資料7

コンテンツモデレーションにおける透明性と説明責任に関するサンタクララ原則

2022年5月12日

事　　務　　局

サンタクララ原則（2.0）について

○　サンタクララ原則は、コンテンツモデレーションにおける透明性及び説明責任に関する原則として、人権団体や学識経験者から構成されるグループによって2018年に策定、2021年に改訂された。
○　FacebookやGoogle、Twitterなど※12のプラットフォーム企業が支持している。
○　この原則は、コンテンツモデレーション携わる企業が、有意義な透明性及び説明責任を確保するための包括的な推奨事項を示すものであり、5つの「基本原則」と3つの「運用原則」、2つの「政府及びその他の国家機関のための原則」で構成される。
○　なお、この原則は規制の雛型を意図したものではなく、規制当局が政策形成においてモデレーションで考慮すべき事項を知るための「ガイド」であるとされている。

※　Apple, Facebook, GitHub, Google, Instagram, LinkedIn, Medium, Reddit, Snap, Tumblr, Twitter, YouTubeの12のプラットフォーム

基本原則（Foundational Principles）

① 人権及びデュープロセス（Human Rights and Due Process）
② 理解しやすいルール及びポリシー（Understandable Rules and policies）
③ 文化的能力（Cultural Competence）
④ コンテンツモデレーションへの国家関与（State Involvement in Content Moderation）
⑤ 確実性及び説明可能性（Integrity and Explainability）

運用原則（Operational Principles）

① 数値（Numbers）
② 告知（Notice）
③ 異議申立（Appeal）

政府及びその他の国家機関のための原則（Principles for Governments and other State Actors）

① 企業の透明性に対する障害の排除（Removing Barriers to Company Transparency）
② 政府自身の透明性向上（Promoting Government Transparency）

サンタクララ原則(2.0)について

① 人権及びデュープロセス(Human Rights and Due Process)

- 企業は、コンテンツモデレーションの全ての過程において人権及びデュープロセスを総合的に考慮し、それがどのように行われているか説明する情報を公表すべき。**(モデレーションにおける表現の自由等への配慮)**
- 企業は、自動化プロセス(人によるレビューの補完の有無は問わない)が品質及び精度に十分高い信頼性がある場合のみ、それを使用したコンテンツの識別や削除、アカウント停止の措置をすべき。**(自動化されたモデレーションの信頼性)**
- 企業はユーザに対し、コンテンツやアカウントが措置を受けた場合でもサポートを受けられるための明確でアクセス可能な方法を提供すべき。**(異議申立や問い合わせに対する受付態勢)**

② 理解しやすいルール及びポリシー(Understandable Rules and policies)

- 企業は、どのような場合にユーザーのコンテンツやアカウントが措置を受けるかについての明確かつ正確なポリシーを、簡単にアクセスできる場所で公開する必要がある。**(ポリシーとその適用の透明性)**

③ 文化的能力(Cultural Competence)

- モデレーションや異議申立てについて意思決定を行う者が、そのモデレーションを行う投稿についての言語、文化、政治的社会的背景を理解していることが求められる。**(言語、文化、政治、社会的理解)**
- 企業は、ポリシーやその適用において文化の多様性やそのサービスが利用される背景が考慮されるようにし、それらの考慮事項が全ての運用にかかる原則にどのように統合されたかについての情報を公表すべき。**(ポリシーへの反映の透明性)**
- 企業は、通報や異議申立てプロセス等が、ユーザが使用する言語で提供されること、また、コンテンツモデレーションの過程において、言語、出身国又は宗教を理由に差別されないことを確保すべき。**(ユーザへの言語的配慮、差別的取り扱いの禁止)**

④ コンテンツモデレーションへの国家関与(State Involvement in Content Moderation)

- 企業は、コンテンツモデレーションの過程への国家関与(現地法の遵守などを含む)に起因するユーザの権利侵害リスクを認識すべき。特に、コンテンツ削除又はアカウント停止を求める国家機関の要求が特に懸念される。**(国家関与への配慮)**

⑤ 確実性及び説明可能性(Integrity and Explainability)

- 企業は、コンテンツモデレーションシステム(精度及び差別禁止の追求、定期的評価の提出、報告や異議申立てメカニズムの公平な提供を含む)が確実かつ有効に機能していることを確認すべき。**(モデレーションの確実性)**
- 企業は、その意思決定の品質を積極的に監視し、高い信頼性レベルを確保すべきであり、システムの精度に関するデータを公表し、そのプロセス及びアルゴリズムシステムを定期的に外部監査に委ねるべき。**(モデレーションの説明可能性)**

サンタクララ原則（2.0）について

① 数値（Numbers）

- データは全て、定期的（四半期ごとが望ましい）に、オープンライセンスに従った機械可読式フォーマットで提示されるべき。
- 企業は、以下の情報を国又は地域別に（可能な場合はルールに違反したカテゴリー別に）公表すべき。
  - 措置を行った投稿及び停止されたアカウントの総数
  - 異議申立てが認められた数（又は割合）及び棄却された数（又は割合）
  - 自動検出によりフラグが立てられた投稿の異議申立てが認められた数（又は割合）及び棄却された数（又は割合）
  - 誤った措置であったことを認識した後、主体的（異議申立てによらず）に投稿やアカウントの復元を行った数
  - ヘイトスピーチのポリシーを実施した数やCOVID-19パンデミック等の危機的時期に発生したコンテンツの削除及び制限に関する数
- 国家機関が関与した意思決定には、以下の特殊要件が適用される。これは国別に公表すべき。
  - 国家機関がコンテンツ又はアカウントに対する措置を要求した数、要求主体の具体的名称
  - コンテンツにフラグを設定したのは裁判所命令もしくは他の国家機関か
  - 国家機関による措置の要求の数及び措置を行わなかった数
  - フラグの設定理由、措置を行った理由（ポリシー違反か現地法違反か）
- 企業は、フラグ設定プロセスの濫用防止のため、以下の情報の収集及び報告を検討すべき。
  - 一定期間を通じて立てられたフラグの総数
  - ボットに起因して立てられたフラグの総数
  - フラグを設定された投稿及びアカウントの総数（違反とされたルールやポリシー別、フラグの申告主体別）
- 企業は、コンテンツモデレーションに要求される自動機能の利用回数に加え、以下の情報を公表すべき。
  - 自動化プロセスがいつ、どのように使用されるのか
  - 自動化プロセスが使用されるコンテンツのカテゴリー、意思決定に自動化プロセスが使用されるか否かの主な基準
  - 自動化プロセスの信頼性、精度、成功率
  - 自動化プロセスの人による監視が含まれる程度
  - 自動化プロセスによりフラグ設定された場合、その異議申立てが認められた又は棄却された、カテゴリー別の数（又は割合）
  - 業界共通のハッシュ共有データベース等への参加状況及びそれらを通じてフラグ設定されたコンテンツに対する対応

② 告知（Notice）

- 企業は、ポリシー違反を理由にコンテンツ削除、アカウント停止等の措置が行われるユーザに対し、その理由について告知しなければならない。（対象となる投稿の明示、違反したとされるポリシーの条項、対象となった経緯、異議申し立ての機会等を含む）

③ 異議申立（Appeal）

- ユーザに対して利用可能な異議申立てプロセスの情報を提供するサポートチャネルへの十分なアクセス能力を提供すべき。

サンタクララ原則(2.0)について

① 企業の透明性に対する障害の排除(Removing Barriers to Company Transparency)

- 政府及びその他の国家機関は、企業が本原則を全面的に遵守するのを妨害する透明性の障害を排除すべき(また、かかる障害の導入を差し控えるべき)。
- 政府及びその他の国家機関は、国家機関から発生するコンテンツ又はアカウントへの措置要求を詳しく説明する情報の公開を企業が禁止されないようにすべき。ただし、かかる禁止に明確な法廷根拠があり、それが合法的目的の遂行に必要かつ比例的な手段である場合を除く。

② 政府自身の透明性向上(Promoting Government Transparency)

- 政府及びその他の国家機関は、コンテンツ又はアカウントに対する措置要求についての法的根拠別に分類したデータを含め、コンテンツモデレーションに関する決定への自らの関与を報告すべき。この報告は、全ての国家機関において説明されるべきであり、必要に応じて、半国家機関も含むべき。
- 政府及びその他の国家機関は、規制及び非規制措置を用いる方法を含め、本原則に従って企業による適切かつ有意義な透明性をどのように奨励すればよいか検討すべき。