

# Google のアプローチ 生成 AI におけるコンテンツの安全性

Chenie Yoon

Google Asia Pacific コンテンツレギュレーションリード  
2024 年 3 月

# アジェンダ

- Google のアプローチ
- 合成メディア検出と来歴情報の技術

# Google のアプローチ

「私たちは、大胆に物事を進めていきますが、その進め方には十分に責任を持つつもりです」



**Sundar Pichai**

CEO of Google and Alphabet

## 2018 年以降、Google の AI 原則に基づく

### AI のあるべき姿

- 1 社会的にとって有益である
- 2 不公平なバイアスの発生、助長を防ぐ
- 3 安全性確保を念頭においた開発と試験
- 4 人々への説明責任
- 5 プライバシー・デザイン原則の適用
- 6 科学的卓越性の探求
- 7 これらの基本理念に沿った利用への技術提供

### Google が追求しない用途

- 1 総合的にみて有害または有害な可能性がある
- 2 危害を与えることを主な目的とするテクノロジー
- 3 国際的に認められた規範に反するような監視
- 4 国際法の理念や人権に反する

Google は 20 年以上  
にわたって AI を開発  
しており、すでに主要  
製品の原動力となっ  
ています



# ここでいう責任とは

1. ポリシー
2. テクノロジー
3. ユーザーのためのコンテキストの提供
4. 業界全体での取り組み


Google のポリシーでは、AI による生成であるかどうかにかかわらず、操作された虚偽のメディアをかなり前から禁止してきました


**Manipulated media**

We do not allow content that:

- deceives users through manipulated media related to politics, social issues, or matters of public concern.

[Tips for understanding this policy](#)

 Learn more about the commonly used policy terms and what they mean in the [glossary](#).

 Give feedback about this article

Was this helpful?





# 個人の顔や声を模倣した、**改変または合成されたコンテンツ**の報告

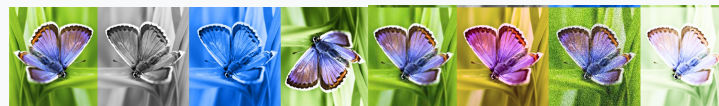
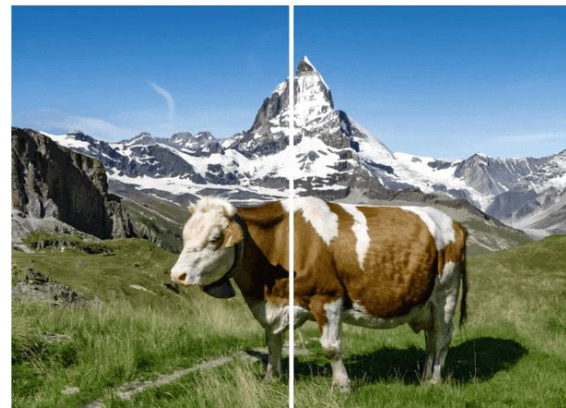
**Someone changed or faked my voice or image without my permission.**

"Altered" or "synthetic" refers to content that looks or sounds like you, but was significantly edited or generated by AI or other tools.

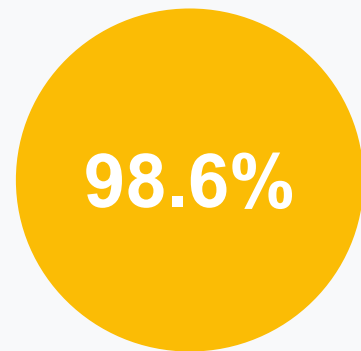
[Learn more about altered or synthetic content labels.](#)

[Report altered or synthetic content](#)

AI が生成した画像に  
ウォーターマーク(電子  
透かし)を入れて識別す  
るツール **SynthID** を発表  
し、Imagen 2 から利用で  
きるようになりました



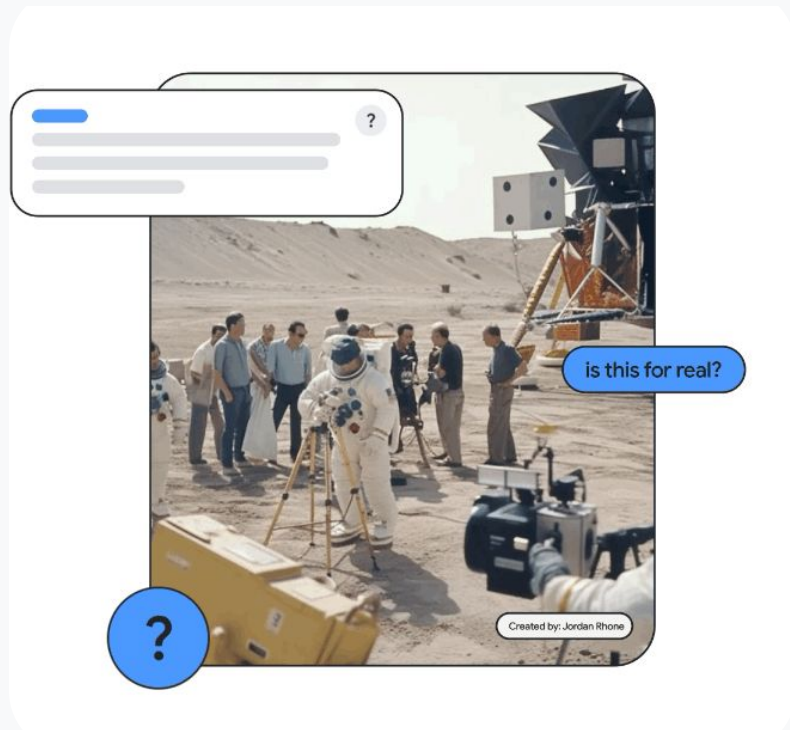
# 合成メディア検出の研究に積極的に取り組んでいます



Google の AudioLM モデルによって  
生成された合成音声の検出精度

サンプルとなったのは、短い話し言葉でした。この有望な研究は、今後、他のタイプの音声に広がることが期待されます

オンラインで目にする  
情報の信頼性を**評価す  
るツール**を提供して  
います



IPTC メタデータや  
SynthID 電子透かしに  
基づき、ある画像が AI  
によって生成されたも  
のと判明している場  
合、その旨を開示しま  
す



実物のように見えるコンテンツについては、  
改変または合成されたメディア（生成 AI を  
使用したものを含む）で作成したコンテンツ  
であることを視聴者に  
開示するよう、クリエイターに求めます

Check out my pizza creation! Saved as private

Details Monetization Ad suitability Video elements Checks Visibility

My video contains paid promotion like a product placement, sponsorship, or endorsement

By selecting this box, you confirm that the paid promotion follows our ad policies and any applicable laws and regulations. [Learn more](#)

**Altered content**

Do any of the following describe your content?

- Makes a real person appear to say or do something they didn't say or do
- Alters footage of a real event or place
- Generates a realistic-looking scene that didn't actually occur

Yes

No

To follow YouTube's policy, you're required to tell us if your content is altered or synthetic and seems real. This includes realistic sounds or visuals made with AI or other tools. Selecting "yes" adds a label to your content. [Learn more](#)

**Automatic chapters**

Allow automatic chapters (when available and eligible)

Video link: <https://youtu.be/245fvdgb>

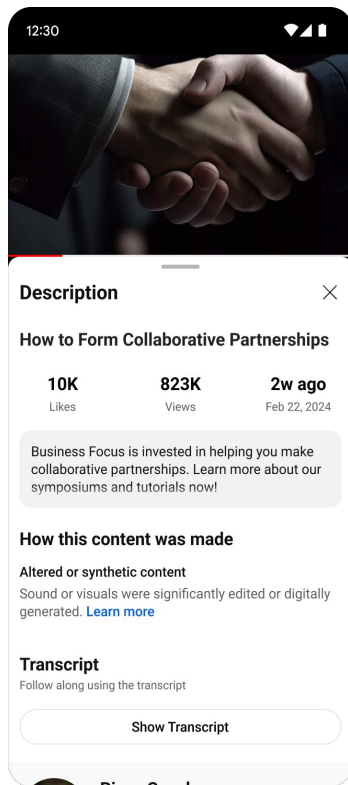
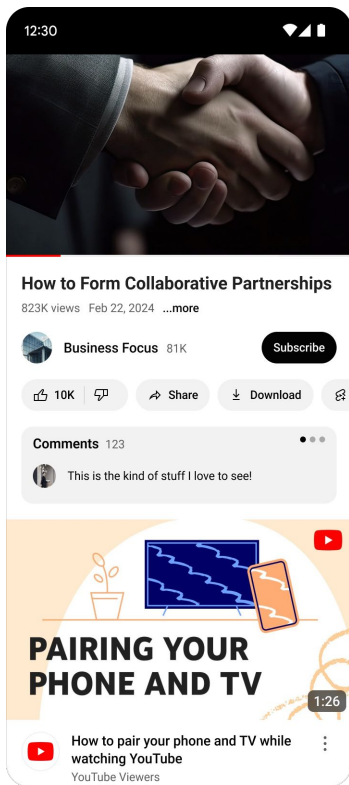
Filename: Check out my pizza creation!.mp4

Uploading ... 50% done, 6 minutes left

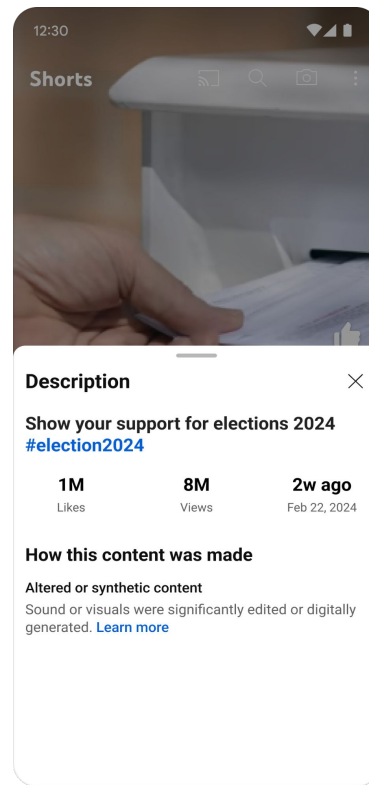
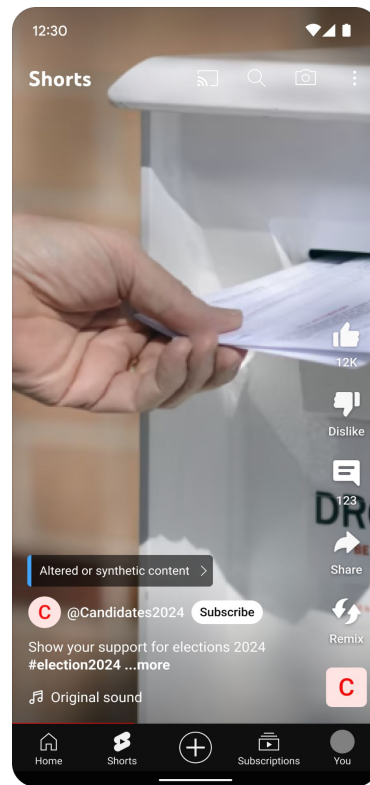
NEXT



## ユーザー コンテキスト



説明欄に表示されるラベルの例



動画プレーヤーに表示されるラベルの例

# 偽・誤情報問題啓発キャンペーン



ほんとかな？が、あなたを守る。



協力：総務省・国際大学GLOCOM



## ①フェイクニュースが 身近に存在する

人を混乱させるために流される誤情報・偽情報が存在する事を、身近な自分事として捉える



## ②ファクトチェック をしよう

立ち止まり、ファクトチェックをする大切さを知る

- 他の情報と比べてみる
- 情報の発信元を確かめる
- いつ頃書かれたものか確かめる
- 一次情報を確かめる



## ③拡散することの リスクを知ろう

軽い気持ちで拡散する事で、自分のリスクになることを知る

- 法的責任を問われる
- 人を傷つける
- 信用を失う
- 拡散に加担し社会への迷惑となる





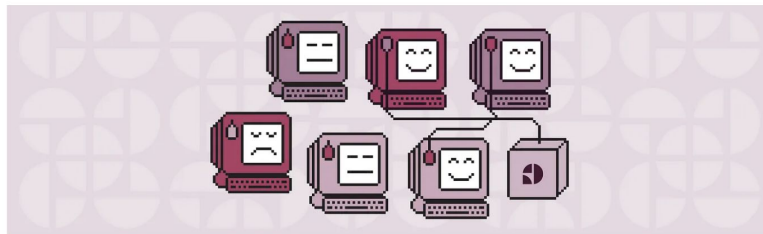
# エコシステム全体の安全性を高めるツールや知見を共有します



Jigsaw

Apr 21 · 6 min read · Listen

## Reducing Toxicity in Large Language Models with Perspective API



開発者が利用できる Perspective API は、害悪をフィルタリングします

# 業界、市民社会、学術 界とともに責任ある AI を構築します

ML  
• Commons



Frontier Model Forum のパートナー

ANTHROPIC



# 私たちは、C2PA に参加し、来歴情報の主だった標準仕様の開発を進めています

## Google to join C2PA to help increase transparency around digital content

### Google Joins C2PA Steering Committee

*Google to join C2PA to help increase transparency around digital content*

SAN FRANCISCO, Calif. — February 08, 2024

Today, the Coalition for Content Provenance and Authenticity ([C2PA](#)), a global standards body advancing transparency online through certifying the provenance of digital content, announced that Google has joined C2PA as a steering committee member.

Google joining the C2PA marks a significant moment for bringing more transparency to digital content. In joining, Google will help to further the adoption of [Content Credentials](#), the C2PA's technical standard for tamper-resistant metadata that can be attached to digital content, showing how and when the content was created or modified. Alongside other steering committee members including Adobe, BBC, Intel, Microsoft, Publicis Groupe, Sony and Truepic, Google will collaborate to further develop the C2PA's technical standard for digital content provenance. With this, Google is also actively exploring how to incorporate Content Credentials into its own products and services in the future.

Additionally, Google's participation, which also includes YouTube, will help to drive broader awareness of Content Credentials as a key resource to help people around the world understand the content they're viewing and increase trust in the digital ecosystem.

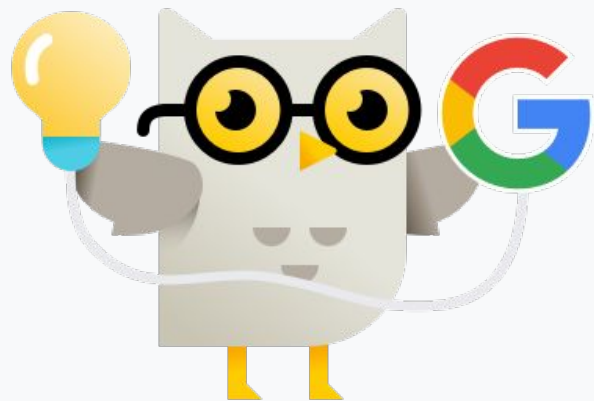
"At Google, a critical part of our responsible approach to AI involves working with others in the industry to help increase transparency around digital content," says Laurie Richardson, VP of Trust and Safety at Google. "This is why we are excited to join the committee and incorporate the latest version of the C2PA standard. It builds on our work in this space – including Google DeepMind's SynthID, Search's About this Image and YouTube's labels denoting content that is altered or synthetic — to provide important context to people, helping them make more informed decisions."

"It is more important than ever to have a transparent approach to digital content that empowers people to make decisions," said Andrew Jenks, C2PA Chair. "The C2PA standards are undoubtedly leading the charge in this endeavor, and we're thrilled with the growth and adoption. Google's membership is an important validation for the C2PA's approach. We encourage others to join us in expanding the use of

# 合成メディア検出と来歴情報の技術

「合成メディア」または「ディープフェイク」とは、生成 AI を使ってゼロから作成・編集されたコンテンツのことです

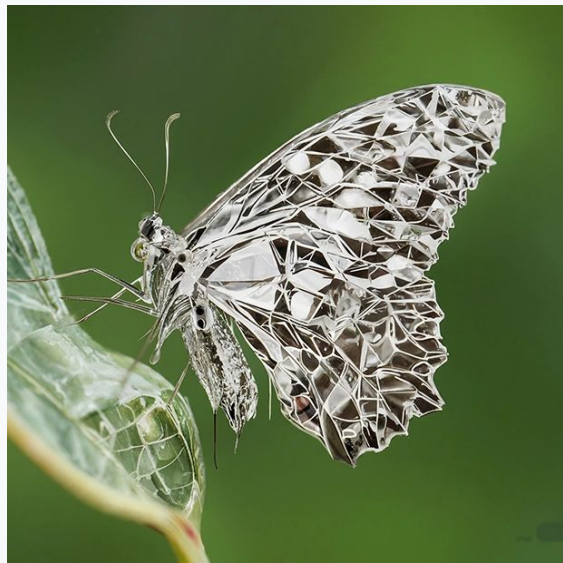
コンテキスト、ランキン  
グ、ポリシーといった既  
存ツールは、合成メデ  
ィアが招くリスクへの対  
処に役立ちます



本日は、オンラインコンテンツがどのように作成されたか(来歴)を評価するための技術的な取り組みに注目します



ウォーターマーク(電子透かし)は、メディアの一部として、目に見える形・見えない形で来歴情報を追加します





フィンガープリンティング  
は、「ハッシュ値」を作成  
します。これは他のコン  
テンツのハッシュ値と後  
で比較・照合できます

## Perceptual hashing

[Add languages](#)

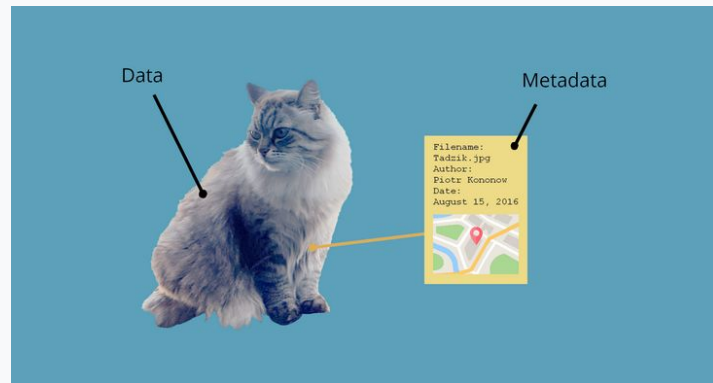
[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#)

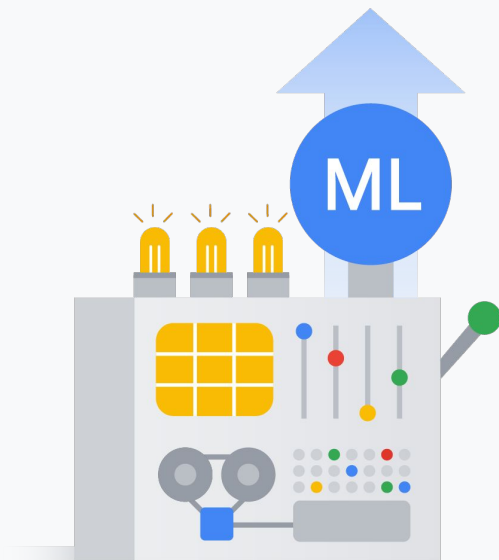
From Wikipedia, the free encyclopedia

**Perceptual hashing** is the use of a [fingerprinting algorithm](#) that produces a snippet, hash, or [fingerprint](#) of various forms of [multimedia](#).<sup>[1][2]</sup> A perceptual hash is a type of [locality-sensitive hash](#), which is analogous if [features](#) of the multimedia are similar. This is in contrast to [cryptographic hashing](#), which relies on the [avalanche effect](#) of a small change in input value creating a drastic change in output value. Perceptual hash functions are widely used in finding cases of online [copyright infringement](#) as well as in [digital forensics](#) because of the ability to have a correlation between hashes so similar data can be found (for instance with a differing [watermark](#)).

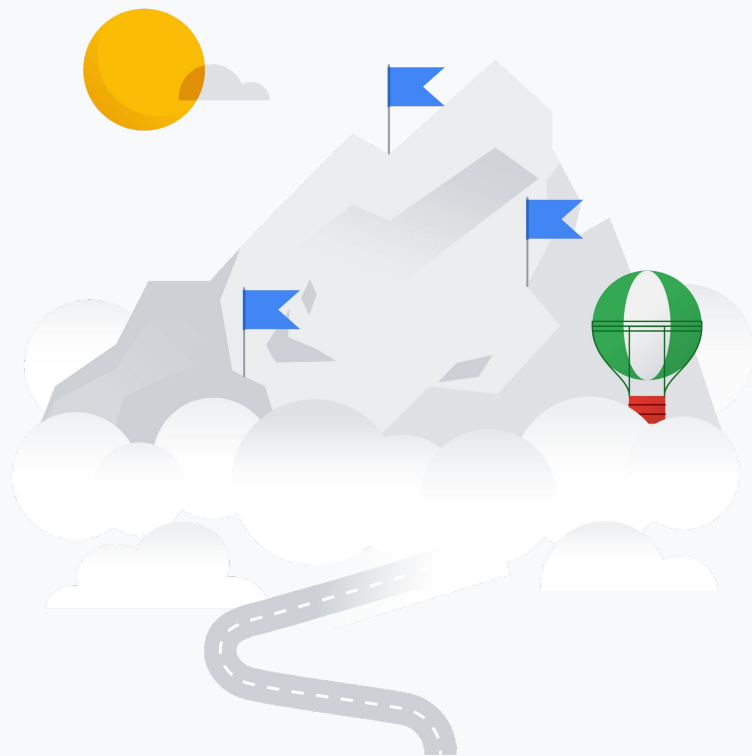
メタデータは、生成日  
やサムネイルなどの  
「データに関するデー  
タ」を署名入り・署名な  
しでファイルに付加しま  
す



**検出分類器**は、1つまたは複数の生成 AI システムによる出力を判別するよう訓練されたモデルです



どの方法も便利ですが、**限界**があります。  
組み合わせても確実な  
解決法にはなりません





# Thank you

