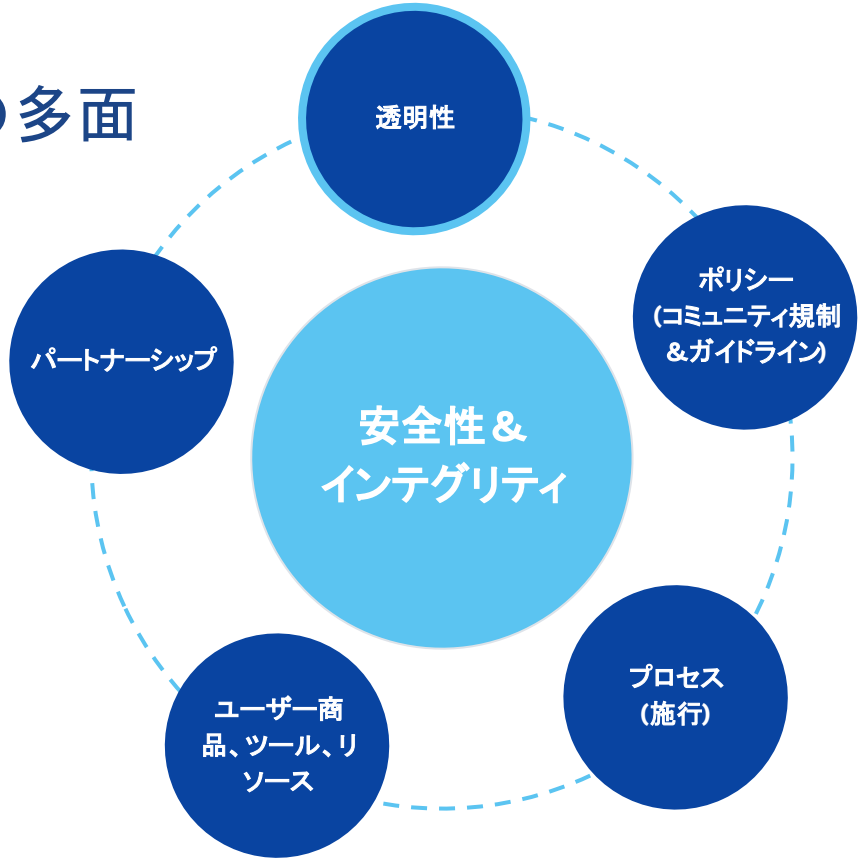
資料4

Metaのコンテンツ、安全性、透明性へのアプローチ

Meta

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+

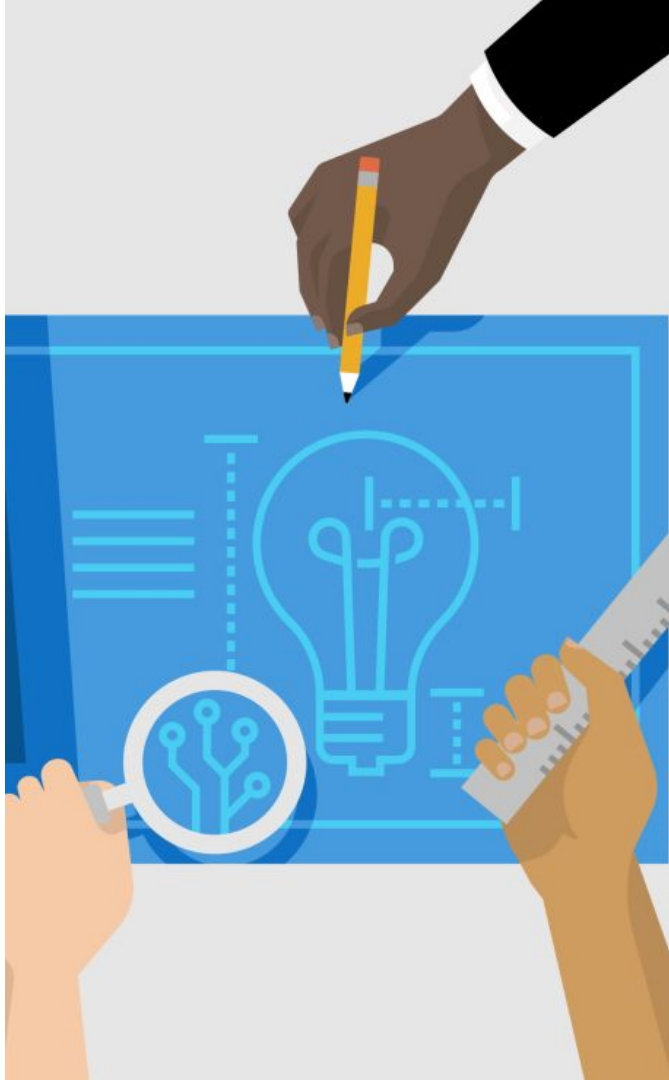
安全性、セキュリティへの投資額

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

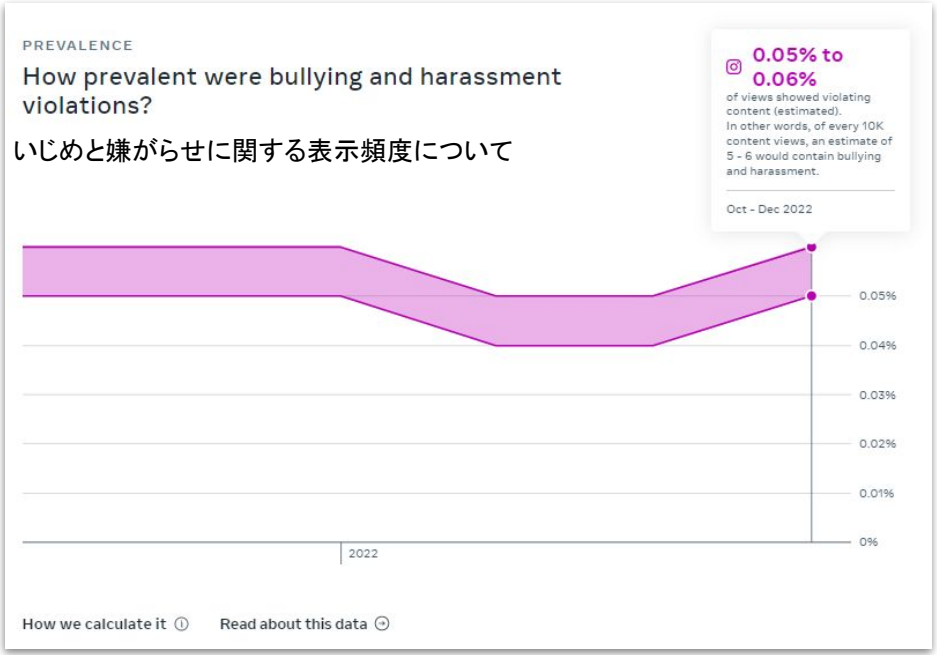


透明性センター

- 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



いじめと嫌がらせに関する表示頻度について

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

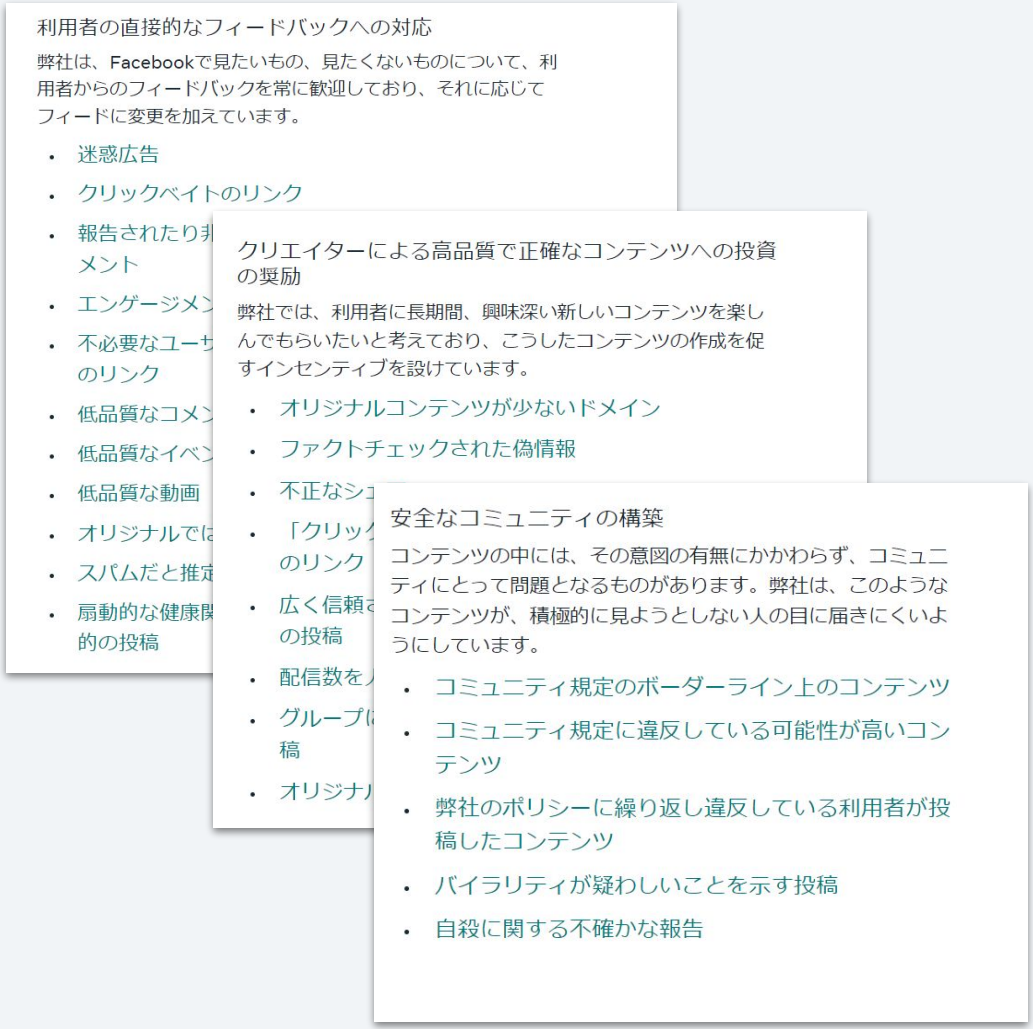
コミュニティ規定施行レポートリンク
transparency.fb.com/data/community-standards-enforcement/

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:

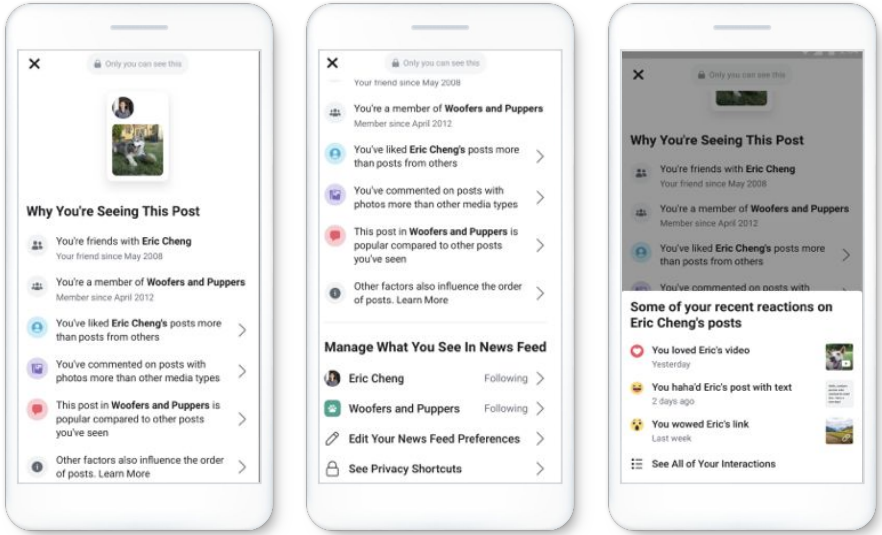
- クリックベイト詐欺のリンク
- エンゲージメントベイト詐欺
- 低品質な動画
- コミュニティ規定違反と思われるコンテンツ
- ポリシー違反を繰り返す利用者のコンテンツ
- 違反ギリギリのコンテンツ

ガイドライン:
https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote



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

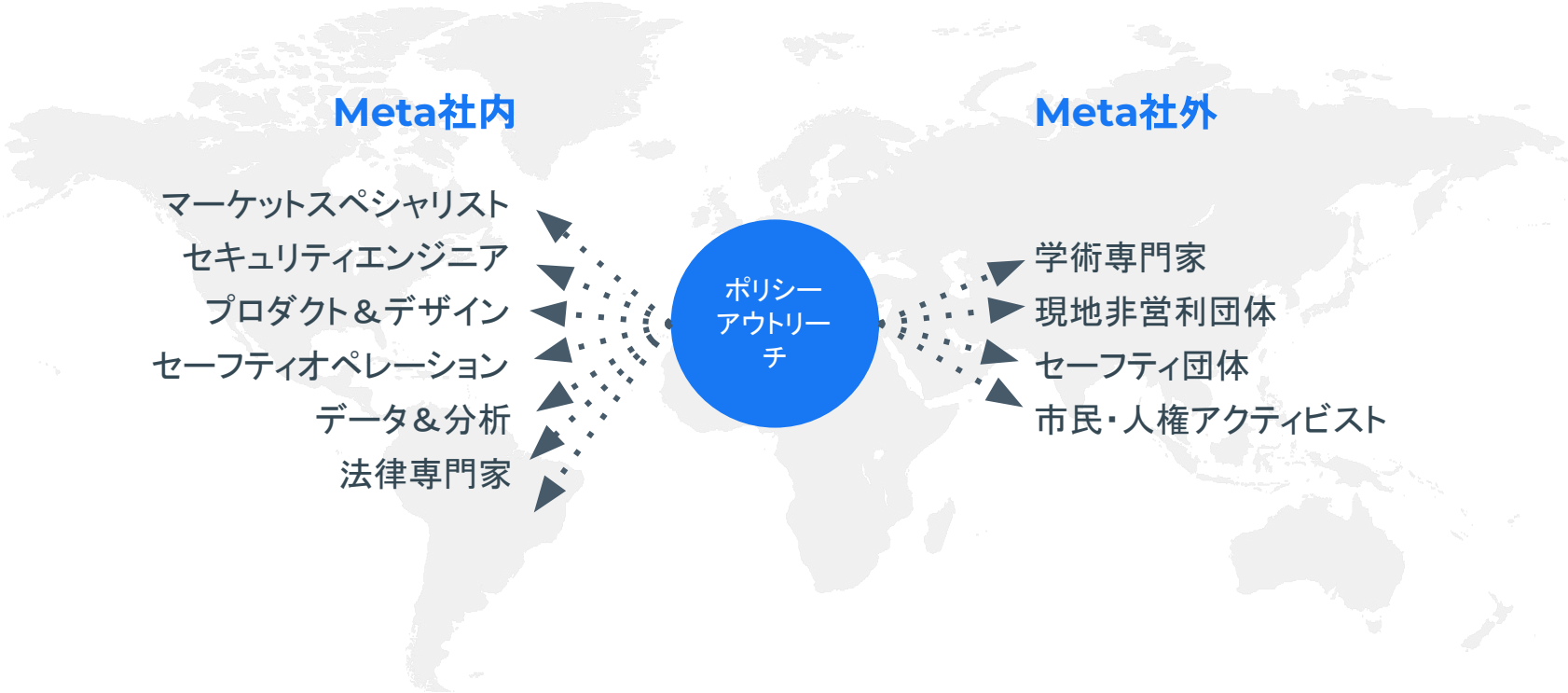
いじめと嫌がらせ

ポリシーの詳細　ユーザーエクスペリエンス　データ

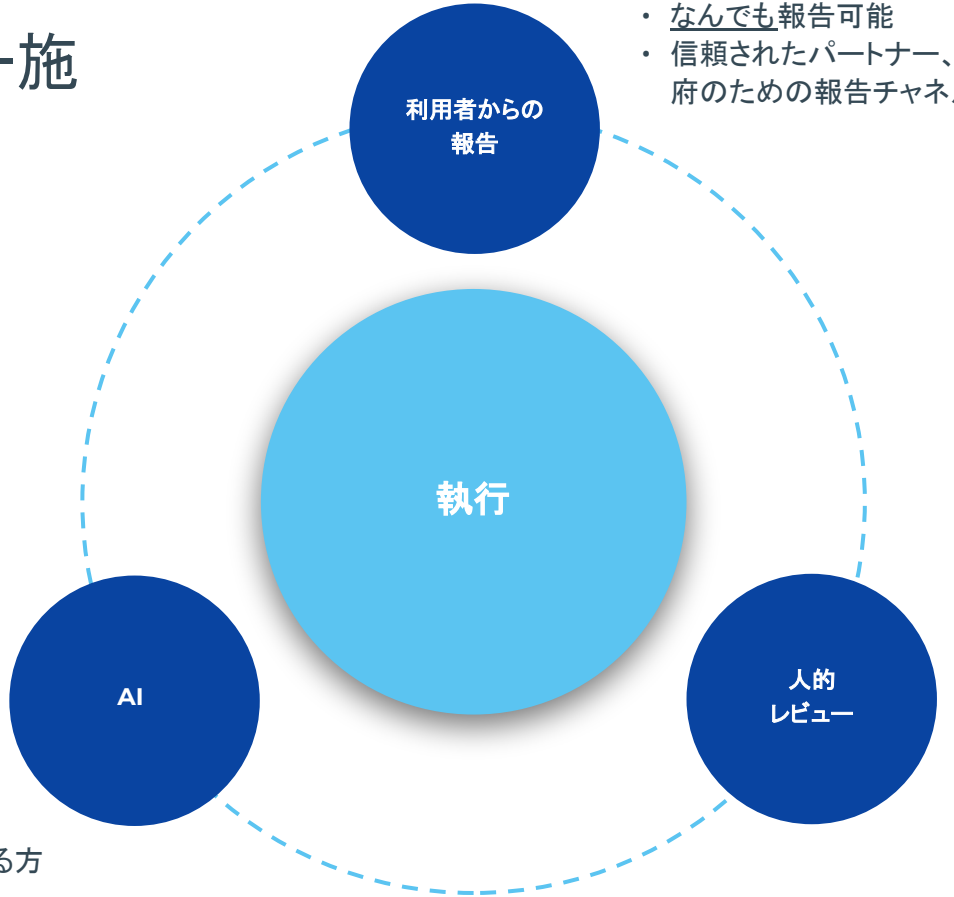
全ポリシーのリスト: facebook.com/policies_center/

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



- なんでも報告可能
- 信頼されたパートナー、法執行機関、政府のための報告チャネルを含む

- プロアクティブ検知
- コンテンツモデレーションの主たる方法
- レビューするコンテンツの優先付け

- 15,000+ コンテンツレビュアー
- 24x7 体制
- 日本語を含め、多言語に対応

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）

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

PARTNERSHIPS

ストップ
いじめ！ナビ
stopijime.jp

KIDSDOOR
NPO法人 キッズドア

3keys
～きっかけ・きづき・きぼう～

sodateage

NPOカタリバ

零

TECH
COALITION

unicef

Digital Trust
& Safety Partnership

NATIONAL CENTER FOR
MISSING &
EXPLOITED
CHILDREN

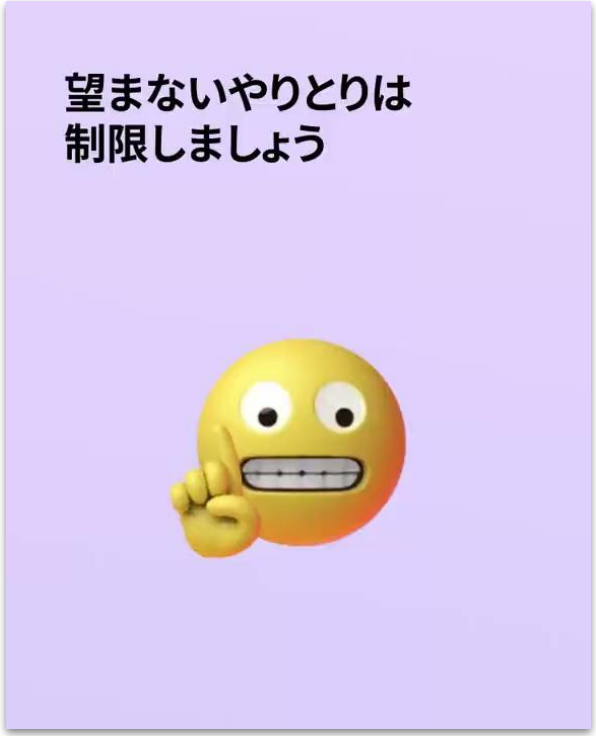
MEDIA
LITERACY
COUNCIL

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

「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



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


Meta