

Facebookは、ポリシー違反によってサービスの利用者に生じる影響を最小限に抑えることを目標としています。この目標に向けた取り組みの効果を測るために違反コンテンツの表示頻度を測定します。

●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」ということとなります。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。

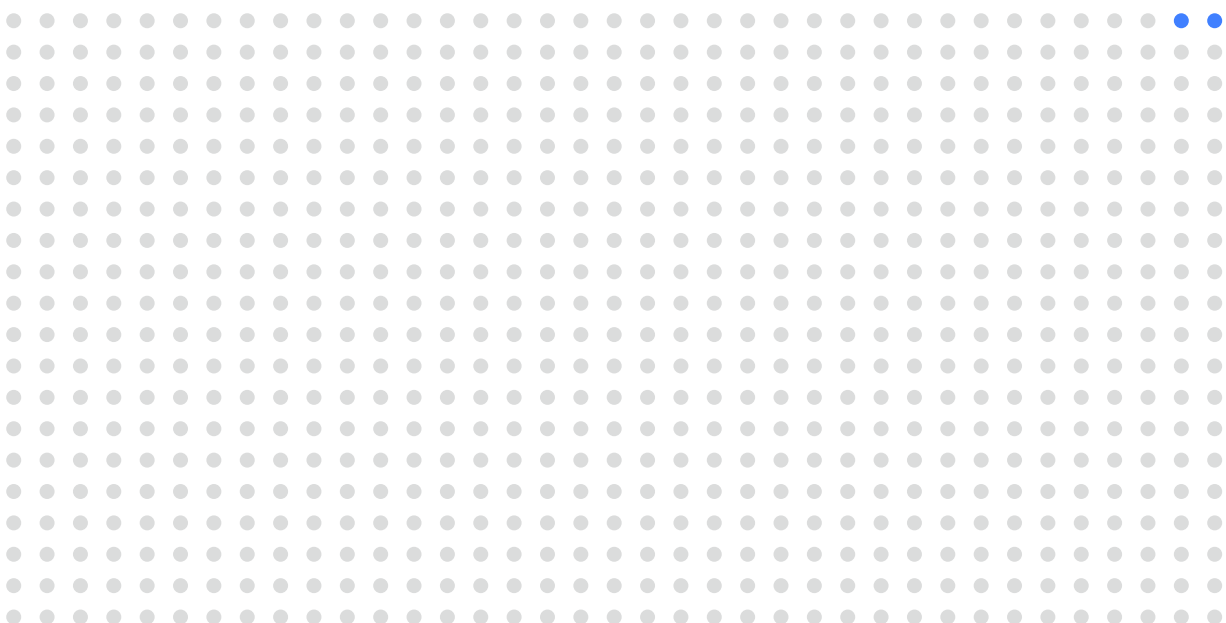
●表示頻度の測定方法

違反コンテンツの表示頻度は、FacebookまたはInstagramにおけるコンテンツ表示のサンプルを用いて推定します。この値は、違反コンテンツの推定表示数を、FacebookまたはInstagram上の全コンテンツの推定表示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の表示頻度が0.18%～0.20%である場合、10,000回表示されるコンテンツのうち、平均で18件～20件が成人のヌードと性的行為に関する弊社規定に違反しているということを意味します。

1ドット = 10件の表示

表示総数 10,000件

違反コンテンツ表示 20件



表示頻度が0.20%であれば、10,000件の表示のうち20件で違反コンテンツが表示されることを意味します。数値が小さい場合であっても、利用者に重大な影響を及ぼす場合があります。

弊社のサービスにおいて発生頻度が非常に低いタイプの違反もあります。違反コンテンツを利用者が目にする可能性は非常に低く、また、利用者がそれを見る前に弊社はそうしたコンテンツの大半を削除します。結果として、多くの場合、表示頻度を正確に推定するのに十分な数の違反サンプルを得ることができません。こうした場合には、弊社ポリシーに違反するコンテンツを利用者が見る頻度の上限値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%だった場合には、当該期間におけるFacebookまたはInstagram上での10,000件の表示のうち、テロリストのプロパガンダに関するポリシーに違反するコンテンツを含む表示が4件以下だと弊社が推定していることを意味します。

ある種の違反の表示頻度が非常に低く、弊社が提示できるのが上限値のみである場合、報告対象期間の間に、100分の数ポイントの変化がこの上限値に生じ得る点に留意することが重要です。しかし、こうした小さな変化は統計上有意でない場合があります。このような場合、こうした小さな変化は、サービスでの違反コンテンツの表示頻度に実際に生じた差異を示すものではありません。

●表示頻度の測定理由

弊社は、コンテンツがFacebookまたはInstagramの利用者に与えた影響の規模を判断するため、コンテンツの投稿数ではなく、コンテンツが表示された頻度を推定しています。1件の違反コンテンツが公開されたとしても、それが1,000回、100万回にわたり表示される場合もあれば、一切表示されない場合もあります。公開された違反コンテンツ数ではなく違反コンテンツの表示について測定すると、コミュニティへの影響をより反映したものとなります。ただし、表示頻度の数値が小さくても、弊社サービスではコンテンツの表示数全体の規模が大きいため、大きな影響を及ぼす場合もあります。

弊社は、利用者のスクリーンに1件のコンテンツが表示された場合に、コンテンツが1回表示されたものとして記録します。利用者が次のような行為を行った場合に表示が1回発生したものとして扱われます。

- 投稿を見た場合（投稿内に複数のコンテンツが含まれる場合であっても、表示数はその投稿に割り当てられます）
- クリックして写真や動画プレイヤーを拡大した場合（表示数は、その写真や動画に割り当てられます）

●表示頻度を推定する際のサンプリングの利用方法

弊社では、FacebookまたはInstagramで表示されるコンテンツをサンプル抽出することにより表示頻度を算出しています。

そのため、表示されたサンプルとその中で表示されるコンテンツについては、手動で確認しています。その後、サンプルは、弊社のポリシーに従って違反に該当するもの、該当しないものに分類されます。サンプリングを担当するチームは、サンプル抽出された表示で、投稿内のすべてのコンテンツが表示されない場合であっても、その投稿全体に違反がないか審査します。

そして、違反コンテンツに該当するサンプルの割合から、全体における違反コンテンツの表示割合を推定します。違反の各種類について、FacebookまたはInstagramのすべての部分からサンプルを抽出しているわけではない点にご留意ください。

一定の種類違反に関して、弊社では、階層化されたサンプリングを行っており、文脈上、コンテンツの表示に違反が含まれている可能性が高いことが示されている場合には、サンプルレートが増やされます。例えば、ニュースフィードよりもグループにおける違反の表示頻度が高い場合、ニュースフィードよりも高い確率でグループの表示からサンプルを抽出します。この手法は、サンプリングの不確実性の軽減を目的の1つとして採用されています。不確実性は、指標の値を一定の幅を持った値域で示すことによって表現しています。例えば、成人のヌードと性的行為について10,000件当たり18件～20件が違反していると表すのはその一例です。この幅により、値に95%の信頼性を与えることができます。これは、弊社が毎回異なるサンプルを用いてこの測定を100回行った場合、うち95回で真正な数値が導き出されていると考えられることを意味します。

表示の頻度が非常に低いタイプの違反では、表示頻度の正確な値を推定するのに、サンプリングには非常に多くの

コンテンツのサンプル数が必要となります。こうした種類の違反に関しては、層化抽出法を用いるのではなく、ランダム・サンプリングを行います。この種の違反に関しては、推定できるのは上限値のみです。つまり、違反コンテンツの表示頻度が示された値以下であることについて信頼できるものの、どの程度下回るかを正確に述べることはできないということです。この上限に関する信頼性も95%です。コンテンツがFacebookまたはInstagramの利用者に与えた影響の規模を判断するため、コンテンツの投稿数ではなく、コンテンツが表示された頻度。1件の違反コンテンツが公開されたとしても、それが1,000回、100万回にわたり表示される場合もあれば、一切表示されない場合もあります。公開された違反コンテンツ数ではなく違反コンテンツの表示について測定すると、コミュニティへの影響をより反映したものとなります。ただし、表示頻度の数値が小さくても、弊社サービスではコンテンツの表示数全体の規模が大きいため、大きな影響を及ぼす場合もあります。

●注意事項

サンプルコンテンツの分類を行う担当者が、違反コンテンツを違反していないと分類したり、違反にならないコンテンツを違反として分類したりするなど、誤った判断をする場合もあります。これらの誤認の相対的な割合が、表示頻度測定に影響を及ぼすことがあります。弊社では監査によりエラーを測定し、そのエラーを考慮して算出した表示頻度を調整します。暴力や過激な描写を含むコンテンツなどの分野に関しては、ある利用者の心を乱すおそれのある投稿に弊社がぼかしを入れた場合、表示頻度の計算には、ぼかしが入られる前のコンテンツの表示のみが算入されます。

●Facebookにおける偽アカウントの表示頻度

Facebookにおける偽アカウントの表示頻度は、アクティブな偽のFacebookアカウントの割合を月ごとに推定したものです。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、偽アカウントが利用者の目に触れたり、交流したりしなかった場合であっても、利用者に対する影響とFacebook上のアクティブな偽アカウント数が比例しているという前提のもとに使用されます。

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ーをサンプル抽出し、そのアカウントが偽のアカウントであるか否かを分類します。弊社では、測定日から過去30日の間に弊社ウェブサイトやモバイルデバイスを通じてFacebookにログインもしくは訪問した、またはMessengerアプリを利用したFacebookの登録利用者を月間アクティブユーザー（MAU）と定義します。

措置を講じたコンテンツ

弊社規定に対する違反を理由として弊社が措置を講じたコンテンツ（投稿、写真、動画、コメントなど）やアカウントの数を測定しています。この指標は、弊社による施行活動の規模を示すものです。「措置を講じる」には、FacebookまたはInstagramからコンテンツを削除すること、一部の利用者にとって不快である写真や動画に警告付きでぼかしを入れること、またはアカウントを停止することが含まれます。法執行機関に付託するコンテンツは、集計には加えません。

措置を講じたコンテンツに関する指標を、弊社が弊社コミュニティにおける違反やこれらの違反による影響をいかに効果的に検知しているかを示すものと捉えるのは簡単ですが、措置を講じたコンテンツ数は、弊社の違反コンテンツの抑制に向けた取り組みの成果を部分的に表すものに過ぎません。この指標には、違反の検知に要した時間や、FacebookまたはInstagram上で違反コンテンツが表示されている間に利用者がそれを目にした回数は反映されていません。

この指標は、弊社の支配が及ばない外部要因によって上下する場合があります。例えば、スパム送信者が同一の悪意あるURLを掲載した1,000万件の投稿をシェアするというサイバー攻撃が発生したとします。URLを検知した後、弊社はただちにこの1,000万件の投稿を削除します。1,000万件のコンテンツに措置が講じられた結果、措置を講じたコンテンツ数は極端に増加します。しかしながら、この値は、スパムに対する弊社の対策が向上したことを必ずしも表すものではありません。むしろ、スパム送信者がその月に、洗練されていない、検知が容易なスパムによってFacebookを攻撃したことを示しています。措置を講じたコンテンツ数は、スパムが実際に利用者にも及ぼした影響の度合いを示すものでもありません。利用者が実際にそのスパムを目にした回数は、数回、数百回、数千回のいずれであるか把握できないからです（こうした情報は、表示頻度の値として入手できます）。サイバー攻撃が沈静化した後には、弊社の検知技術が向上していても、措置を講じたコンテンツ数は劇的に減少します。



1件のコンテンツは、投稿、写真、動画やコメントなど、複数の要素で構成される場合があります。

●コンテンツおよび措置の集計方法

個々のコンテンツを数える方法は複雑であり、時間とともに発展してきました。弊社は2018年7月に、Facebookのポリシーに違反するとして措置を講じた個別のコンテンツ数を明確化するため、採用する手法を更新しました。今後も、最も正確で有意義な指標を提示する取り組みの一環として、手法を継続的に発展、向上させていきます。一般的に弊社が目指すことは、Facebookのポリシーの違反に関して弊社が措置を講じたコンテンツの正確な合計数を提供することです。

FacebookとInstagramでは、コンテンツを数える方法にいくつか違いがあります。

Facebookでは、写真や動画がない投稿、または写真や動画が1つだけの投稿は、1件のコンテンツとして集計されます。すなわち、写真1枚の投稿であって違反があるもの、テキストのみの投稿であって違反があるもの、テキストと写真1枚の投稿であって、そのいずれか、または両方に違反があるものはすべて、削除された場合には措置を講じたコンテンツが1件だったと数えます。

1件のFacebook投稿に複数の写真や動画が含まれる場合、写真や動画それぞれを1件のコンテンツとして数えます。

例えば、4枚の写真を含む1つのFacebook投稿から、違反のある写真2枚を削除した場合、削除した各写真について1件ずつ数に入れ、措置を講じたコンテンツが2件だったと数えます。投稿全体を削除した場合には、その投稿自体も1件として数に入れます。例えば、4枚の写真とともにFacebook投稿自体を削除した場合、各写真について1件ずつ、投稿自体について1件を数に入れ、措置を講じたコンテンツが5件だったと数えます。投稿から削除したのが添付された写真と動画のみであった場合、当該コンテンツ数のみを数えます。

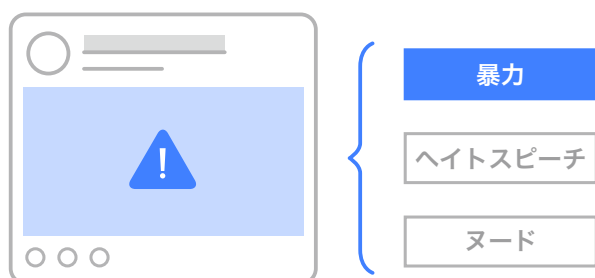
Instagramでは、投稿に違反コンテンツが含まれる場合には投稿全体を削除しており、これを措置を講じたコンテンツ1件と数えています。投稿に含まれる写真や動画の数は関係ありません。

時として、1つのコンテンツが複数の規定に違反していることがあります。指標測定の目的に沿うよう、弊社では、措置の測定は主な違反内容に対するもの1件にとどめています。通常、この措置は、最も深刻な規定の違反に対するものとなります。その他に、違反の主な理由に関する判断を審査担当者に委ねる場合もあります。

●違反の分類方法

弊社がコンテンツに対して措置を講じる際は必ず、違反のあったポリシーでコンテンツを分類します。報告を確認する際に、審査担当者は、そのコンテンツが弊社のポリシーに違反しているか否かを最初に評価します。違反していると判断した場合は、違反の種類に応じて分類します。

以前は、審査の判断においては、審査担当者に違反の分類を求めていませんでした。代わりに、利用者が報告する際に弊社に提供する情報に依拠していました。2017年に、弊社は、審査担当者のコンテンツ削除理由に関してより詳細な情報を記録するよう審査プロセスを更新し、これにより、さらに正確な指標を策定できるようになりました。また、弊社の検知技術のアップデートにより、違反の自動検知、フラグ、削除においても、審査担当者による判断と同じ分類が使われるようになりました。



特定の規定違反に関して措置が講じられたコンテンツを集計するには、措置を講じるたびに違反を分類する必要があります。

●偽アカウントとしてFacebook上で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」は、偽アカウントであることが理由で弊社が停止したアカウントの数です。

●注意事項

弊社では、スパム送信者が何度も投稿を試みたり、偽アカウントの作成を試みたりしたことを検知した時点でこれらをブロックします。そのため、この時点でブロックされたコンテンツや、作成がブロックされたアカウントは、措置を講じたコンテンツや措置を講じたアカウントに含まれません。これらのブロックを含めると、停止された偽アカウントや削除されたスパムコンテンツの数は1日あたり何百万件の規模で劇的に増加するためです。

施行の対象がURLである場合には、当該URLのリンクを含む現在のコンテンツまたは今後のコンテンツを削除します。措置を講じたコンテンツ数については、利用者がFacebookで当該コンテンツの表示を試みたかどうかに基づいて測定します。

●アカウント、ページ、グループおよびイベントへの対応措置の測定方法

コンテンツの多くは、Facebookにおける利用者のアカウント、ページ、グループまたはイベントの中に存在しています。アカウント、ページ、グループまたはイベントの中のコンテンツや行為に基づいて、アカウント、ページ、グループまたはイベントが全体として弊社のポリシーに違反したと判断される場合もあります。通常、アカウント、ページ、グループまたはイベントに規定違反があるかを判断する際に、その中のすべてのコンテンツを審査するとは限りません。アカウント、ページ、グループまたはイベントが停止された場合、自動的に、そこにあるすべてのコンテンツが利用者から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盛り込まれた指標においては、弊社の審査において違反があると判断し、明確な措置を講じたアカウント、ページ、グループまたはイベントのコンテンツのみを数に入れています。そのコンテンツを含むアカウント、ページ、グループまたはイベントが停止されたことによって自動的に削除されたコンテンツを数に入れることはありません。

Facebookの偽アカウントに関するものを除き、このレポートには現在のところ、措置を講じたアカウント、ページ、グループまたはイベントに関する指標は盛り込まれていません。レポートに含まれるのは、措置を講じたアカウント、ページ、グループまたはイベントの中のコンテンツのみです。

事前対応率

更新日：2021年7月29日

この指標は、利用者から報告を受ける前に弊社が検知・フラグし措置を講じたすべてのコンテンツとアカウントの割合を示すものです。この指標は、弊社による違反検知の効率性を示すものです。



機械学習技術の強化は、検知速度の向上に不可欠です。

機械学習と、違反コンテンツを審査し対応を行う熟練の専門家チームをバランスよく活用しています。

一部の違反については、違反のおそれがあるコンテンツを高い確率で事前に検知でき、ほとんどのコンテンツに関して、利用者による報告より前に弊社がフラグ・検知しています。この傾向は、規定に違反するコンテンツを自動的に特定する機械学習技術が構築できているケースで特に顕著です。

このような技術の発展は将来的に大いに期待されるのですが、すべての種類の違反に効果を発揮するまでには至っていません。例えば、文脈やニュアンスを解釈する能力には、特にテキストがベースのコンテンツではいまだに限界があります。これは、一定の種類の違反を事前に検知する際の課題となっています。

この指標は、外部要因によって上下する場合があります。例えば、スパム送信者が同一の悪意あるURLを掲載した1,000万件の投稿をシェアするというサイバー攻撃が発生したとします。利用者からの報告を受ける前に悪意のあるURLを弊社が検知した場合、サイバー攻撃が継続している間は、弊社の検知技術の水準に変化がなくても事前対応率は継続的に上昇し、攻撃の終了後に自動的に減少します。この指標は、弊社のプロセスやツールに生じた変更に応じて上下する場合があります。例えば、弊社の検知技術が向上すれば指標も向上しますが、利用者による報告が改善され弊社が事前検知に頼ることが少なくなれば、指標も低下します。

この指標は措置を講じたコンテンツ数を基にしたものであるため、同じような考慮すべき事項が多く適用されます。事前対応率に、違反コンテンツの検知に要した時間や、検知前の表示回数は反映されていません。また、弊社が検知できなかった違反の総数や、そのような違反コンテンツの表示回数も反映されていません。弊社が事前検知するコンテンツの割合が極めて高く、一部のカテゴリでは99%に達していても、残りの小さな割合の違反が利用者に重大な影響を及ぼす場合もあります。

●事前対応率の算出方法

この割合は、FacebookまたはInstagramの利用者から報告を受ける前に弊社が検知・フラグし措置を講じたコンテンツの件数を、弊社が措置を講じたコンテンツの総数で割って算出しています。

Facebookの偽アカウントの場合、この指標は利用者から報告を受ける前に弊社が偽アカウントであることを検知・フラグし停止したFacebookアカウントの割合を示すものとして算出されます。この指標は、利用者から報告を受ける前に弊社が検知・フラグし停止した偽アカウントの件数を、弊社が停止した偽アカウントの総数で割ったものです。

●注意事項

事前対応率の算出では、利用者から寄せられた報告を限定的に数えています。例えば、利用者があるページに関して報告した後、そのページが審査されている間に弊社が当該ページ内の何らかの違反コンテンツを特定して措置を講じた場合、そのコンテンツには事前対応したものとして報告されます。ただし、利用者から違反コンテンツについて具体的な追加報告があった場合はその限りではありません。このように利用者の報告を受けたコンテンツを限定的に数える方法は理想的ではないものの、現時点でのベストな手法となります。

異議申し立て済みのコンテンツ

Facebookにおけるポリシー違反に関して、弊社ポリシーに対する違反を理由として弊社が措置を講じた後に利用者から異議申し立てを受けたコンテンツ（投稿、写真、動画、コメントなど）の数を測定しています。

利用者がFacebookにおける決定に異議申し立てを行うには、自らのコンテンツが削除または警告付きでぼかしが入れられたことについて弊社からの通知を受けた後、「審査のリクエスト」を選択します。審査がリクエストされた場合、Facebookがその投稿を再度審査し、それが弊社のコミュニティ規定に従っているかを判断します。弊社の判断が誤っていると思われる場合、利用者はこのプロセスを通して弊社に知らせることができます。このようなプロセスは、公正なシステムの構築に不可欠です。

この指標を、コンテンツに対する弊社の決定の正確性を示すものと解釈すべきではありません。利用者は、さまざまな理由に基づいて異議を申し立てることができるからです。

この指標では、各四半期（例えば、1月1日から3月31日までの期間）に異議が申し立てられたコンテンツの総数を報告します。この数値は、同一の四半期に措置が講じられたコンテンツや、復元されたコンテンツの数値と直接比較できるものではありません。復元されたコンテンツの中には過去の四半期に異議申し立てがなされているものもあれば、異議申し立ての中には翌四半期に復元されるものもあります。

この指標は、外部要因や内部のプロセスによって上下する場合があります。例えば、オフラインでの出来事やスパム攻撃によりFacebook上の違反投稿が増加したと仮定します。その結果、弊社が措置を講じる投稿数は通常よりも増加します。弊社が措置を講じるコンテンツの増加に伴い、比例して通常より多くの異議が申し立てられる場合があります。この異議の急増は、Facebookによる誤った判断が増加したことを意味するものではなく、単に、弊社の決定に対して異議申し立てを選択した利用者が増えたことを示します。

1件のコンテンツは、投稿、写真、動画やコメントなど、複数の要素で構成される場合があります。個々のコンテンツを数える方法は複雑であり、時間とともに発展してきました。措置を講じたコンテンツの指標について詳しくはこちら。

●Facebookの異議申し立てプロセスの仕組み

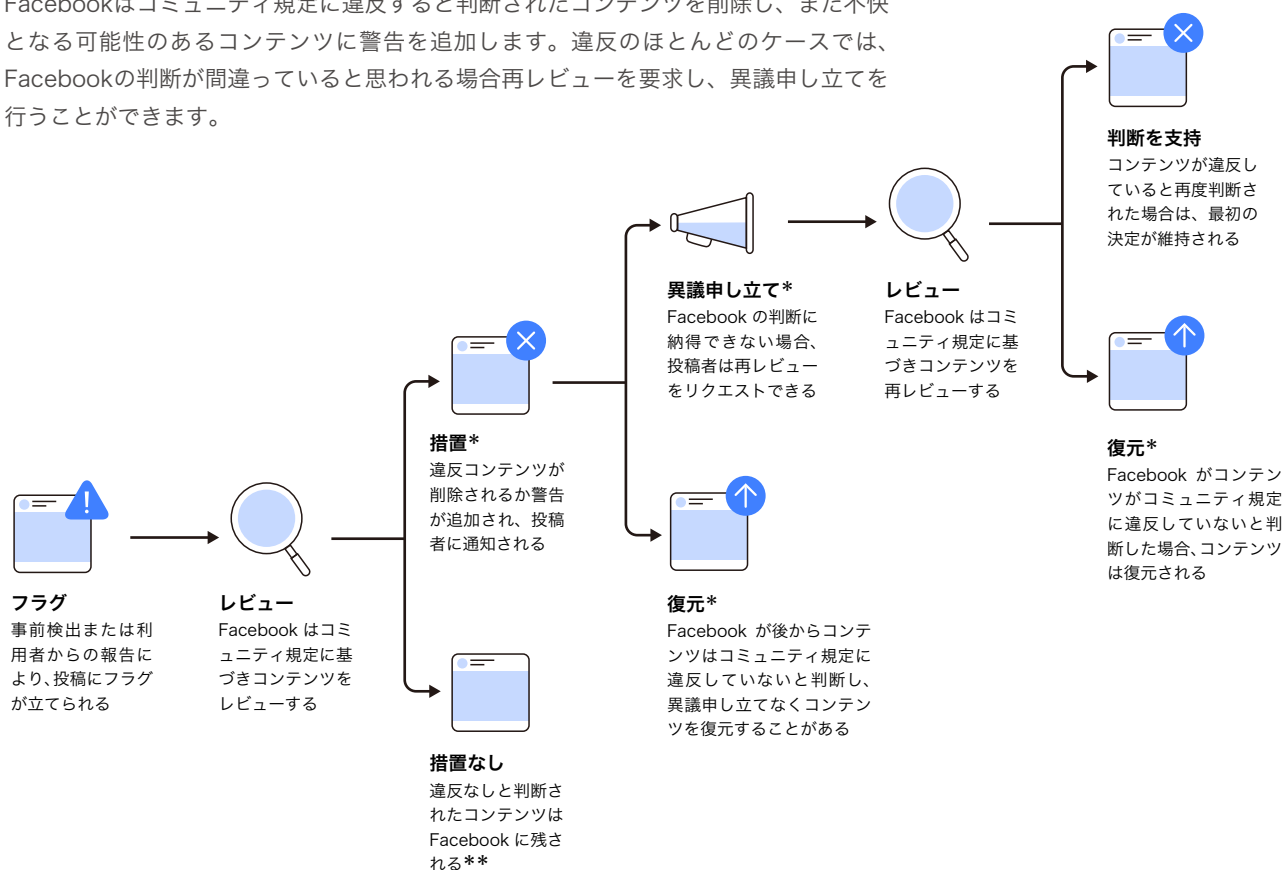
利用者が公開した投稿について、弊社ポリシーへの違反を理由としてFacebookから削除することを弊社が決定したとします。投稿者には、この決定が通知されるとともに、審査をリクエストするか決定を承諾するかを選択権が与えられます。

審査のリクエストを選択した場合、コンテンツが再審査されます。再審査中は、Facebook上で他の利用者がそのコンテンツを見ることはできません。審査担当者には、その投稿が以前に審査されたものであることは知らされません。

審査担当者が元の決定に同意した場合、そのコンテンツはFacebookから削除されたままになります。審査担当者が当初の審査結果に同意せず、削除すべきでない判断した場合、当該コンテンツは別の審査担当者によって3度目の審査かけられます。この審査担当者の判断によって、コンテンツがサービス上に残されるべきか否かが決まります。

Facebookのコンテンツの異議申し立てと復元プロセスの仕組み

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。



* これらの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。

●異議申し立ての対象

現在、Facebook上の大多数の違反タイプについて異議申し立ての機会が与えられています。児童の搾取に関連する画像など、重大な安全上の懸念がある違反については、異議を申し立てることができません。

弊社では、措置が講じられたコンテンツに関する異議申し立てだけでなく、報告されたものの措置が講じられなかったコンテンツに関する異議申し立ての機会提供も開始しています。こうした報告者による異議申し立ては、コミュニティ規定施行レポートに盛り込まれていません。

●アカウント、ページ、グループおよびイベントに対する異議申し立ての測定方法

現時点でコミュニティ規定施行レポートには、アカウント、ページ、グループおよびイベントに対する弊社の対応措置に関する異議申し立ての指標は含まれていません。

復元されたコンテンツ

ポリシー違反に関して、弊社が当初の対応措置後に復元したコンテンツ（投稿、写真、動画、コメントなど）の数を測定しています。

「復元する」とは、過去に削除されたコンテンツを戻したり、一度警告付きでぼかしが入られたコンテンツから警告とぼかしを外したりすることを意味します。

異議申し立てに対応して復元したコンテンツ、さらには異議申し立てが直接なかったものの弊社が復元したコンテンツを報告します。次のような場合、Facebookは異議申し立てがなくてもコンテンツを復元します。

- 同一コンテンツの複数の投稿を削除する際に誤りがあった場合。この場合、弊社の決定に1人が異議を申し立てれば、すべての投稿を復元します。
- 投稿者が異議申し立てを行う前に、レビュー時の誤判断が判明した場合
- 悪意があると弊社が特定するリンクを含んだ投稿を削除した後、そのリンクが有害ではなくなったことが分かった場合。この場合、弊社は投稿を復元することができます。これは、特にスパムについて当てはまります。

この指標は、コンテンツに対して措置を講じる際に生じた弊社の誤りを示すものと解釈されがちですが、上記の悪意のあるリンクの例のとおり、投稿の復元は必ずしも、誤りがあったことを意味するものではありません。

各四半期（例えば、1月1日から3月31日までの期間）にFacebookが復元したコンテンツの総数を報告します。この数値は、同一の四半期に措置が講じられたコンテンツや、異議が申し立てられたコンテンツと直接比較できるものではありません。例えば、復元されたコンテンツの中には過去の期間に異議申し立てがなされているものもあれば、異議申し立ての中には次の期間に復元されるものもあります。

1件のコンテンツは、投稿、写真、動画やコメントなど、複数の要素で構成される場合があります。個々のコンテンツを数える方法は複雑であり、時間とともに発展してきました。措置を講じたコンテンツの指標について詳しくはこちら。

●アカウント、ページ、グループおよびイベントに対する異議申し立ての測定方法

現時点でコミュニティ規定施行レポートには、弊社により復元されたアカウント、ページ、グループおよびイベントに関する指標は含まれていません。