Metaによる Misinformation及び Disinformationへの対応について

13 April 2023

Meta

本日のアジェンダ

Misinformationへの対抗措置

敵対的脅威を及ぼそうとする個人・集団への対応
(Disinformation)

Misinformation and Disinformationに係る課題へのその他の方策

Misinformationへの対抗措置

Meta

ABCフレームワーク（Actor, Behavior, Content ）

行為者に係る問題

偽アカウント／濫用的アカウント

乗っ取り

なりすまし

虚偽表示

行為に係る問題

いじめやいやがらせ

現実性の高い暴力による脅迫

チャイルドグルーミング

金融詐欺

スパム

不正なエンゲージメント稼ぎ

組織的偽装行為

コンテンツに係る問題

Misinformation

ヌード・ポルノ

自殺自傷

暴力や過激な描写

ヘイトスピーチ

児童に対する性的搾取

削除

コミュニティ規定に違反するコンテンツやアカウント（有害な誤情報を含む。）を削除

抑制

質の低いコンテンツ（誤情報を含む）の配信を抑制

情報提供

利用者が何を読み、何を信頼し、何を共有すべきかを判断できるためにより多くのコンテキストを情報提供

削除

- 差し迫った物理的な危害を及ぼす危険性のある誤った情報
- 操作・誘導された映像／「ディープフェイク」について
- 暴力を助長する陰謀行為ネットワーク (QAnon)
- 投票妨害

+その他コミュニティ規定に違反するいっさいのもの

コンテンツ配信ガイドライン

コミュニティ規定でどのようなコンテンツがFacebookから削除されるかを説明しているのと同様に、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：

●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

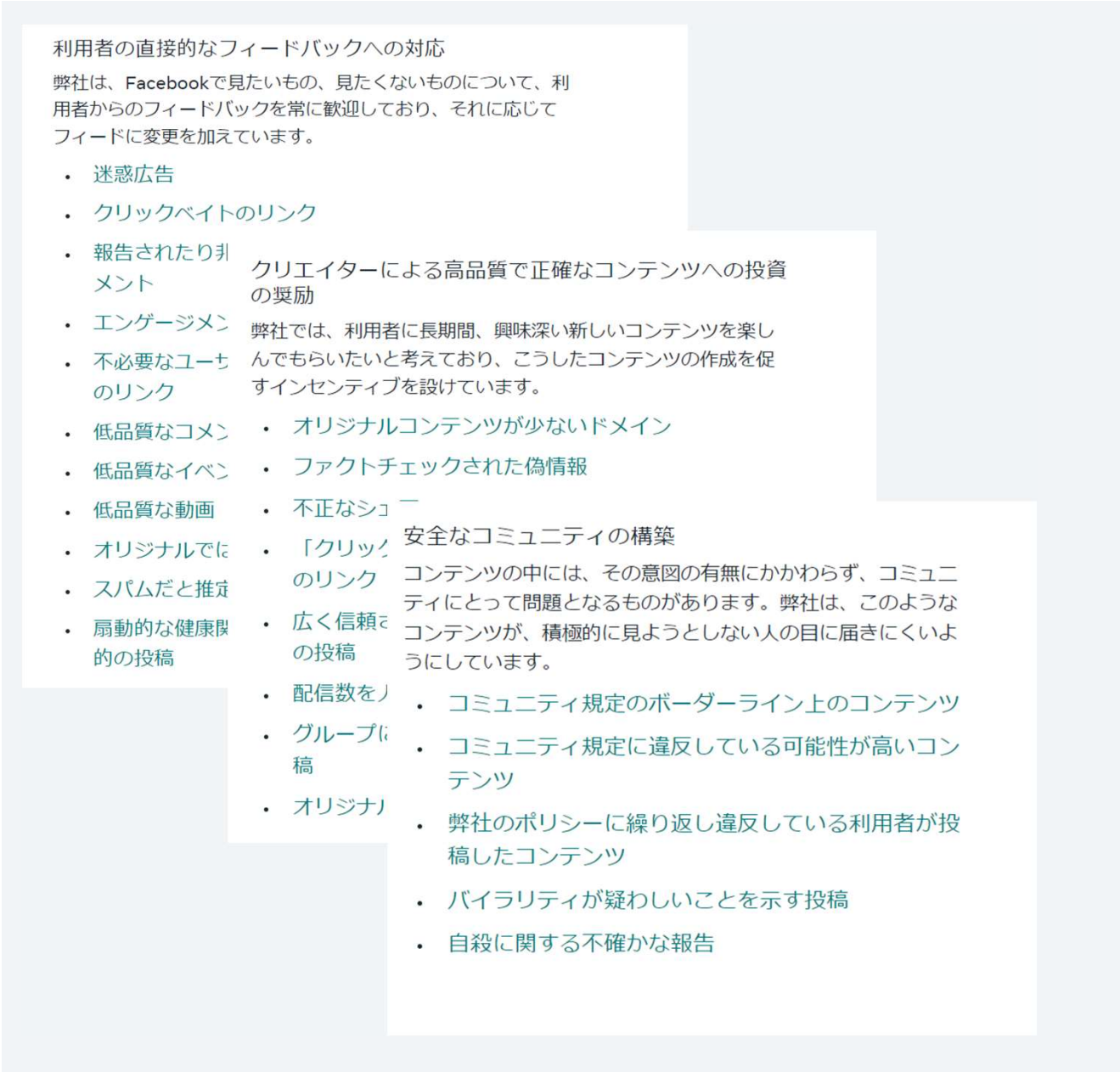
●低品質な動画

●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

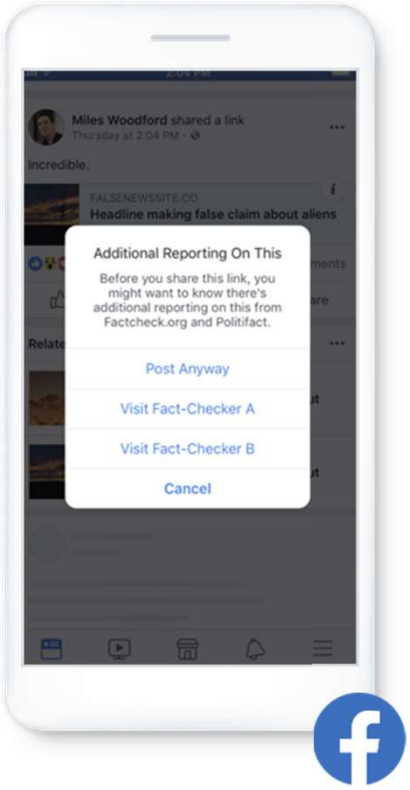
詳細:
https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote



情報提供

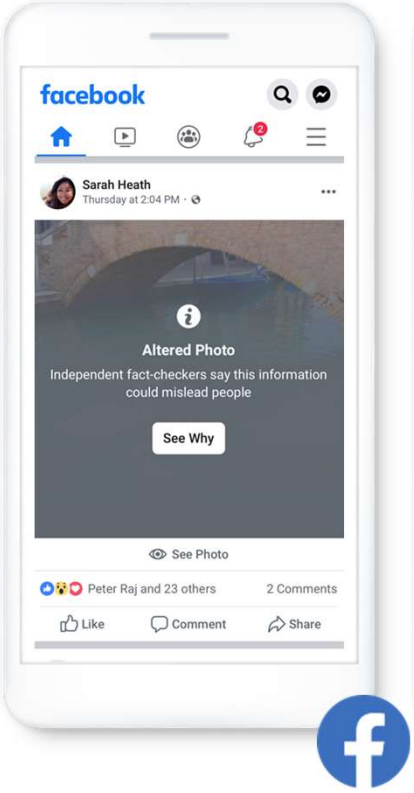
ラベル付けと通知

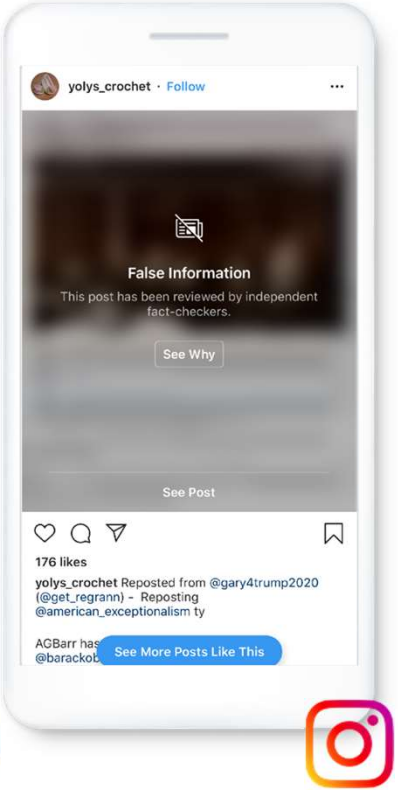
シェアする際の通知



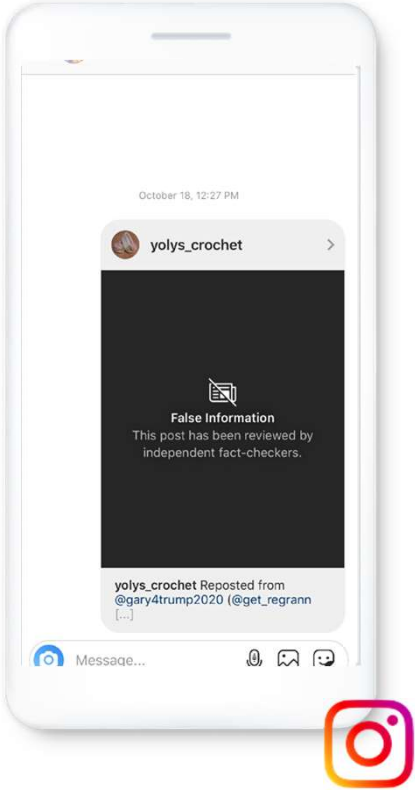
利用者への通知

警告・注意ラベル





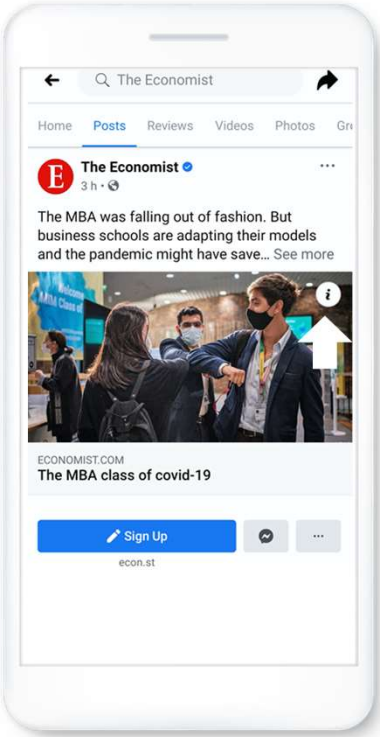
ファクトチェックの通知



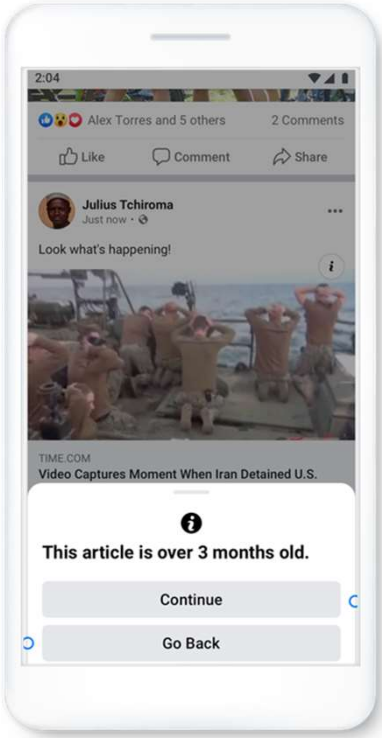
情報提供

コンテキストの提供と追加情報の提供

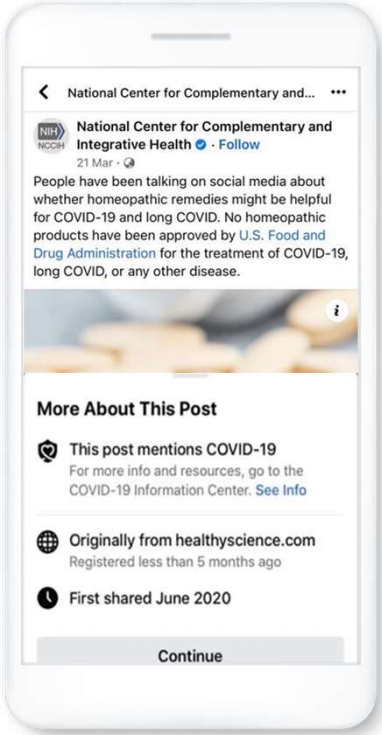
コンテキストボタン



”Old news”通知



新型コロナウィルス感染症



情報提供

リテラシー向上に向けて

www.wethinkdigital.fb.com

みんなのデジタル教室

日本での取り組み

日本では2020年12月に開始、NPO法人企業教育研究会と共同で日本に合わせたコンテンツを制作

全国の中学校、高校で2トピックの授業を提供し、**24,000名**以上が受講

86%の受講者がインターネットやアプリ、SNSの利用の仕方について考えが変わったと回答

「リソース」ページでは、安心安全のためのInstagramの機能の設定方法を紹介する動画なども提供

Learn.
Share.
Grow.

FactChecking Fundamentals

- Metaの支援を受けて、IFCNがアジア太平洋州のジャーナリストを対象として提供するもの。
- 日本語を含め、15か国語に対応。
- 無料。
- 本コース修了した受講生には、その成果を称える修了証を授与。

内容

(1)ファクトチェックの概要、
(2)検証、デバンキング（虚偽の暴露）について、
(3)健康・保健に関するmis/disinformation

https://www.poynter.org/shop/fact-checking/fact-checking-fundamentals-japanese/

敵対的脅威を及ぼそうとする個人・集団への対応 (Disinformation)

Meta

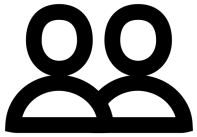
偽アカウントの摘発

正真性（Authenticity）

Metaのセキュリティシステムはバックグランドで毎秒数百万回の探知を実行。

偽アカウントの特定

AIにより、報告を受ける前に99.3%以上の偽アカウントを特定・削除



人の目による監視

安心・安全に係るスタッフを40,000人以上に三倍増。

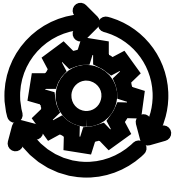
敵対的脅威となる勢力の探知と排除

探知

悪影響を及ぼす選挙関連の行為は人の目により警戒探知

対抗措置

疑わしい活動に対してセキュリティ対応チームが対応・違反アカウントの削除



適応化・順応化

世界規模で行われる事象から得られる知見を活かして継続的に改善

コスト引き上げ と 利得の引き下げ

- 該当勢力による作戦実施を困難にする
- 作戦の一挙殲滅
- 採った措置について情報公開
- プラットフォームを横断した協力関係

Misinformation及び Disinformationに係る課題へのその他の方策

Meta

業界自主行動規範

＋ さまざまなステークホルダーとの**継続的な対話とコラボレーション**の促進。

＋ 政策決定に必要な証拠を収集するために、**モニタリング、観察、施行**の期間の提供。

＋ **グローバルな業界標準**やベストプラクティスの枠組み、国際人権法に沿った業界標準を当該国に適切な設定。

＋ 言論・表現の自由、安全、プライバシーなどの価値観の間の保護のバランスを考慮し、**システムおよびリスクベースのアプローチ**を採用。

＋ 変化し続ける新たな脅威に対して、反復、改善、適応し、迅速に対応するための**より一層の柔軟性**の向上。

＋ 政府や業界が、misinformationに対して、強権的な規制を導入することなく対処が可能に。


Meta