参考資料１

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（仮訳）

はじめに

2018年、米国における大規模なコンテンツモデレーション（Content Moderation at Scale）カンファレンスと並行して、人権団体、人権活動家及び人権問題学識経験者で構成されるグループは、ますます攻撃的になるインターネットプラットフォームによるユーザ生成コンテンツのモデレーションに対し、有意義な透明性及び説明責任を最大限確保するための3つの原則を策定及び提起した。この原則は、カリフォルニア州サンタクララで開催された当該グループの第1回会議にちなんで名付けられ、影響を受ける発言者に有意義なデュープロセスを提供し、コンテンツガイドラインの施行が公正且つ公平であり、比例原則に則っており、ユーザの権利を尊重したものであることをより確実にするために、コンテンツモデレーションに携わる企業が取るべき最初のステップについての推奨事項を示している。これがサンタクララ原則の第1版（1.0）であった。

2018年からこれまでに、大企業12社—Apple、Facebook (Meta)、Google、Reddit、Twitter及びGithubなど—がサンタクララ原則を支持しており、透明性とプロセス上の保護施策を提供する企業の全体数は今も増え続けている。また、こうした大企業の多くが提供する透明性とプロセス上の保護施策のレベルも向上しつつある。

加えて、上記の企業が社会に果たす役割の重要性も高まりつつあり、これによって、説明責任を果たすために下す決定において十分なレベルの透明性を企業が提供する責任はますます大きくなっている。

こうしたことを理由に、複数の人権団体、人権活動家及び人権問題学識経験者で構成される広範なグループが結成され、2020年から2021年にかけて共同作業を行った結果、このサンタクララ原則の第2版（2.0）が策定された。この原則は、50を超える団体及び個人が参加した広範にわたる協議及び、綿密な起草及びレビュープロセスを経て策定された。このサンタクララ原則の第2版は、世界各国から経験と専門知識を引き出したことで、グローバルコミュニティの期待とニーズをより強く反映するものになっている。

このサンタクララ原則の第2版は、基本原則と運用原則で構成される。基本原則は、ビジネスモデル、創業年数及び規模に関係なく、コンテンツモデレーションに携わる時に全ての企業が考慮するべき包括的且つ分野横断的原則である。基本原則は、この原則をいかに施行するかについて、個々の原則及びガイダンスを明記する。運用原則では大企業や成熟企業向けに、コンテンツモデレーションの特定の段階及び側面におけるより詳細な期待事項を設定する。中小、新規及び資金力のない企業も、この運用原則を指針として、また、今後のコンプライアンスを報告するために利用したいと思っていただきたい。第1版で設定された最低基準と異なり、この第2版は、有意義な透明性及び説明責任の確保に厳密に必要なのはどの情報なのかについてより具体的に示している。

このサンタクララ原則の第2版では、「コンテンツ」とみなされるもの及び企業が講じる

べき「措置」に対し、透明性が要求される範囲が拡大されている。「コンテンツ」という用語は、広告を含め、有償か無償かを問わずユーザが1つのサービスに生成する全てのコンテンツを指す。「措置」及び「措置を受けた」という用語は、ルール及びポリシーの不遵守を理由に企業がユーザのコンテンツ又はアカウントに施行するあらゆる形態の施行措置を意味する。これには、コンテンツの削除、アルゴリズムの変更によるコンテンツのランクダウン及びアカウントの（一時的又は永久的）停止などが挙げられる（ただし、これに限らない）。

このサンタクララ原則の第2版は、企業が人権を尊重し、その説明責任を高める責任に従い、人権活動家の活動を支援するのに役立つように策定された。

参加団体

Access Now

ACLU Foundation of Northern California ACLU Foundation of Southern California ARTICLE 19

Brennan Center for Justice

Center for Democracy & Technology Electronic Frontier Foundation Global Partners Digital

InternetLab

National Coalition Against Censorship New America's Open Technology Institute Ranking Digital Rights

Red en Defensa de los Derechos Digitales WITNESS

基本原則

１．人権及びデュープロセス

原則：企業はコンテンツモデレーションプロセスの全段階で、人権とデュープロセスが統合的に考慮されるようにするべきであり、この2つが統合的に考慮された経緯を説明する情報を公表するべきである。企業は、プロセスの品質及び精度に十分高い信頼性がある自動化プロセス（人的レビューで補完するものか否かは問わない）のみを用いてコンテンツを識別又は削除したり、アカウントを停止したりするべきである。企業は、ユーザに対し、コンテンツ及びアカウントが措置を受けた場合でもサポートをうけるための明確でアクセス可能な方法を提供するべきである。

施行：以下を伝えるなどの方法で、コンテンツモデレーションプロセスの全段階で人権とデュープロセスが統合的に考慮されたことをユーザが確信を持てるようにするべきである。

- 企業はそのルール及びポリシーの策定においてどのように人権（特に、表現の自由に対する権利及び差別を受けない権利）を考慮したか。

- 企業はそのルール及びポリシーを施行するにあたってデュープロセスの重要性をどのように考慮したか、また、特に、このプロセスはどの程度確実であり公正に管理されているか。

- 企業がコンテンツモデレーションに自動化プロセスを使用する程度及び企業がかかる使用においてどのように人権を考慮したか。

２．理解しやすいルール及びポリシー

原則：企業は、ユーザのコンテンツ又はアカウントに措置が施行されるのはどのような時かに関する明確で厳密なルール及びポリシーを、簡単にアクセスできる主要なロケーションに公表するべきである。

施行：ユーザは以下について、簡単に理解できるべきである。

- 企業が禁止しており、削除されることになるコンテンツの類型並びに許容されるコンテンツ及び許容されないコンテンツの詳細なガイダンス及び事例。

- アルゴリズムによる表示の降格といった削除以外の措置を企業が講じるコンテンツの類型及び、コンテンツと措置の個々の類型に関する詳細なガイダンス及び事例。

- 企業がユーザのアカウントを一時的又は永久的に停止する状況。

３．文化的能力

原則：文化的能力としては、特に、モデレーション及び異議申立てについて意思決定を行う者が、モデレーションする投稿について、言語、文化及び政治社会的背景から理解していることが要求される。企業は、そのルール及びポリシー、並びにその適用において、文化の多様性及び、企業のプラットフォームやサービスが提供及び利用される背景が考慮されるようにし且つ、この考慮事項が全ての運用原則にどう組み込まれたかについて情報を公表するべきである。企業は、通報、告知及び異議申立てプロセスをユーザがサービスと情報をやりとりする言語で提供されること及び、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ユーザが差別されないことを確保するべきである。

施行：ユーザは、ユーザが使用する言語又は方言で、ルール及びポリシー、告知、異議申立て及び通報メカニズムにアクセスできるべきである。以下であることを、ユーザが確信できるようにするべきである。

- モデレーションに関する意思決定は、当該言語又は方言に精通した者によって行われる。
- モデレーションに関する意思決定は、当該の地域的又は文化的背景を十分認識した状況で行われる。
- 企業は、コンテンツモデレーターの言語及び地理的分布を実証する数字といった、サービスの提供先であるユーザに対する言語、地域及び文化的な適性を実証するデータを報告すること。

４．コンテンツモデレーションへの国家関与

原則：企業は、コンテンツモデレーションプロセスへの国家関与に特に起因して発生する、ユーザの権利侵害リスクを認識するべきである。これには、現地法の遵守又はその他の国益への貢献のいずれかを目的とする、企業のルール及びポリシーの策定及び施行に対する国家の関与を含む。コンテンツの削除又はアカウントの停止を求める国家機関（政府機関、規制当局、法執行機関及び裁判所を含む）の要求及び要請が特に懸念される。

施行：ユーザは、国家機関が、自身のコンテンツ又はアカウントへの措置を要求した場合やそれに関与した場合、そのことを知るべきである。ユーザは、企業が、その措置の施行は関連法によって要求されたもの考えているかについても知るべきである。透明性報告の一貫として、国が法の下にコンテンツの制限を要求していると現時点で報告している企業もあれば、公にも措置を受けたユーザにも国家関与を報告していない企業もあるが、自社のルール及びポリシーの施行に国家が関与している場合は、企業はユーザにこれを全て明確に報告するべきである。

ユーザは、特に、以下にアクセスできるべきである。

- 現地法の要件を反映するためのルール又はポリシーの詳細（グローバルに適用されるか、特定の法域で適用されるかは問わない）。
- コンテンツやアカウントへのフラグ設定や、企業が講じるその他の措置に関する公式または非公式の企業と国家機関との連携または合意に関する詳細
- 企業のルールあるいはポリシー、又は現地法のいずれかに基づいて、国会機関によりフラグ設定されたコンテンツ又はアカウントが評価されるプロセスの詳細。
- 投稿及びアカウントに対する措置の施行を国が要求する際の詳細。

５．確実性及び説明可能性

原則：企業は、自動化及び非自動化コンポーネントの両方を含む、コンテンツモデレーションシステムが確実且つ有効に機能していることを確認するべきである。これには、検出方法における精度及び差別禁止の追求、定期的評価の提出及び、告知及び異議申立てメカニズムの公平な提供などが挙げられる。企業は、その意思決定の品質を積極的に監視して、高い信頼性レベルを確保するべきであり、また、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である。企業は、措置の要求が真正であり、ボット又は組織的攻撃に起因するものにならないように努力するべきである。

自動化システムには固有の懸念が多数あり、企業は、システムを信頼できる場合に限定して、且つ、透明かつ説明可能な方法でこれを活用するべきである。

施行：コンテンツに対する決定が慎重に且つ人権を尊重して下されることをユーザが確信できるようにするべきである。ユーザは、コンテンツモデレーションの決定はどの時点で下されたか又は、どのような時に自動化ツールに支援されたかを知るべきであり、コンテンツ関連の自動化プロセスに用いられる意思決定ロジックを詳細に理解するべきである。企業は、企業によるユーザ管理を実現するどのような管理機能にユーザはアクセスしているのか、コンテンツはアルゴリズムシステムを利用してどのように監督されるか及び、この管理機能がユーザのオンライン経験にどのような影響をもたらすかを明確に説明するべきである。

運用原則

１．数値

数値に関する原則では、自身の発言に対する決定を理解しようとするユーザ及び社会全体の両方にとっての、コンテンツモデレーションの透明性の重要性が考慮されている。企業は、ユーザ及び研究者がシステムを理解及び信頼できるように、企業のルール及びポリシーの違反を理由にユーザのコンテンツ及びアカウントに企業が講じることができる全ての措置に関する情報を報告するべきである。

企業は、措置を受けたコンテンツ及びアカウントの数に関する情報を、以下の観点に即して国や地域ごとに、また、可能な場合は、ルールに違反したカテゴリー別に公表するべきである。

- 措置を受けたコンテンツ及び停止されたアカウントの総数。
- コンテンツ措置又はアカウント停止の決定に対する異議申立ての数。
- 一連のコンテンツ又はアカウントの再開に至った成功した異議申立ての数（又は割合）及び、棄却された異議申立ての数（又は割合）。
- 自動検出機能によりフラグが立てられたコンテンツの、成功した異議申立て及び棄却された異議申立ての数（又は割合）。
- 措置の実行又は停止が誤りであったと認められた末、異議申立てを経ずに、企業が事前対応的に再開した投稿又はアカウントの数。
- 明らかな場合は、ヘイトスピーチのポリシーが施行された標的集団又は特徴別の件数（ただし、企業はこの目的のために標的集団に関するデータを収集してはならない）。
- COVID-19のパンデミック期間及び武力紛争期間等の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連する数。

国家機関が関与した意思決定には、報告に関する以下の特殊要件が適用される。この要件は国別に示されるべきである。

- 国家機関がコンテンツ又はアカウントに対する措置の施行を要求又は要請した数。
- 個々の要請に対する、国家機関の具体的名称。
- コンテンツにフラグを設定したのは裁判所命令／判事であるか或いはその他の国家機関であるか。
- 国家機関による措置施行の要求又は要請の件数及び、結果的に施行に至らなかった要求又は要請の数。

- 個々のフラグの設定理由は、企業のルール及びポリシー違反の申立てであったか（その場合は、どのルール又はポリシーであったか）、現地法違反の申立てであったか（その場合は、現地法のどの規定であったか）或いは、その両方であったか。

- コンテンツに対する措置の施行理由は、企業のルール及びポリシーの違反であったか現地法の違反であったか。

フラグ設定プロセスの今後の濫用は特に懸念されるため、企業は、かかる濫用の頻度をユーザ及び研究者が評価できるようなデータ報告及び、濫用防止に向けて企業が講じる措置を検討するべきである。特殊な地域的背景を踏まえて濫用関連の傾向を特定するのに役立つ具体的な基準及び／又は、定性的報告を考案してもよい。企業は、以下の情報の収集及び報告を検討するべきである。可能な場合は、国又は地域別に検討するべきである。

- 一定期間を通じて立てられたフラグの総数

- ボットに端を発するフラグの総数

- フラグを設定された投稿及びアカウントの総数

  - 違反を申告されたルール及びポリシー別

  - フラグの発生源（国家機関、信頼されるフラグ管理者、ユーザ、自動検出など）別

自動化プロセスがコンテンツモデレーションに果たす役割が大きくなったことにより、企業のプロセス及びシステムを包括的に理解するには、自動化意思決定ツールの利用における透明性が不可欠である。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、企業は以下に関する情報も公表するべきである。

- コンテンツに措置を実行する際に、自動化プロセスはいつ、どのように（単独又は人の監視を伴って）使用されるか。

- 自動化プロセスが使用されるコンテンツのカテゴリー及び類型。

- 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。これには、経時的変化及び、言語及びコンテンツカテゴリー間の差異などを含む。

- コンテンツモデレーションの自動的決定の人的レビューをユーザが要求できる可能性を含む、自動化プロセスに人による監視が含まれる程度。

- コンテンツ又はアカウントが自動検出機能によりフラグ設定された場合の、成功した異議申立て又は棄却された異議申立ての、コンテンツのフォーマット及び違反カテゴリー別の数（又は割合）。

- 業界共通のハッシュ共有データベース又はその他のイニシアティブへの参加及び、かかるイニシアティブを通じてフラグ設定されたコンテンツに対する企業の対応方法。

データは全て、定期報告の中で、望ましくは四半期ごとに、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である。

２．告知

企業は、サービスのルール及びポリシー違反を理由にコンテンツが削除される、アカウントが停止される又は、その他の何らかの措置を講じる際には、個々のユーザに対し、削除、停止又は措置の理由について告知を提供しなければならない。企業はそのルール及びポリシーの中で、このルールの例外、例えば、コンテンツがスパム、フィッシング又はマルウェアに相当する時の対応を明確に設定するべきである。

投稿に措置が施行された理由についてユーザに告知を提示する際は、企業は告知に以下などを記載するようにするべきである。

- URL、コンテンツの抜粋及び／又は、措置を受けたコンテンツを特定できるだけのその他の十分な情報。

- コンテンツの違反が判明したガイドラインの具体的な条項。

- コンテンツが検出及び削除された経緯（他のユーザ又は信頼できるフラグ管理者による通報、自動検出、外部の法的又はその他の申し立て）

- フラグ設定又は措置の施行命令における国家機関の関与に関する具体的な情報。国家機関によりフラグが設定されたコンテンツはそのようなものとして特定されるべきであり、法律で禁止される場合を除き、国家機関を具体的に特定するべきである。コンテンツが企業のルール又はポリシーではなく、現地法に違反すると申し立てられる場合は、現地法の該当する規定をユーザに伝えるべきである。

適切な告知に対する他の基準には以下などが挙げられる。

- 告知は時宜にかなったものであり且つ、期限又は関連する手続き上の要件を含め、ユーザが決定に異議を申し立てられるプロセスの説明を記載するべきである。

- 告知は、ユーザのアカウントが停止又は終了される場合でもアクセス可能な耐久性のある形式で提供されるべきである。

- コンテンツにフラグを立てるユーザには、ユーザが報告したコンテンツのログ及び、モデレーションプロセスの結果を提示するべきである。

- 告知は、最初の投稿に使用された言語又は、ユーザが選択するユーザインターフェ

ース言語で行うべきである。

- 告知では、ユーザに対し、利用可能なユーザサポートチャネルに関する情報及びこれにアクセスする方法を提示するべきである。

- 告知は、必要に応じて、グループ管理者及びフラグ管理者といった他の関係者にも提示されるべきである。これには、削除されたコンテンツの当初の場所への告知の投稿を含む。

３．異議申立て

異議申立ての原則は、企業の説明義務、レビュー及びユーザに提供される異議申立てプロセスを網羅する。ユーザには、措置の施行決定及び措置の最初の施行決定が下された際に利用可能な異議申立てプロセスについて情報が提供され、サポートチャネルへの十分なアクセス能力が提供されるべきである。企業は、コンテンツの削除、フラグが立てられたコンテンツの存続、アカウントの停止又は、表現の自由に対する権利を含む、ユーザの人権に影響を及ぼすその他の種類の措置の施行決定に適時に異議申立てを行う有意な機会を提供するべきである。企業は、比例原則に従って、コンテンツの削除及びアカウント停止といった最も厳しい制限に対する異議申立ての機会の提示を優先するべきである。

企業は異議申立てに以下を組み込むようにするべきである。

- 異議申立てを利用する個人に提供されるタイムライン及び進捗状況の追跡方法の詳細を記載する、明瞭且つユーザが簡単にアクセスできるプロセス。

- 当初の決定に関与していない個人又は複数個人で構成されるパネルによる人的レビュー。

- このレビューに参加する個人又はパネルは、異議申立てに関連するコンテンツの言語及び文化的背景に精通していること。

- このレビューで検討されることになる、異議申立てを裏付ける追加情報をユーザが提出する機会。

- レビュー結果の通知及び、ユーザがその決定を理解できるだけの十分な理由の記述。

長期的には、独立したレビュープロセスも、ユーザが救済を求めることを可能にするための重要な要素になる可能性がある。かかるプロセスがある場合は、企業はこのプロセスへのアクセスについてユーザに情報を提供するべきである。企業は、独立したレビュープロセスを統制する又はこれに影響を与える程度までサンタクララ原則を盛り込むようにするとともに、透明性の定期的報告、異議申立ての現状についてのユーザに対する明確な情報及び、決定の論理的根拠を確実に提供するべきである。

特定の状況、例えば、影響を受けるユーザが削除制度の濫用の標的になり得る場合又は、影響を受けるコンテンツが選挙期間中の政治的コンテンツのような一刻を争うものである場合は、企業は、異議申立てプロセスを迅速に処理すべきか否かを検討するべきである。異議申立てプロセスを迅速に処理する場合は、企業は、この処理が実施される状況及び、ユーザが異議申し立ての迅速な処理を要求できる可能性について、明確なルール及びポリシーを提示するべきである。

政府及び他の国家機関のための原則

政府は言うまでもなく、様々な国際法律文書、例えば、世界人権宣言の第19条に基づいて、全ての個人の表現の自由を尊重する義務を有する。ゆえに、国家機関は、反対派、反政府主義者、社会運動又は何人を検閲する意図で企業のコンテンツモデレーションシステムを悪用したり操作したりしてはならない。

透明性に関して言うと、企業による透明性は、コンテンツモデレーションプロセスにおける信用及び信頼を確保する上で極めて重要な要素である。ただし、政府は、透明性の妨げになる自らの役割を認識した上で、これを最小限に留めるとともに、コンテンツの削除又は制限に対する政府要求についても透明性を提供しなければならない。

１．企業の透明性を阻む障害の排除

政府及びその他の国家機関は、企業が上記の原則を全面的に遵守することを妨げる透明性への障害を排除するべきである（また、かかる障害の導入を差し控えるべきである）。

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の削除又は措置施行の要請又は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るべきである。ただし、かかる禁止に明確な法的根拠が存在し、それが合法的目的の遂行の必要且つ比例的な手段である場合を除く。

２．政府の透明性の促進

政府及びその他の国家機関は、コンテンツに対する措置施行又はアカウントの停止を求める要求又は要請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関与を自ら報告するべきである。この報告では、できれば統合報告書において、全ての国家機関について説明するべきであり、必要に応じて、半国家機関も含まれるべきである。

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、上記の原則に従って企業による適切且つ有意義な透明性をどのように奨励すればよいかを検討するべきである。

謝辞

コメントの提出、グループ協議への参加及び予備作業に対するレビュー及びコメントを賜った団体及び個人の方々全てに感謝の意を表する。コメントは、7amleh、Association for Progressive Communications、Centre for Internet & Society、Facebook/Meta、Fundación Acceso、GitHub、Institute for Research on Internet and Society、InternetLab、Laboratório de Políticas Públicas e Internet (LAPIN)、Lawyers Hub、Montreal AI Ethics Institute、PEN America、Point of View、Public Knowledge、Taiwan Association for Human Rights、The Dialogue、Usuarios Digitales等の諸団体からいただいた。協議を調整及び主催された、また、他の方法でご協力いただいた方々及び団体を以下に列挙するが、これだけに限らない。ALT Advisory、Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE)、UNESCO、Irina Raicu、Eduardo Celeste、Derechos Digitales、Robert Gorw、Ivar A.M. Hartmann、Amélie Heldt、Tomiwa Ilori、Julian Jaursch、Clara Iglesias Keller、Paddy Leerssen、Martin J. Riedl、Christian Strippel、Daphne Keller。

サンタクララ原則2.0は、Swedish Postcode Foundationからも協力をいただいた。

最後に、サンタクララ原則1.0の作成及び支援にご参加いただいた以下の諸団体及び方々に感謝を申し上げたい。ACLU Foundation of Northern California、Center for Democracy & Technology, Electronic Frontier Foundation、New America's Open Technology Institute、Irina Raicu、Nicolas Suzor、Sarah Myers West及び、Sarah T. Roberts。また、大規模なコンテンツモデレーションと削除（Content Moderation & Removal at Scale）カンファレンスを主催していただいたサンタクララ大学のハイテク法研究所及び、この1.0を生み出したワークショップの開催にご協力いただいたEric Goldmanにも謝意を表したい。このワークショップは、ペンシルベニア大学のInternet Policy Observatory atのご支援もあって実現された。Suzor氏は オーストラリア研究会議のDECRA Fellowship (プロジェクト番号DE160101542)を取得されている。

（https://santaclaraprinciples.org/より総務省仮訳として作成（2022年5月12日））