表示頻度

更新日：2021年5月19日

Facebookは、ポリシー違反がサービス利用者にもたらす影響を最小限に抑えることを目指しています。この目標に対する成果を評価するため、違反コンテンツの表示頻度を測定しています。

●表示頻度とは

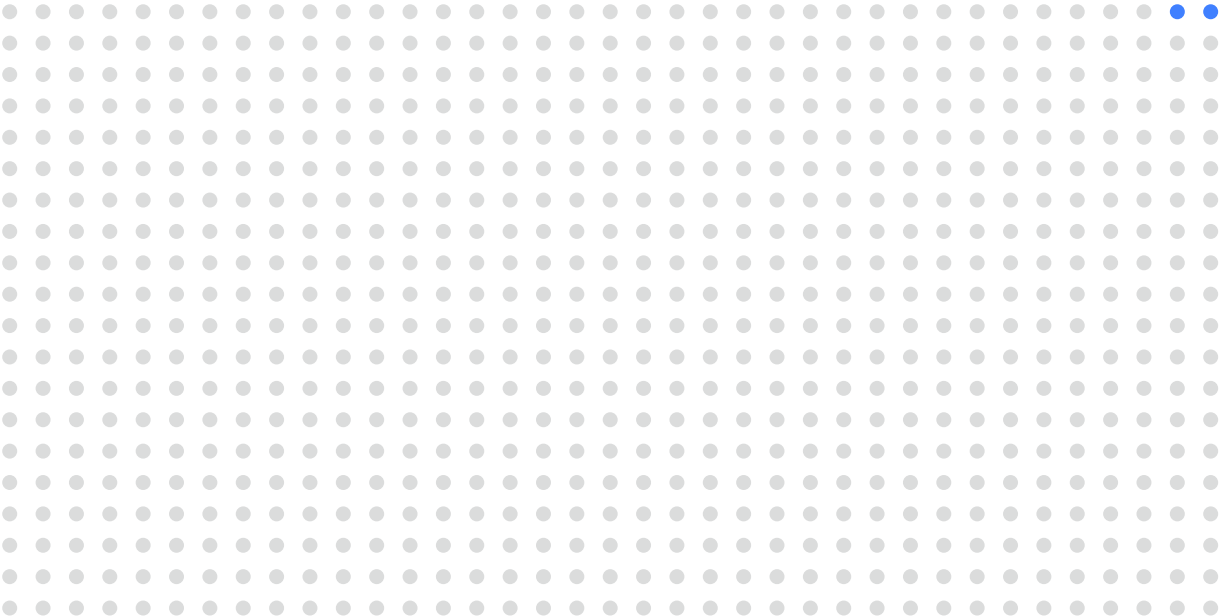
表示頻度とは、FacebookまたはInstagramにおけるコンテンツ閲覧をすべて検討し、そのうち違反コンテンツの閲覧であると推定される割合を測定します（閲覧の定義については「閲覧数の表示頻度を測定する理由」で詳しく説明しています）。この指標は、違反コンテンツによる影響はそのコンテンツの閲覧回数に比例すると仮定しています。

表示頻度とは、別の考え方でとらえれば「違反コンテンツの閲覧をどれだけ防ぐことができなかったか」ということになります。これには、違反を早期に発見できなかった場合や完全に見逃してしまった場合があります。

●表示頻度の算出方法

違反コンテンツの表示頻度は、FacebookまたはInstagram全体のコンテンツ閲覧のサンプルを使用して推定されます。この値は、違反コンテンツを表示した閲覧数の推定値を、FacebookまたはInstagramのコンテンツ閲覧回総数の推定数で割って算出しています。例えば、成人のヌードと性的行為に関するコンテンツの表示頻度が0.18%~0.20%だった場合、10,000回閲覧されたコンテンツのうち、成人のヌードと性的行為に関するポリシーに違反したコンテンツが平均18 ～ 20回あったことになります。

1ドット＝閲覧数10回
総閲覧数10,000回
違反コンテンツ閲覧　20回



表示頻度が0.20%であった場合、10,000回閲覧されたコンテンツのうち、違反コンテンツが20回あったことになります。
数字は非常に小さいですが、わずかな数であっても利用者に大きな影響を及ぼします。

違反の種類によっては、弊社のサービス上で発生する頻度が非常に低いものがあります。違反コンテンツが閲覧される可能性は非常に低く、閲覧される前にそのほとんどが削除されます。そのため、表示頻度の正確な推定に十分な数の違反サンプルが見つからない場合が多くあります。このような場合には、ポリシーに違反するコンテンツを利用者が目にする頻度の上限を推定することができます。例えば、テロリストのプロパガンダに関する上限値が0.04%だった場合、その期間のFacebookまたはInstagramでの10,000回の閲覧のうち、テロリストのプロパガンダに関するポリシーに違反したコンテンツは4回未満であったと推定されます。

ここで大切なことは、違反コンテンツの表示頻度が非常に低く上限値しか表示できない場合、レポート期間中にこの上限値が数100分の1ポイント変化する可能性があるということです。しかしこのような小さな変化は、統計的に有意とはならないでしょう。このような小さな変化は、サービス上でのこの違反コンテンツの表示頻度における実際の違いを示すものではありません。

●表示頻度を測定する理由

投稿されたコンテンツ量ではなく、コンテンツの閲覧回数を推定しているのは、そのコンテンツがFacebookまたはInstagramで利用者にどれだけ影響を与えたかを判断したいからです。違反コンテンツは1回公開されると、1,000回、100万回と閲覧されることもあれば、まったく閲覧されない可能性もあります。違反コンテンツの公開量ではなく、違反コンテンツの閲覧数を測定する方が、コミュニティへの影響をより正確に反映することができるのです。違反コンテンツの表示頻度は少なくても、弊社サービスへの影響は大きくなる可能性があります。弊社のサービスにおけるコンテンツの全体的な閲覧数が多いためです。

利用者の画面にコンテンツが表示されるとコンテンツ閲覧数を記録します。具体的には、次のような場合があります。

- 投稿を閲覧した時ー投稿に複数のコンテンツが含まれていても、その投稿自体に閲覧数が割り当てられます
- 写真や動画プレーヤーをクリックして拡大した時ーその写真や動画に閲覧数が割り当てられます

●表示頻度の推定にサンプリングを使用する

FacebookまたはInstagram上のコンテンツの閲覧数をサンプリングし、表示頻度を推定しています。

このため、閲覧数のサンプルとそこに表示されているコンテンツを手作業でレビューします。弊社のポリシーに基づいて、そのサンプルが違反している／違反していないとのラベル付けをします。このサンプリングを行うチームは、サンプリングした閲覧で投稿のすべてのコンテンツが表示されていなくても、投稿全体に違反がないかどうか確認します。

これらサンプル中の違反コンテンツの一部を活用し、違反コンテンツの全閲覧の割合を推定します。なお、すべての違反カテゴリーに関してFacebookおよびInstagramのすべての部分からサンプリングするわけではありません。

特定の違反カテゴリーには層別サンプリングを使用し、コンテンツ閲覧に違反が含まれている可能性が高いことを文脈が示していた場合にサンプル率を高めています。例えば、違反行為がニュースフィードよりもグループでより頻繁に閲覧されていた場合、グループでの閲覧をニュースフィードでの閲覧よりも高い割合でサンプリングします。その理由のひとつは、サンプリングによる不確実性を減らすためです。この不確実性を表現するために、例えば10,000回の閲覧のうち18~20回が成人のヌードと性的行為に関する違反であるというように、値の範囲を示しています。この範囲は95%の信頼性を反映しています。つまり、毎回異なるサンプルを用いて100回測定した場合、真の数値は100回のうち95回がこの範囲内に収まると予想されます。

閲覧頻度が非常に低い違反カテゴリーでは、正確な表示頻度を推定するためにサンプリングに非常に多くのコンテンツサンプルが必要となります。このようなカテゴリーの違反行為については、層別サンプリングではなく、ランダムサンプリングを実施します。このようなカテゴリーでは、上限を推定することしかできません。つまり、違反している閲覧の表示頻度がその上限以下であることは間違いありませんが、どの程度低いかは正確に示すことはできません。この上限値に対する信頼性も95%です。投稿されたコンテンツの量ではなく閲覧の頻度を調べるのは、そのコンテンツがFacebookおよびInstagram上でどれだけ利用者に影響を与えたかを判断するためです。違反コンテンツは、

一度公開されても、1,000回、100万回、あるいはまったく閲覧されない可能性もあります。違反コンテンツの公開量ではなく、違反コンテンツの閲覧数を測定する方が、コミュニティへの影響をより正確に反映することができるのです。違反コンテンツの表示頻度は少なくても、弊社サービスへの影響は大きくなる可能性があります。弊社のサービスにおけるコンテンツの全体的な閲覧数が多いためです。

●警告

サンプルをラベル付けする際、違反コンテンツを非違反コンテンツとラベル付けしたり、あるいはその逆が起こることがあります。このようなミスの割合が相対的に表示頻度に影響する可能性があります。Facebookでは監査を実施してミスを測定し、表示頻度算出を調整しています。暴力や過激な描写を含むコンテンツに関しては、不快に思う利用者がいる可能性のある投稿に警告カバーを表示していますが、表示頻度の算出はカバーを追加する前のコンテンツ閲覧を考慮しています。

●Facebook上における偽アカウントの割合

Facebook上における偽アカウントの割合は、月あたりのアクティブなFacebook偽アカウントの割合を推定したものです。違反コンテンツの表示頻度とは異なり、偽アカウントの表示頻度は、利用者がそれらのアカウントとの接触を持つことがなくても、Facebook上のアクティブな偽アカウントの数に比例して利用者への影響があると仮定しています。

偽アカウントの表示頻度を推定するため、弊社は月間アクティブユーザをサンプリングし、偽アカウントかどうかラベル付けしています。過去30日間にウェブサイトまたはモバイルデバイスからログインしてFacebookを訪問した登録済みのFacebookユーザ、またMessengerアプリを使用した（Facebookユーザとして登録済みの）ユーザを月間アクティブユーザと定義しています。

措置を講じたコンテンツ

更新日：2021年5月19日

私たちは、Facebookの基準に反する行為を行ったコンテンツ（投稿、写真、動画、コメント等）またはアカウントの数を測定しています。この指標は、弊社の実施規模を示しています。措置を講じるとは、FacebookやInstagramからコンテンツを削除すること、利用者を不快にさせる可能性のある写真や動画上に警告カバーを表示すること、アカウントを無効にすることなどです。法的処置に至った場合はそのコンテンツは追加カウントしていません。

措置を講じたコンテンツという指標が、違反発見の効率性や違反がコミュニティに与えた影響を示していると思われるかもしれません。しかし、措置を講じたコンテンツの量は、ストーリーの一部に過ぎず、違反行為発見に要した時間や、FacebookやInstagram上の違反行為を利用者が見た回数は反映されません。

この指標は、コントロールできない外的要因によって増減します。例えばサイバー攻撃で、スパマーが同じ悪意のあるURLを含む1,000万件の投稿をシェアしたとします。私たちはこのURLを検出した後、その1,000万件の投稿を削除します。措置を講じたコンテンツ数は1,000万件となり、非常に高い数値となります。しかしこの数字は、必ずしも私たちがスパムに対処する能力を高めたことを反映しているのではなく、スパマーがその月に、検出しやすい単純なスパムでFacebookを攻撃することを決めたことを反映しています。また、措置を講じたコンテンツとは、そのスパムが実際どれだけ利用者に影響したかを示すものでもありません。利用者はそのスパムを数回見たかもしれないし、数百回、数千回見たかもしれません（この情報は表示頻度に表れます）。サイバー攻撃の後は、今後の検出力が向上したとしても、措置を講じたコンテンツの数は劇的に減少する可能性があります。

コンテンツは、投稿、写真、動画、コメント等さまざまなものが含まれます。

●コンテンツと措置の数え方

コンテンツの数え方は複雑で、時間をかけて進化してきました。2018年7月には方法を更新し、弊社のポリシーに違反したとして措置を講じたコンテンツの数を明確にし、また今後も最も正確で意味のある指標を提供するというコミットメントの一環として、方法の成熟と改善を続けていきます。全体的には、ポリシー違反で措置を講じたコンテンツの総数を正確に表すことを目指しています。

FacebookとInstagramでは、コンテンツの数え方にいくつか違いがあります。

Facebookでは、写真や動画のない投稿や写真や動画が1枚の投稿は、1つのコンテンツとして数えます。つまり、写真が1枚で違反している投稿、文章で違反している投稿、文章と写真1枚で、どちらかあるいはどちらも違反している投稿の場合、削除されても1つのコンテンツとして数えます。

Facebookへの1件の投稿に複数の写真や動画が含まれている場合、各写真や各動画を1つのコンテンツとして数えます。例えば、Facebookへの投稿1件に4枚の写真が含まれており、そのうち2枚を違反として削除した場合、削除した写真ごとにカウントして2つのコンテンツに措置を講じたと数えます（写真2枚がそれぞれカウントされる）。投稿全体を削除した場合は、その投稿もカウントします。例えば、Facebookへの投稿1件に4枚の写真が含まれており、その投稿全体を削除した場合、措置を講じたコンテンツは5つになります（写真4枚と投稿1件がそれぞれカウントされる）。投稿から一部の写真や動画のみを削除した場合は、削除したコンテンツだけを数えます。

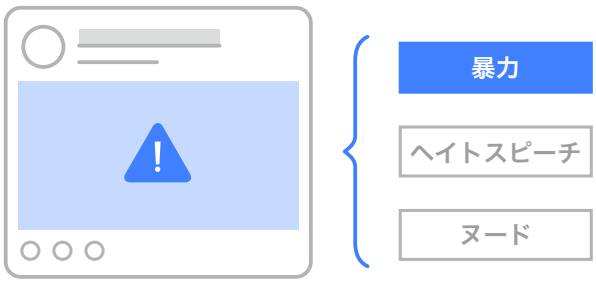
Instagramでは、違反するコンテンツが含まれている場合、投稿全体を削除し、投稿に含まれていた写真や動画の数に関わらず、措置を講じたコンテンツ1件としてカウントします。

時には、1つのコンテンツが複数の基準に違反している場合があります。このような場合、測定を目的としているため、主要な違反のみを措置の理由として考えます。通常、主要な違反とは最も深刻な基準の違反とします。その他のケースでは、レビュー担当者に違反の主な理由についての判断を任せます。

●違反行為の表示方法

コンテンツに措置を講じる場合、そのコンテンツが違反したポリシーでラベル付けします。レビュワーがレポートを見る時、まずそのコンテンツがポリシーに違反しているかどうか選択します。違反と判断した場合は、違反グループにラベル付けします。

これまでは、レビュワーが判断する際、違反をラベル付けする必要はありませんでした。代わりに、利用者がレポートを提出する際の情報を参考にしていました。2017年、レビュープロセスを改善し、レビュワーがコンテンツを削除した理由をより詳細に記録することで、より正確な指標を確立することができました。検出テクノロジーも改善し、違反が発見された時、フラグが立てられた時、あるいは削除された時に、レビュワーの判断と同じ違反ラベルを使用してラベル付けをするようにしました。



特定の基準違反について措置を講じたコンテンツを数えるために、措置を講じるごとにその違反をラベル付けする必要があります。

●Facebook上で偽アカウントとして措置を講じたアカウント

偽アカウントについては、「措置を講じたコンテンツ」ではなく「措置を講じたアカウント」として報告しています。「措置を講じたアカウント」とは、偽アカウントとして無効にしたアカウントの数です。

●警告

措置を講じたコンテンツとアカウントには、そもそもコンテンツやアカウントが作成されないようにブロックしたケースは含まれていません。ブロックするケースには、高頻度で投稿しようとするスパマーや偽アカウントの作成を検出した場合があります。このようなブロックしたケースを含めた場合、無効化した偽アカウントや削除したスパムコンテンツ数が劇的に増加してしまいます（1日数百万件が見込まれます）。

URLを強制的に削除すると、そのリンクを含む現在または将来のコンテンツがすべて削除されます。利用者がこのコンテンツをFacebook上で表示しようと試みたかに基づいて措置を講じたコンテンツの数を測定します。

●措置を講じたアカウント、ページ、グループ、イベントを測定する方法

Facebookのユーザアカウント、ページ、グループ、イベント内には大量のコンテンツが存在します。これらのオブジェクトのひとつが全体としてその中のコンテンツや行動に基づいて、Facebookのポリシーに違反することがあります。通常、あらゆるコンテンツをレビューしなくても、そのアカウント、ページ、グループ、イベントが基準に違反しているかどうか判断することができます。アカウント、ページ、グループ、イベントを無効化した場合、利用者はその中のすべてのコンテンツに自動的にアクセスできなくなります。

コミュニティ規定施行レポートに記載されている指標では、アカウント、ページ、グループ、イベントに含まれるコンテンツのうち、それらのオブジェクトのレビュー中に違反していると判断し、明示的に対処したもののみを数えています。そのコンテンツを含むアカウント、ページ、グループ、イベントを無効にした際自動的に削除されたコンテンツについては数えていません。

Facebookの偽アカウントを除き、このレポートでは措置を講じたアカウント、ページ、グループ、イベントに関連する指標は現在含んでおらず、これらオブジェクト内のコンテンツのみが対象となっています。

事前対応率

更新日：2021年5月19日

この指標は、利用者の報告を受ける前にFacebookが発見しフラグを立て、措置を講じたあらゆるコンテンツまたはアカウントの割合を示しています。この指標は、弊社の違反検出の効率性を示す指標として使用しています。

機械学習技術への投資は、より迅速に検出を行うために重要です。

弊社では、機械学習と訓練された専門家チームがバランス良く違反コンテンツのレビューと措置に取り組んでいます。

違反の可能性のあるコンテンツは、違反によっては高い割合で事前に検出することができるため、ほとんどのコンテンツは利用者に発見される前にFacebookが発見しフラグを立てています。これは特に、基準に違反する可能性のあるコンテンツを自動的に特定する機械学習技術を構築できた場合に当てはまります。

このような技術は非常に有望ですが、あらゆる種類の違反に効果を発揮するにはまだ数年かかります。例えば、機械が文脈やニュアンスを理解するにはまだ限界があり、特にテキストベースのコンテンツの場合はこれが顕著です。このため、特定の違反の事前検出に関してはさらに課題があります。

この指標は外的要因によって増減します。例えばサイバー攻撃で、スパマーが同じ悪意のあるURLを含む1,000万件の投稿をシェアしたとします。利用者から報告される前に弊社が悪意のあるURLを検出した場合、事前対応率はサイバー攻撃中は上がり、その後下がります。これは期間中に検出テクノロジーが変わらなくても同じです。またこの指標は、プロセスの方法やツールの変更によっても増減します。例えば、検出テクノロジーが向上すれば事前対応率は上がり、利用者からの報告が改善されて事前検出への依存度が下がれば、事前対応率は下がります。

この指標は措置を講じたコンテンツの量に基づいているため、同様に考慮する点が多くあります。Facebookの事前対応率には、違反コンテンツを検出するために要した時間や検出される前に閲覧された回数は反映されていません。また、検出できなかった違反の数や、そのコンテンツが閲覧された回数も反映されていません。弊社が事前に検出するコンテンツの割合は非常に高く、99%に達するカテゴリ—もありますが、残りのわずかな割合でも利用者に大きな影響を与える可能性があります。

●事前対応率の算出方法

この割合は、FacebookやInstagramの利用者から報告を受ける前にFacebookが発見しフラグを立て、措置を講じたコンテンツの数を、措置を講じたコンテンツの総数で割って算出されています。

Facebookの偽アカウントに関しては、利用者から報告を受ける前にFacebookが発見しフラグを立て、偽アカウントであることを理由に無効化されたアカウントの割合として算出しています。この指標は、利用者が報告する前にFacebookが発見しフラグを立て、無効化されたアカウントの数を、偽アカウントであることを理由に無効化されたアカウントの総数で割って算出しています。

●警告

事前対応率は、利用者からの報告をコンテンツに厳密に帰属させ算出しています。例えば、あるページが報告され、そのページのレビューを行い、ページ内の違反コンテンツを特定し対処した場合、そのコンテンツに事前にフラグを立てたことを報告します（ただし、利用者からの報告が追加であった場合を除く）。このように厳密な方法で利用者からの報告を帰属させることは理想的ではありませんが、それを上回る方法はまだ見つかっていません。

異議を申し立てたコンテンツ

更新日：2021年5月19日

Facebookでのポリシー違反に関して、ポリシー違反として措置を講じた後に利用者が異議を申し立てたコンテンツ（投稿、写真、動画、コメント）の数を測定しています。

Facebookの決定に異議を申し立てるには、コンテンツが削除された旨、あるいは警告表示となった旨が通知された後に、利用者は「レビューをリクエストする」というオプションを選択します。レビューがリクエストされると、Facebookは投稿を再度レビューし、Facebookのコミュニティ規定に従っているか判断します。利用者は、このプロセスを通して弊社が間違っていると考えていることを伝えることができます。その結果、私たちは公正なシステムを構築することができるのです。

この指標は、コンテンツに対する弊社の判断が正しいことを示すものではありません。利用者はさまざまな理由で異議申し立てを行う可能性があるからです。

Facebookは各四半期（1月1日から3月31日等）に異議申し立てが提出されたコンテンツの総数を報告します。つまり、その数字は、措置を講じた、あるいは同じ四半期に復元されたコンテンツを直接比較することはできません。復元されたコンテンツの中には、前期に申し立てられたものもあるでしょうし、今期申し立てたコンテンツでも復元は来期になるということもあるでしょう。

この数字は、外的要因あるいは内的プロセスによって増減する可能性があります。例えば、オフラインイベントやスパム攻撃によりFacebook上で違反投稿が増えたとします。その結果、Facebookは通常より多くの投稿に措置を講じます。多くのコンテンツに措置を講じると、比較的多い数の異議申し立てが出される可能性があります。異議申し立ての急増は、Facebookがより多く間違った判断をしたということではなく、より多くの利用者が私たちの判断に異議を申し立てたことになります。

コンテンツは、投稿、写真、動画、コメントなどさまざまなものがあります。それぞれのコンテンツを数える方法は複雑で、時間とともに進化しています。措置を講じたコンテンツの指標についてはこちらをご覧ください。

●Facebookの異議申し立てプロセスについて

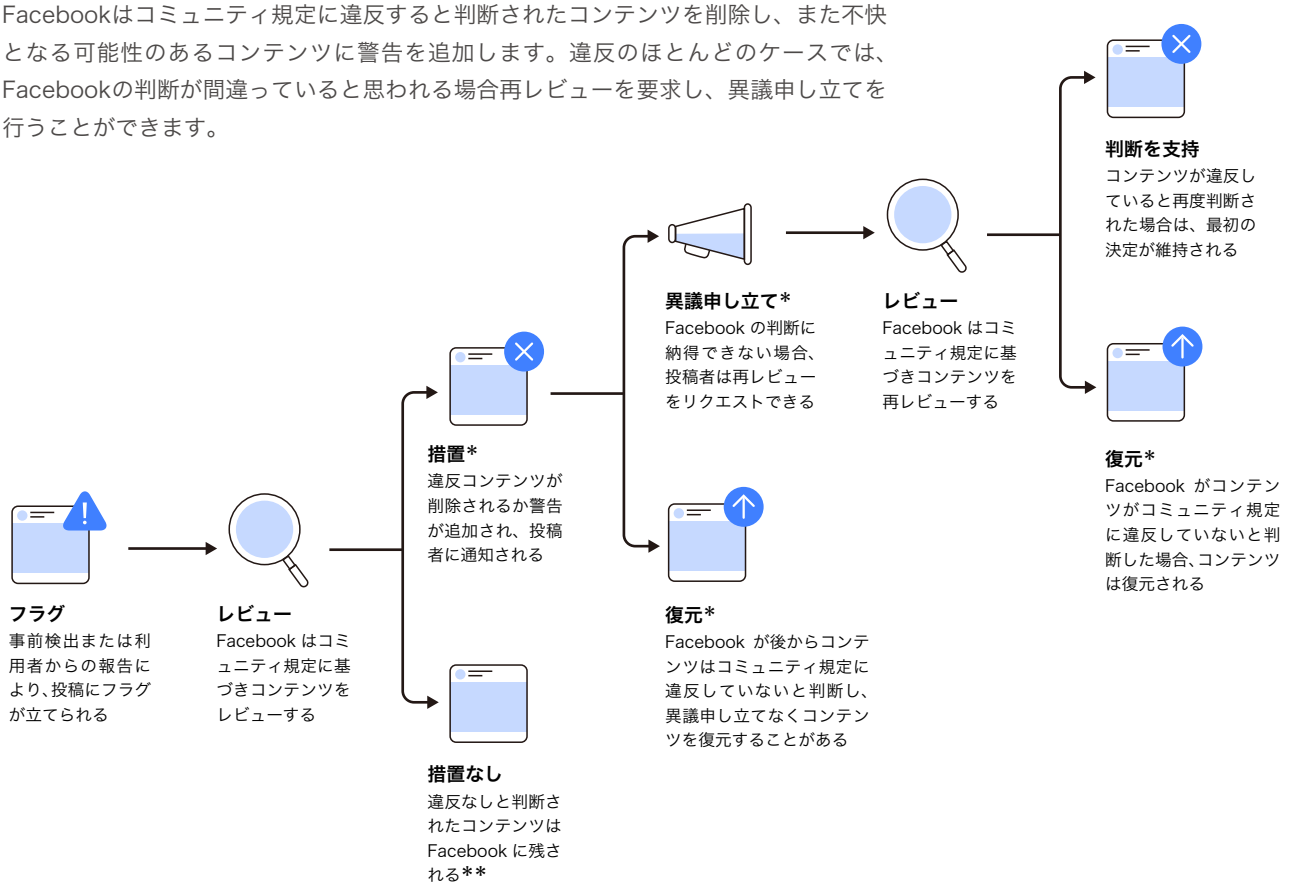
ある記事が、Facebookのポリシーに違反しているため削除すると判断されたとします。投稿者は通知を受け、レビュー要求か決定の受け入れか選択肢を与えられます。

投稿者がレビューを選択すると、コンテンツは改めてレビューのために再提出されます。再レビューの間はそのコンテンツはFacebook上の他の利用者に表示されません。レビュワーは、その投稿が以前レビューされたことがあることを知らされていません。

レビュワーが最初の判断に同意した場合、コンテンツはFacebookから削除されたままになります。しかし、レビュワーが判断を覆し削除すべきではなかったと判断した場合には、コンテンツは第3のレビュワーのもとに送られます。このレビュワーの判断により、コンテンツをFacebook上に残すかどうかが決定されます。

Facebookのコンテンツの異議申し立てと復元プロセスについて

Facebookはコミュニティ規定に違反すると判断されたコンテンツを削除し、また不快となる可能性のあるコンテンツに警告を追加します。違反のほとんどのケースでは、Facebookの判断が間違っていると思われる場合再レビューを要求し、異議申し立てを行うことができます。



＊　これら指標はコミュニティ規定施行レポートで報告されています。

＊＊ 近年、Facebook は報告されたコンテンツに措置が講じられなかった際にも、異議申し立ての提案を開始しました。このような異議申し立ての指標はレポートに記載されていません。

●異議申し立てができる内容

現在、Facebookではほとんどの違反行為に対して異議申し立てを行うことができます。ただし、児童搾取の画像など、安全性が非常に懸念される違反については、異議申し立てを行うことはできません。

また、措置を講じたコンテンツだけではなく、報告はされたが措置は講じられていないコンテンツに対しても異議申し立てを受け始めました。このような報告のみに対する異議申し立ては、コミュニティ規定施行レポートには記載されていません。

●アカウント、ページ、グループ、イベントに対する異議申し立ての測定方法

現在コミュニティ規定施行レポートには、弊社が措置を講じたアカウント、ページ、グループ、イベントに対する異議申し立ての指標は記載されていません。

復元されたコンテンツ

更新日：2021年5月19日

ポリシー違反については、最初に措置を講じた後復元したコンテンツ（投稿、写真、動画、コメント等）の数を測定しています。

「復元」とは以前に削減したコンテンツを元に戻したり、警告カバーを表示させたコンテンツから警告表示を外したりすることを指します。

私たちは、異議申し立てに基づいて復元したコンテンツと直接異議申し立てはなかったが復元したコンテンツを報告します。異議申し立てがなかったがコンテンツを復元する場合には次のような理由があります。

- 同じ内容の複数の投稿を誤って削除してしまった場合。この場合、1人が異議申し立てをすれば、すべての投稿を復元することができます。
- 投稿者が異議申し立てをする前に、レビューに誤りがあったことが判明しコンテンツを復元した場合。
- 悪質と思われるリンクを含む投稿を削除し、その後そのリンクが悪質ではないことが分かった場合。この場合、該当する投稿を復元することができます。これは特にスパムに当てはまります。

この指標をコンテンツに措置を講じる際に起きる誤りを表示するものと思われるかもしれませんが、前述の悪質なリンクの例のように、投稿を復元するということが必ずしも誤りがあったことを示すとは限りません。

私たちは、各四半期（例えば1月1日から3月31日）にFacebookが復元した総コンテンツ量を報告しています。つまり、この数字は同じ四半期に措置を講じたり異議を申し立てるコンテンツとは直接比較できないことに注意してください。例えば、復元されたコンテンツの中には前期に異議を申し立てられたものが含まれているかもしれませんし、異議を申し立てたコンテンツの中には次期に復元されるものがあるかもしれません。

コンテンツは、投稿、写真、動画、コメントなどさまざまなものがあります。それぞれのコンテンツを数える方法は複雑で、時間とともに進化しています。措置を講じたコンテンツの指標についてはこちらをご覧ください。

●アカウント、ページ、グループ、イベントに対する異議申し立ての測定方法

現在コミュニティ規定施行レポートには、弊社が復元したアカウント、ページ、グループ、イベントに対する指標は記載されていません。