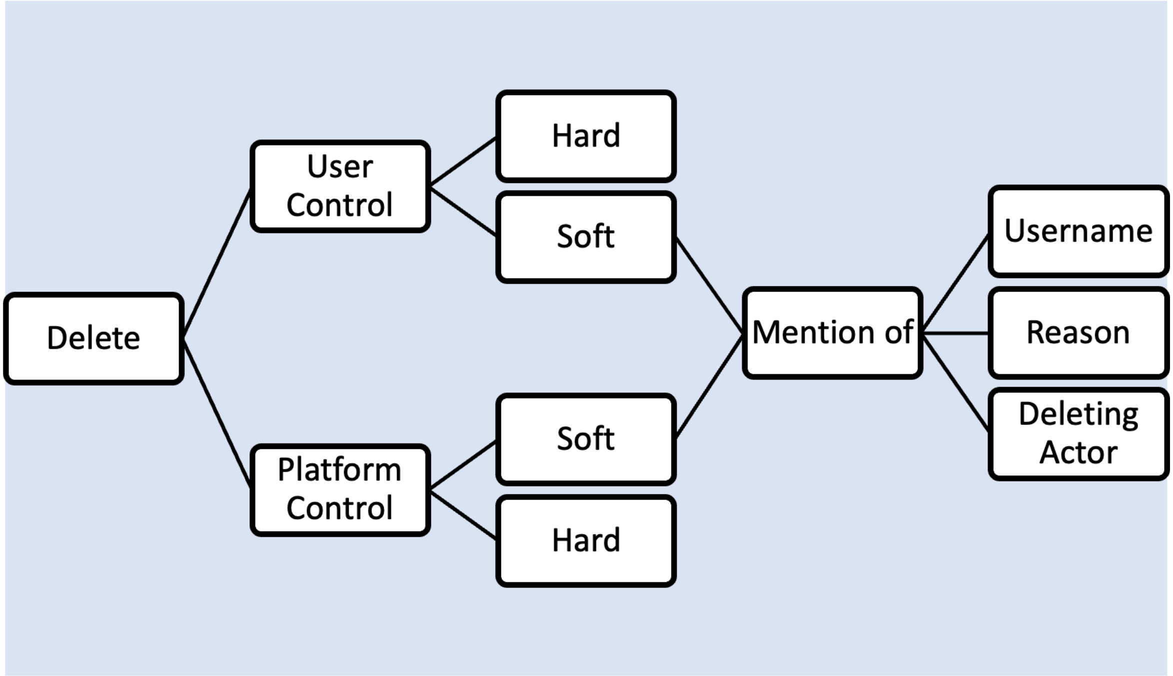


Figure 4: Visualization of deletion choices.

6. Recommendation

The following section outlines our recommendations for how platforms should conduct harmful content deletions. We argue that platforms should generally adopt a soft deletion approach but also recognize user control as the exception to the rule. Finally, we propose what information platforms should include in their soft deletion notices.

6.1. Residual Soft Deletion Approach

We recommend that social media platforms generally implement soft deletion when removing harmful content. Impacted social media users have stressed the importance of preserving context when deleting content, especially on platforms that support dialogue between users. Soft deletion provides context by informing users visiting the discussion thread with deleted content that something has happened. Any notice of deletion already provides greater context and transparency than in cases of hard deletion, where no trace of a former post is visible to other users, and thus, context and understandability are lost.

6.2. User Control as Exception to Soft Deletion

While we recommend that platforms generally implement soft deletions, impacted users should be given control by being able to opt-out and instead choose hard deletion. This is important because affected users' needs vary in situations where harmful content is deleted.

On the one hand, as reflected in our interview with an impacted user, hard deletion can lead to complicated situations for the affected users. They may be dealing with a persistent troll who revisits the same conversation despite having posts deleted. Because deleted posts are not visible, there is no visual means to increase the accountability of a user who violates the rules. Hard deletion can also change the natural flow of a conversation in a way that makes it unintelligible. Thus, an impacted user may prefer a soft deletion to increase accountability and retain context.

However, in some situations, soft deletion may perpetuate offensive behavior. An example would be harmful comments made by a student's classmates. If the content was deleted but the posts were still visible, it may be evident that there was a practice of collectively insulting a person. While the harmful words would be removed, there would still be an indication of the bullying. In these situations, the affected user may prefer to have the comments deleted entirely so they are not constantly reminded of the incident. Therefore, impacted users should be given the choice to opt out of the soft deletion practice and choose hard deletion instead.

6.3. Soft Deletion Notice Design

In the case of soft deletion, we recommend that platforms implement soft deletion notices that include a specific reason for the content removal, state the platform's active role in the deletion, and retain the offender's username. These recommended soft deletion notice building blocks are examined in detail below:

1. Include a specific reason for deletion concerning the relevant laws, terms, or policies: We recommend that deletion notices include the specific grounds for the deletion, making clear whether the content was found to be incompatible with platform policies or illegal. The deletion notice should reference the applicable section of either the platform's terms and policies or the relevant law. The grounds should be as specific as possible, as users have expressed frustration with reasons for deletion that are too ambiguous, e.g., references to entire policies or generic statements regarding violations of broad laws.[19] Platforms may further clarify the grounds for the content removal by relying on the violation categories outlined in the DSA Transparency Database to codify the type of incompatibility or illegality. These categories include, e.g., illegal or harmful speech, online bullying/intimidation, and adverse effects on civic discourse or elections. Ideally, this information regarding the grounds for deletion is accompanied by an explanation that offers greater insight into why this post was specifically found incompatible or illegal. On the one hand, specifying the grounds and explanation for removal serves as a justification for the decision to delete, thereby reassuring other users that there was sufficient cause. On the other hand, the reason for deletion acts as a transparency mechanism, offering context. Importantly, it also builds awareness of relevant terms, policies, and laws that govern content on the platform and enables users to learn from moderation decisions.[20] Users might thus familiarize themselves with the terms and policies they otherwise skim, better equipping them to report harmful content they encounter in the future.

2. State the platform's active role in deletion: Many deletion notices that are currently

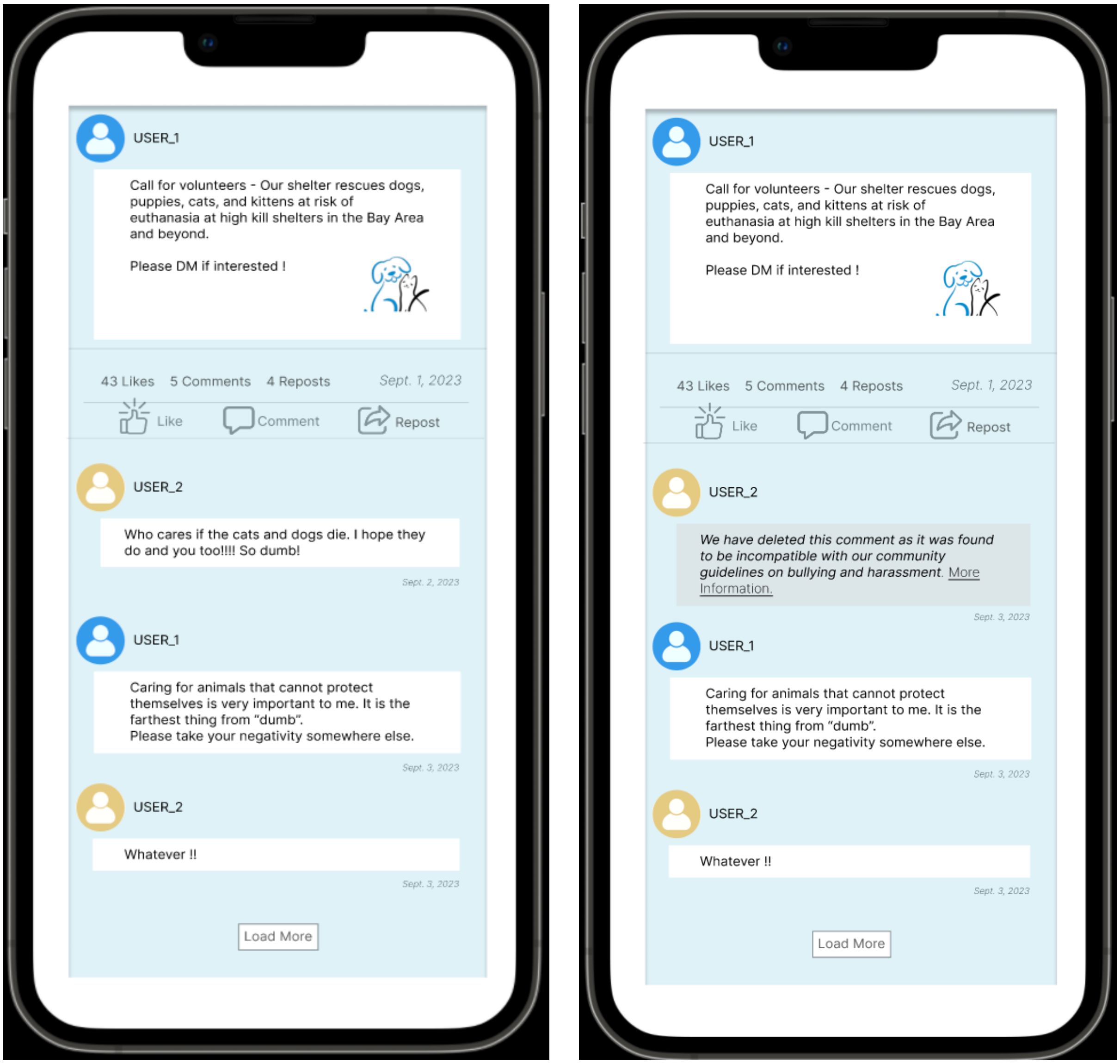
[19] Myers West, S.. (2018) 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.', *New Media & Society*, *20*(11), p.4366-4383. doi:10.1177/1461444818773059.

[20] Jhaver, S., Bruckman, A. and Gilbert, E. (2019) 'Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), p. 150:1-150:27. doi:10.1145/3359252.

in use do not explicitly state who executed the deletion. These notices only state that something has been deleted (see Figure 2 and Figure 3) without mentioning an enforcement actor. We recommend formulating the deletion notice in active voice, with the explicit statement that the platform, e.g., “We”, deleted the content for the given reason. This formulation increases the visibility of the platform's action, signaling to users that the platform takes accountability for harmful content. It may also inspire or incentivize other users to report harmful content, contributing to a safer online environment.

3. Retain the username: Soft deletion notices should retain the author’s username to prevent “repeat offenders”, who may, despite having comments deleted earlier, return to the thread to comment later. By including their username, these returning users can be easily identified. Other users can then also adopt mechanisms to avoid harassment or unwanted interaction from the user, e.g., by blocking them.

Below is a visualization of how the deletion notice can be constructed if the impacted user opts to have the offensive content soft deleted. In this case, additional information can be accessed by those interested in learning more by clicking on “More Information”. Here, platforms can state the incompatibility category according to the DSA’s codification scheme, e.g., online bullying/intimidation, link the relevant section of the community guidelines policy on bullying and harassment, and explain why the content is considered incompatible with the community guidelines on bullying and harassment.



USER_1

Call for volunteers - Our shelter rescues dogs, puppies, cats, and kittens at risk of euthanasia at high kill shelters in the Bay Area and beyond.

Please DM if interested !

43 Likes 5 Comments 4 Reposts Sept. 1, 2023

USER_2

Who cares if the cats and dogs die. I hope they do and you too!!!! So dumb!

Sept. 2, 2023

USER_1

Caring for animals that cannot protect themselves is very important to me. It is the farthest thing from "dumb".
Please take your negativity somewhere else.

Sept. 3, 2023

USER_2

Whatever !!

Sept. 3, 2023

USER_1

Call for volunteers - Our shelter rescues dogs, puppies, cats, and kittens at risk of euthanasia at high kill shelters in the Bay Area and beyond.

Please DM if interested !

43 Likes 5 Comments 4 Reposts Sept. 1, 2023

USER_2

We have deleted this comment as it was found to be incompatible with our community guidelines on bullying and harassment. More Information.

Sept. 3, 2023

USER_1

Caring for animals that cannot protect themselves is very important to me. It is the farthest thing from "dumb".
Please take your negativity somewhere else.

Sept. 3, 2023

USER_2

Whatever !!

Sept. 3, 2023

Figure 5: Visualization of Recommended Soft Deletion Notice (Own work)

7. Conclusion

Exposure to malicious content online is a serious issue on social media and can adversely affect an individual's physical and mental well-being. To combat this, social media platforms have designated content moderation mechanisms for handling reports of abuse and otherwise harmful content. In most scenarios, the offensive content will be deleted from the user interface if found to violate platform policies or other laws. While much of the discourse around content moderation has focused on *what* to delete, insights from a social media user directly impacted by harmful content highlighted the significance of *how* platforms delete.

This policy paper has examined current social media platform policies and alternative design options for deleting harmful content online. We identified two main options for how deletions are displayed to platform users, namely, hard and soft deletion. The first, currently practiced by most platforms, would be to completely delete the whole post, removing any trace of its prior existence from the user interface. The alternative is to pursue a soft deletion, where the harmful content is deleted but not the entire post. In this case, a notice of deletion replaces the deleted content. As the hard deletion approach can lead to loss of context, and users have indicated they prefer context retention in content moderation scenarios, we have recommended that platforms generally pursue soft deletions.

However, because affected users have different needs and experiences, we recommend that platforms give affected users control by empowering them with the choice to opt out of soft deletion. Whereas platforms currently decide how the deletion will be enforced and displayed to other users, this would give impacted users greater control. While some impacted users may prefer soft deletion, as it retains greater context and may discourage posters of malicious content from repeating the offense, others may find evidence of deleted comments on a thread to be an unwanted reminder of the incident. Therefore, it is crucial to consult affected individuals.

Finally, in the case of soft deletion, we recommend platforms make visible a deletion notice that includes a specific reason for the deletion that references the relevant sections of laws and/or platform policies, clearly states its role in the deletion, and retains the username of the user that posted the harmful content.

8. References

Amnesty International (2018) *#Toxictwitter: Violence and abuse against women online*, *Amnesty International*. Available at: [#Toxictwitter: Violence and abuse against women online - Amnesty International] (Accessed: 21 November 2023).

Community Guidelines (no date) *Help center*. Available at: [Community Guidelines | Instagram Help Center] (Accessed: 21 November 2023).

Djeffal, C., Hysa, D., Herpers, D., Mette, L. and Kearney, C. (2023) 'REMODE: Re-Designing Content Moderation', TUM Think Tank. doi:10.13140/RG.2.2.27262.56646.

European Commission (no date) *DSA Transparency Database FAQ*. Available at: [DSA Transparency Database FAQ] (Accessed: 3 April 2024).

Felicitas, A. (no date) *Instagram warning message - your account may be deleted*, *AdvertiseMint*. Available at: [Instagram's Warning Notification Gives At-Risk Accounts a Second Chance] (Accessed: 21 November 2023).

Gillespie, T. (2019) *Custodians of the internet* [Preprint]. doi:10.12987/9780300235029.

Guterres, A. (2023) *Secretary-general's opening remarks at Press Briefing on policy brief on information integrity on Digital platforms secretary-general*, *United Nations*. Available at: [Secretary-General's opening remarks at a press briefing on Policy Brief on Information Integrity on Digital Platforms] (Accessed: 21 November 2023).

Haimson, O.L. *et al.* (2021) 'Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), pp. 1–35. doi:10.1145/3479610.

Hochschule für Politik München an der Technischen Universität München (no date) *REMODE. Re-Booting Content Moderation*. Available at: [REMODE Re-Booting Content Moderation] (Accessed: 21 November 2023).

Jhaver, S., Bruckman, A. and Gilbert, E. (2019) 'Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), p. 150:1-150:27. doi:10.1145/3359252.

Madrigal, D.H. and Thakur, D. (2022) *An unrepresentative democracy: How disinformation and online abuse hinder women of color political candidates in the United States*, *Center for Democracy and Technology*. Available at: [An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States] (Accessed: 21 November 2023).

Meta (no date) *Taking down violating content*, *Transparency Center*. Available at: [Taking down violating content | Transparency Center (Accessed: 21 November 2023).

Myers West, S.. (2018) ‘Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.’, *New Media & Society*, *20*(11), p.4366-4383. doi:10.1177/1461444818773059.

Risch, J. and Krestel, R. (2018) *Delete or not delete? semi-automatic comment moderation for the Newsroom*, *ACL Anthology*. Available at: Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom - ACL Anthology (Accessed: 21 November 2023).

Saha, K., Chandrasekharan, E. and De Choudhury, M. (2019) *Prevalence and psychological effects of hateful speech in online college communities*, *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference*. Available at: Prevalence and Psychological Effects of Hateful Speech in Online College Communities - PMC (Accessed: 21 November 2023).

Stockinger, A., Schäfer, S. and Lecheler, S. (2023) ‘Navigating the gray areas of content moderation: Professional moderators’ perspectives on uncivil user comments and the role of (AI-based) Technological Tools,’ *New Media &amp; Society* [Preprint]. doi:10.1177/14614448231190901.

X account notices and what they mean - suspensions and more (no date) *Twitter*. Available at: X account notices and what they mean - suspensions and more (Accessed: 21 November 2023).

r/technology. (2022) *Scientists increasingly can’t explain how AI Works*, *Reddit*. Available at: Scientists Increasingly Can’t Explain How AI Works (Accessed: 21 November 2023).

Yılmaz, G.S. *et al.* (2021) *Perceptions of retrospective edits, changes, and deletion on social media*, *Proceedings of the International AAAI Conference on Web and Social Media*. doi: 10.1609/icwsm.v15i1.18108.