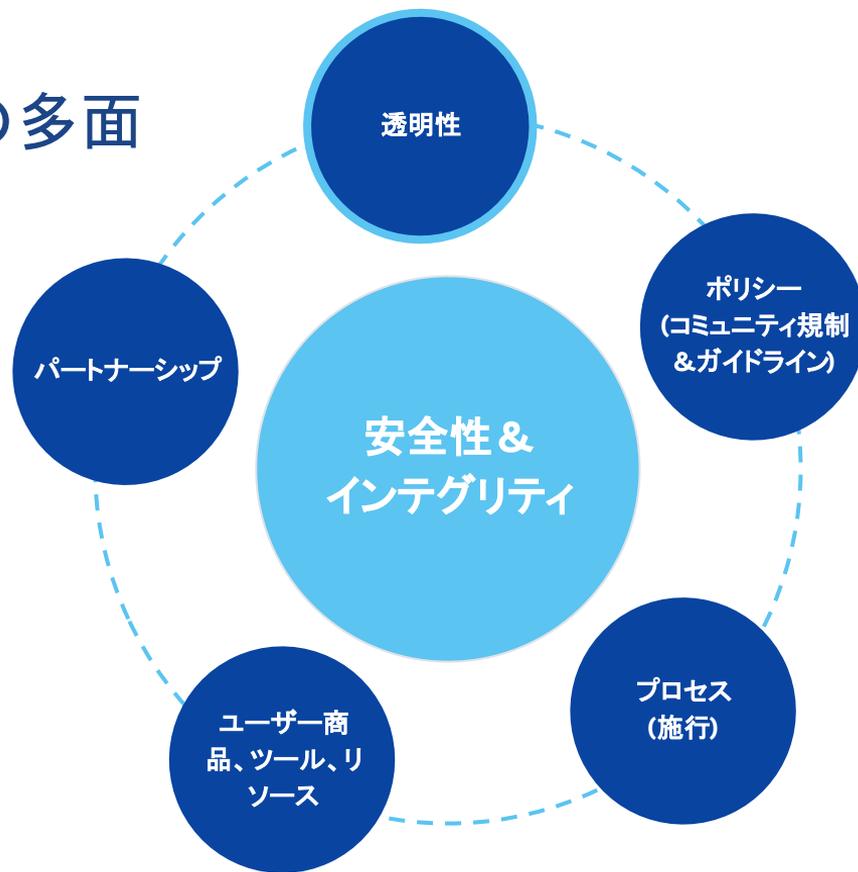


Metaのコンテンツ、安全性、透明 性へのアプローチ

安全性、インテグリティへの多面的アプローチ





40,000 人

グローバルで安全性、セキュリティに従事する人員

40+ チーム

社内でこの業務に関わるチーム

70+ 言語

マーケットスペシャリストチームがカバーする言語

US\$50億+

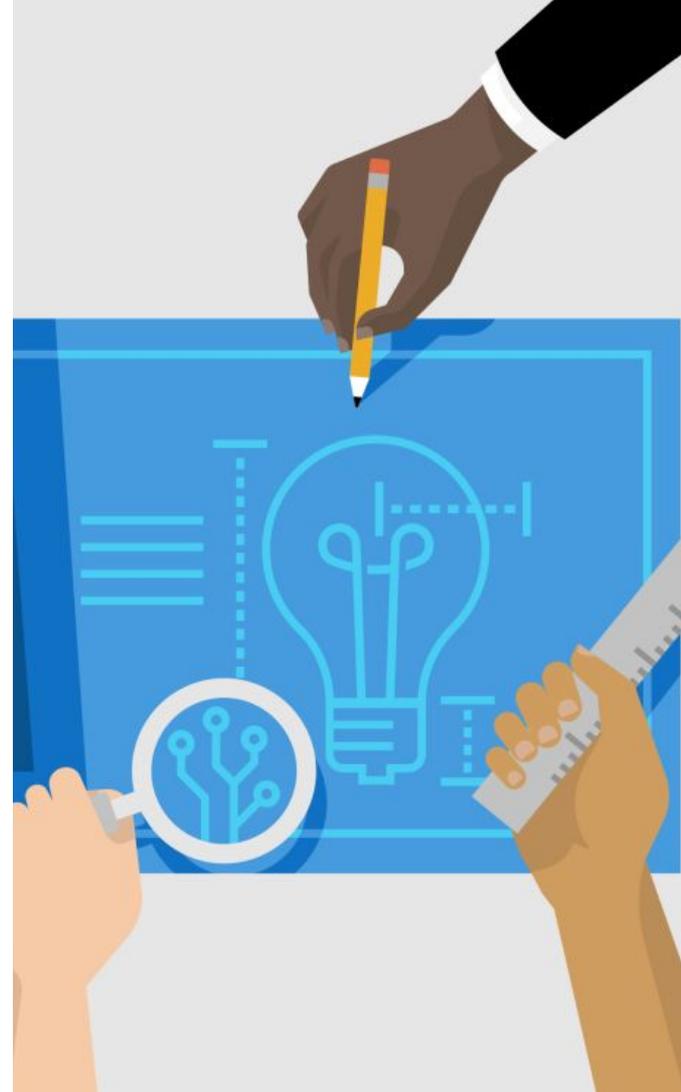
安全性、セキュリティへの投資額

ワールドクラスのAI技術

コンテンツの問題をプロアクティブに検知し対処する技術

透明性に関する考え方

- 透明性とは以下の目的の達成のために **意味をなす**ものであるべき：
 - 公共の利に資する重要かつ関心の高い問題につき、情報提供する
 - 社会に学習の機会を提供する(特に利用者、政策立案者、規制当局、メディア)
 - 利用者をエンパワーし信頼関係を構築する
 - リスク(安全性、セキュリティ、プライバシー、企業秘密など)とのバランスをとる
 - プラットフォーム、デジタルコミュニケーションのグローバルな性質を受け入れる
- データの透明性はプロセスよりも **結果**にフォーカスする - 特に有害コンテンツ表示頻度など
- 透明性は **説明責任の基盤** であり、以下によって促進されるべき：
 - 独立した第三者の評価を受け、グローバルなベストプラクティスに従い、プラットフォームのガバナンスや規制機能に更なる説明義務を課す
 - レポートや、ポリシー、プロセス、システムに関する情報を公開し精査を受ける



透明性センター

- ポリシー
- ポリシー施行
- 偽情報
- 選挙
- フィードのランキング
- セキュリティ
- 監督委員会
- データ

透明性センターリンク: <https://transparency.fb.com>

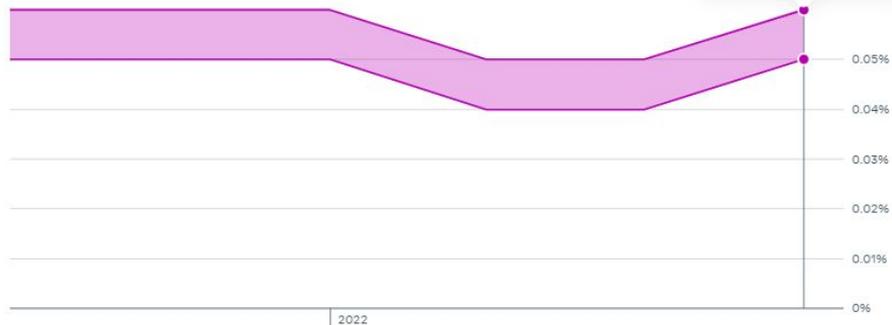


コミュニティ規定施行レポート

PREVALENCE

How prevalent were bullying and harassment violations?

いじめと嫌がらせに関する表示頻度について

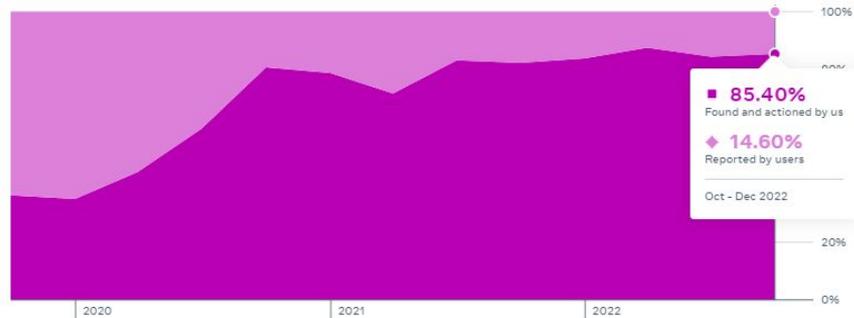


How we calculate it ⓘ Read about this data ⓘ

PROACTIVE RATE

Of the violating content we actioned for bullying and harassment, how much did we find and action before people reported it?

いじめと嫌がらせに関して利用者が報告する前に措置を採ったものの割合について



Found and actioned by us Reported by users

How we calculate it ⓘ Read about this data ⓘ

コミュニティ規定施行レポートリンク

transparency.fb.com/data/community-standards-enforcement/

コンテンツ配信ガイドライン

コミュニティ規定がどのようなコンテンツが Facebook から削除されるかを説明するのと同様に、コンテンツ配信ガイドラインでは、どのようなコンテンツが、問題あるコンテンツまたはクオリティの低いコンテンツとして Facebook 上の配信制限の対象となるのかについて説明しています。例えば：

- クリックベイト詐欺のリンク
- エンゲージメントベイト詐欺
- 低品質な動画
- コミュニティ規定違反と思われるコンテンツ
- ポリシー違反を繰り返す利用者のコンテンツ
- 違反ギリギリのコンテンツ

ガイドライン：

<https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-demote>

利用者の直接的なフィードバックへの対応

弊社は、Facebook で見たいもの、見たくないものについて、利用者からのフィードバックを常に歓迎しており、それに応じてフィードに変更を加えています。

- 迷惑広告
- クリックベイトのリンク
- 報告されたり非推奨コメント
- エンゲージメントの低いリンク
- 不必要なユーザーのリンク
- 低品質なコメント
- 低品質なイベント
- 低品質な動画
- オリジナルではないコンテンツ
- スпамだと推定された投稿
- 扇動的な健康関係の投稿

クリエイターによる高品質で正確なコンテンツへの投資の奨励

弊社では、利用者に長期間、興味深い新しいコンテンツを楽しんでもらいたいと考えており、こうしたコンテンツの作成を促すインセンティブを設けています。

- オリジナルコンテンツが少ないドメイン
- ファクトチェックされた偽情報

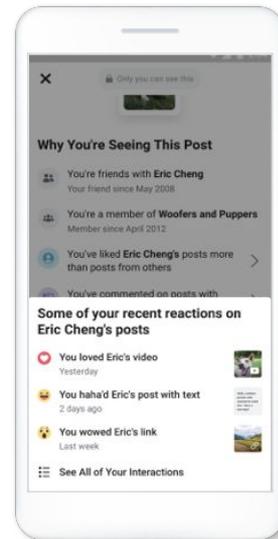
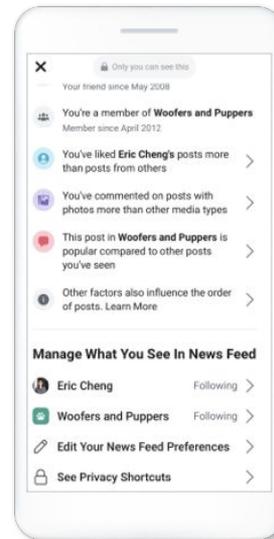
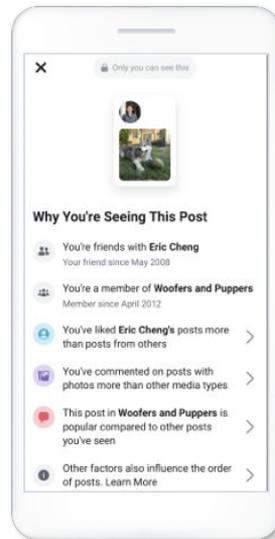
安全なコミュニティの構築

コンテンツの中には、その意図の有無にかかわらず、コミュニティにとって問題となるものがあります。弊社は、このようなコンテンツが、積極的に見ようとしなない人の目に届きにくいようにしています。

- コミュニティ規定のボーダーライン上のコンテンツ
- コミュニティ規定に違反している可能性が高いコンテンツ
- 弊社のポリシーに繰り返し違反している利用者が投稿したコンテンツ
- バイラリティが疑わしいことを示す投稿
- 自殺に関する不確かな報告

この投稿が表示される理由

- 投稿が利用者の **フィードに表示された理由** について、調べることができます。例えばどのようなファクターが寄与したか、利用者のどのようなアクティビティが関係していたかなどです
- コンテンツの順位付において一般的に影響を及ぼすことの大きいデータについて情報提供します
 - 誰が投稿したか
 - どのような投稿
 - どれほど人気のある投稿か
- 利用者自身の行動についての情報も表示されます
- そして設定スクリーンにも直接リンクされています
- この広告が表示される理由にも同様の機能



独立した第三者評価機関

- [Data Transparency Advisory Group \(2019\)](#)
- [EY Assessment of 2021/2022 Community Standards Enforcement Report \(2022\)](#)
- [Digital Trust and Safety Partnership \(DTSP\) \(予定\)](#)

EYによるコミュニティ規定施行レポートの評価結果

	Description	Controls Assessed
Governance	Procedures for the Community Standards Enforcement Report measurement processes	Meta has established a comprehensive risk and control framework by implementing governance controls over Community Standards Enforcement Report measurement and reporting processes to verify accuracy and completeness of metrics.
Data Collection	Collection of prevalence and enforcement data	To maintain data integrity across large scale distributed systems, Meta has implemented measures to verify data between our enforcement systems (systems that take action on content), measurement systems (systems that prepares the data for metric computations) and reporting systems (systems that ultimately publish the end metrics) are consistent.
Data Processing	Preparation and movement of data across our measurement and enforcement pipelines	
Data Aggregation	Aggregating of prevalence and content actioned data for metrics reporting	To minimize the risk of data aggregation errors, Meta has developed procedures where metrics are computed at least twice - one via our measurement systems and separately via independent validation scripts. The results are compared and deviations >0.1% are investigated and corrected before metric publication.
Data Disclosures and Reporting	Analyzing, monitoring and reporting the metric trends and material events that impact the report metrics	Meta has implemented systematic checks and procedures to identify material metric movements. Root cause analysis is conducted for all material movements and disclosures are included in the report to inform the reader of the causes and impact of such events.
Information Technology General Controls	Applications, interfaces, services and tools that record, store, process and report data related to content actioned data	Meta has implemented safeguards to verify access to internal integrity, measurement and reporting systems is restricted and protected from unwarranted access. Changes to such systems follow Management defined system change management processes.

利用条件 & ポリシー



Voice



真正性



安全性



プライバシー



尊厳

いじめと嫌がらせ

ポリシーの詳細 ユーザーエクスペリエンス データ

ポリシーの詳細

変更履歴

今日

現在のバージョン

2023/01/26

2022/09/29

2021/12/23

2021/10/13

過去のデータを表示

ポリシーの基本理念

いじめや嫌がらせは、あらゆる場所で、さまざまな形で起こります。例えば、脅迫したり個人情報を公開したりすること、脅迫的なメッセージを送信すること、悪意をもって一方的に連絡することもいじめや嫌がらせにあたります。Facebookはこのような行為を許容しません。利用者が安心して交流し、尊重されていると感じることができなくなるからです。

Facebookでは、いじめや嫌がらせについて公人・著名人と一般の個人を区別して規定しています。これは、Facebookにおいて議論や意見交換は認められる行為であり、報道で取り上げられたり社会的に注目を集めたりする人については、批判的な意見も含めた議論がされるためです。公人・著名人に関しては、攻撃が重大な場合、また本人が投稿やコメントに直接タグ付けされた場合は削除します。

一般の個人に対しては、保護の範囲が広くなります。例えば、他者の個人的な性的行為に関する主張など、他者に侮辱を与えたり誹謗したりすることを意図したコンテンツは削除されます。Facebookは、いじめや嫌がらせが未成年者に与える精神的影響について認識しています。そのため弊社のポリシーでは、13歳から18歳までの利用者に対する保護を強化しています。

ただし投稿の文脈や意図により認められる場合があります。いじめや嫌がらせへの非難や、意識喚起が目的であることが明確な場合は、投稿やシェアを認めています。事例によっては、本人がいじめや嫌がらせの標的になっているかどうかが確認するため、本人からの報告が必要になる場合があります。いじめや嫌がらせ防止のため、Facebookでは上記の行為やコンテンツの報告に加えて、Facebookが提供しているツールの使用を奨励しています。

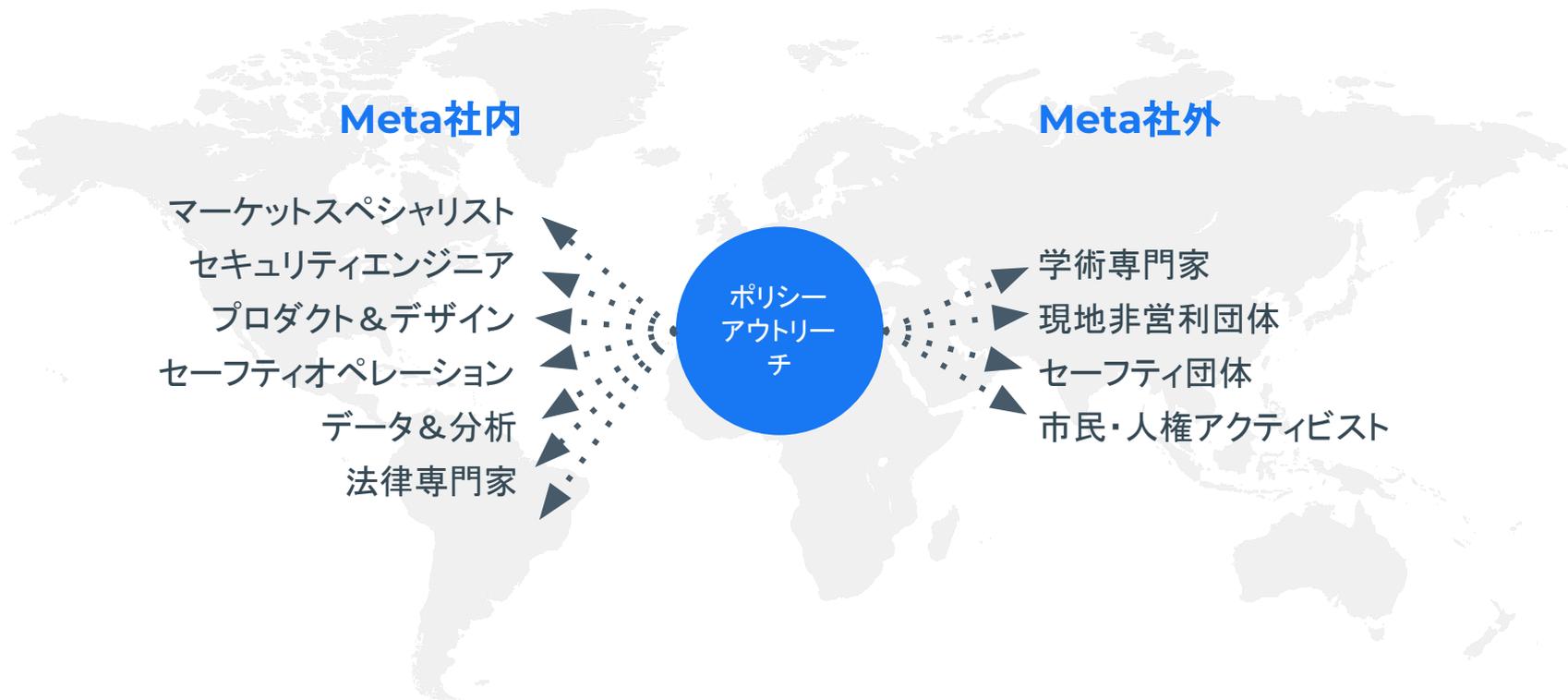
また、Facebookはいじめ防止ホームページを開発しています。このホームページには、いじめやその他の問題についてサポートを必要としている青少年、保護者、教育関係者のためのリソースがまとめられています。いじめをなくするための重要な対話の始め方など、ステップごとにアドバイスを提供しています。いじめと嫌がらせから利用者を守るためのFacebookの取り組みについて、詳しくはこちらをご覧ください。

注: 本ポリシーは、危険な人物および団体に関するポリシーに基づいて指定を受ける組織の個人や、1900年以前に亡くなった個人には適用されません。

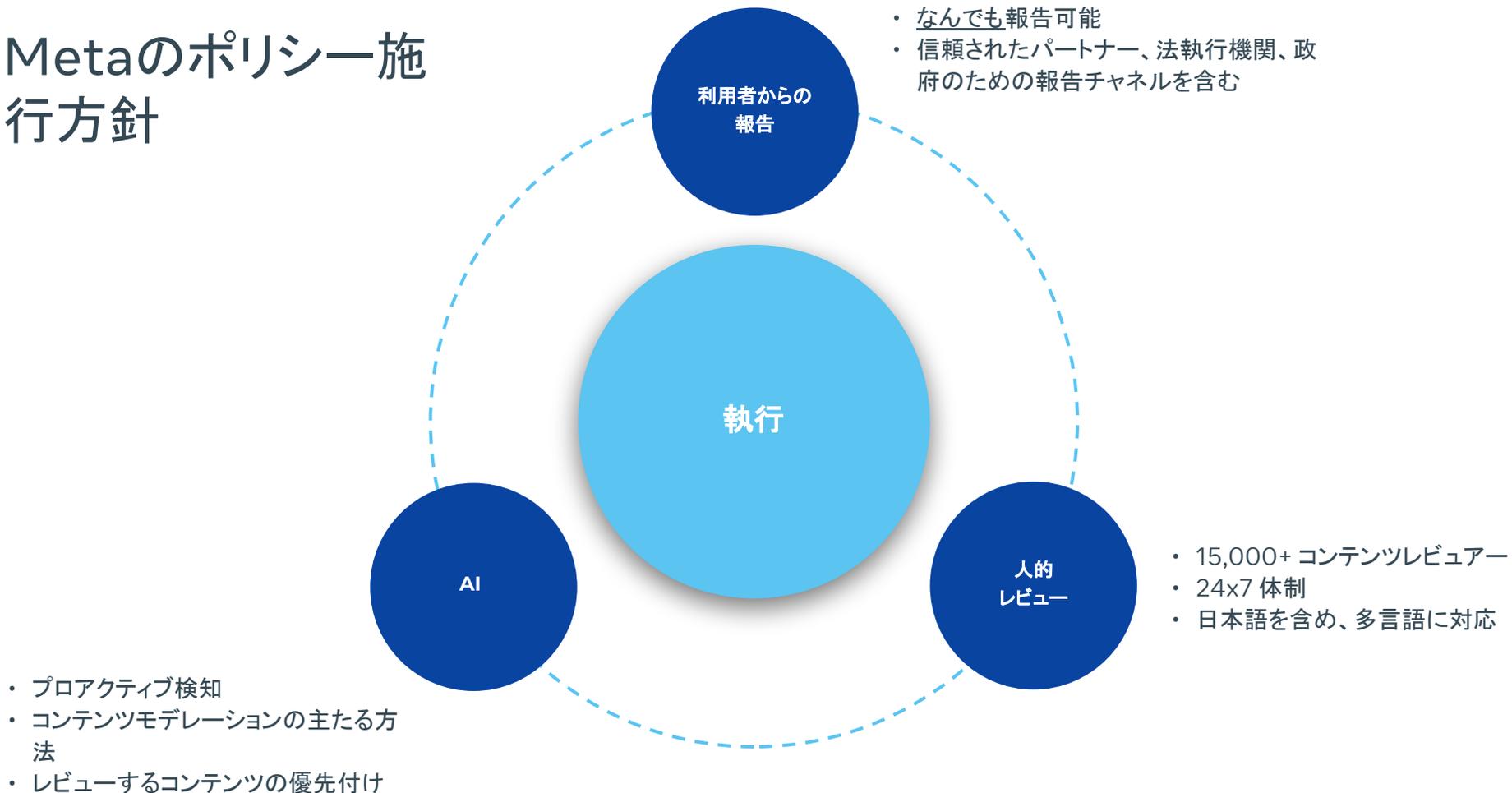
全ポリシーのリスト facebook.com/policies_center/

Metaが協業する人々

Metaは、公正で現実の問題に直結し、各国のニーズに対応したポリシーを作成するため、異なる見解をもつグループは専門家の意見を取り入れています



Metaのポリシー施行方針

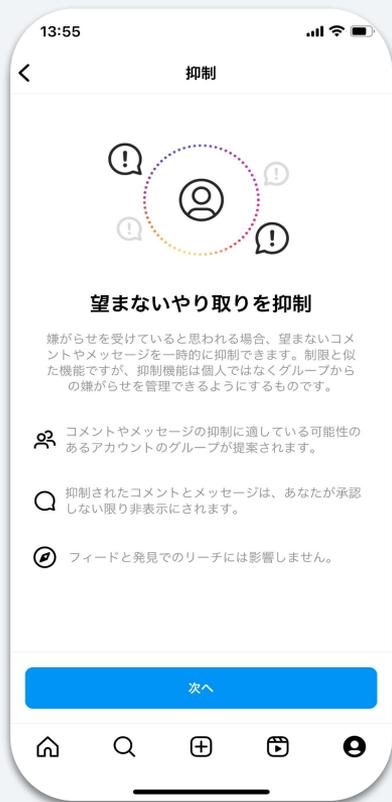


安全性と透明性に関するツール、リソースで利用者をエンパワー

Restrict (制限)



Limits (抑制)



Hidden Words (非表示ワード)



Hidden Words (非表示ワード)



利用者の報告ツール

- 虐待的コンテンツ
- プライバシー違反
- ユーザーの権利違反
- 違法なコンテンツ
- 名誉毀損報告



名誉毀損に関する報告フォーム

このフォームは、Facebookに投稿されている名誉毀損と思われるコンテンツを報告するためのものです。その他の報告は、このフォームでは受け付けていません。コンテンツがFacebookのコミュニティ規定違反となるケースは、名誉毀損以外にもさまざまなものがあります。コミュニティ規定違反を報告する方法について、詳しくはこちらをご覧ください。

法的権利を主張している国
 例えば、主な居住地などです。

報告の委任者

- 自分自身のために報告している。
- 企業・組織の代理で報告している(例:従業員)。
- 権利所有者に代わり報告する許可を得ている弁護士またはその他の代表者である
- 他の人のための報告である

プライバシーの侵害を報告

このお問い合わせ方法は、Facebookでの写真に関するプライバシー権利の侵害を報告するためのものです。そのほかについて報告する場合は、ヘルプセンターにお戻りください。

www.facebook.com/help

公開したくないコンテンツをシェアすると脅され、助けを必要としている場合は、このフォームに示されている手順に従ってください。

www.facebook.com/help/561743407175049

また、Facebookではすべての報告を確認しておりますが、あなたの報告に対処した場合でもお知らせすることはありません。

報告したいものは何ですか？

- 写真
- 動画
- その他

名誉毀損に関する報告フォーム

このフォームは、Facebookに投稿されている名誉毀損と思われるコンテンツを報告するためのものです。その他の報告は、このフォームでは受け付けていません。コンテンツがFacebookのコミュニティ規定違反となるケースは、名誉毀損以外にもさまざまなものがあります。コミュニティ規定違反を報告する方法について、詳しくはこちらをご覧ください。

法的権利を主張している国
 例えば、主な居住地などです。

著作権報告フォーム

著作権とは、映画、音楽、書籍、芸術など、原作者のオリジナル作品を保護する法的権利です。こちらのフォームは、著作権侵害の疑いを報告する場合にのみ使用してください。こちらのフォームを不正利用された場合、アカウントを削除させていただきます。

権利所有者との関係を説明してください。

- 私は権利所有者である
- 自分が所属する団体またはクライアントのための報告である
- 他の人のための報告である

PARTNERSHIPS



安全性についての情報

fb.com/safety

- ・ セーフティーセンター
- ・ いじめをなくそう ハブ

セーフティセンター

Metaは、提供するテクノロジーがすべての人にとってより良いものとなるよう、包括的なアプローチで臨んでいます。Metaのプラットフォームで利用者の安全を守るために行われている取り組みをご紹介します。



いじめをなくそう

この「いじめ防止ホームページ」は、いじめ防止の専門家の協力を得て制作されたものであり、いじめなどのトラブルで助けを必要としている教育関係者と家族のために、役立つリソースがまとめられています。ここでは、いじめられている子どもと大事な対話をはじめの方法、子どもがいじめをしているまたはされていることに悩む親や保護者へのアドバイス、いじめに関わっている生徒を持つ教育関係者向けの指針など、いじめの対策を詳しく説明しています。



いじめや嫌がらせに関しては、背景がきわめて重要です。関わっている子どもたちのことをよく知らず、その状況に隠された微妙なニュアンスが分からなければ、いじめと軽い冗談の区別はなかなかつきません。

Facebookでは、いじめや嫌がらせをきわめて深刻に受け止めており、利用者からの報告とテクノロジーを組み合わせ、そのようなコンテンツを発見・削除しています。この問題に適切に対処することがどれほど重要かを理解していればこそ、常に新しいツールに取り組み、ポリシーの見直しを行うとともに、検出技術にも継続的に投資することで、あらかじめこの問題に最大限の対処ができるようにしているのです。この対策の一環として、50を超える言語で日々寄せられるいじめや嫌がらせの報告を専門家チームが審査しているほか、AI技術を活用して880万件のコンテンツをチェックし、いじめに該当するコンテンツの

「保護者のためのInstagramガイド」や
安心安全のための機能、24のチュートリアル動画を公開




保護者のための
Instagramガイド

望まないやりとりは
制限しましょう



急激に増えるやりとりの抑制



良いガバナンスとは



技術やプラットフォームの違いを認識する



イノベーションの余地を残すためフレキシブルに



必要性に鑑み、釣り合いが取れている



現代のデジタル経済の活力源を維持する



意見、表現の自由、プライバシー、安全性など、人権を守る

