Adobe

データシート

Adobe Firefly：倫理とバイアス

適切な方法で実装された生成 AI は、既存のワークフローに新たな可能性とパワーをもたらし、創造性を増幅します。しかし、境界を押し広げる技術のほとんどと同様に、生成 AI の有望性には、著作権の侵害から有害な影響やバイアスに至るまで、対処すべき課題やリスクが伴います。クラウド技術における 10 年以上にわたる AI イノベーションのレガシーを基盤に、アドビは説明責任、社会的責任、透明性という AI 倫理原則に沿って、Firefly による生成 AI を開発およびデプロイしました。

アドビはクリエイティブな生成 AI を注意深く開発することに注力しており、ユーザーが生成するコンテンツに安心することができます。

安全に商用利用できる Firefly

Firefly が独自開発した生成 AI モデルは安全に商用利用ができます。アドビがライセンスや著作権の問題に責任を持って対応していることを知ることで、お客様は安心して出力を公開することができます。Firefly の生成 AI モデルは、Adobe Stock などのライセンス済みコンテンツや著作権が消滅したパブリックドメインコンテンツから収集した膨大な数の画像、オーディオ、動画を用いて訓練されています。

知的財産（IP）補償

一部の顧客は、万が一 Firefly の出力に関わるクレームが発生した場合に備え、契約上の知的財産（IP）補償が付帯する権利を購入する機会を得ることができます。これは、Firefly の出力が第三者の特許権、著作権、商標権、パブリシティ権、またはプライバシー権を直接侵害または違反していると主張する請求に対して、アドビが法的防御および金銭的損害賠償の責任を負うことを意味します。アドビの Firefly 出力に関する責任は、顧客がアドビと締結している既存契約における技術ベースの知的財産権請求に対する制限と一致します。アドビの補償は、Firefly 出力の修正、Firefly 出力が使用されるコンテキスト、カスタムトレーニングのために提供されるコンテンツ、または顧客契約に違反する Firefly 出力の使用を含む、顧客のアクションまたはコンテンツに基づく、またはそれらに起因する請求には適用されません。

Firefly の商用利用の安全性と補償について詳細を見る

ビジネスにとって安全

AI 機能のリリースに先立ち、アドビのテストチームは有害な出力や悪影響を軽減するため詳細なプロセスに従っています。具体的には、分類およびフィルタリングが必要な言語の優先順位付け、プロンプトのブロックリストおよび拒否リストの作成、職場での閲覧に適さないコンテンツ（NSFW）のインスタンス低減を実施します。

Firefly のコンテンツモデレーションについて詳細を見る

説明責任

アドビでは、AI が支援するツールの成果に対して責任を自覚し、以下のような AI 倫理への取り組みを維持するための標準化されたプロセスを導入しています。

- トレーニングデータと AI モデルを追跡するためのガバナンス手順の**確立**
- エンジニアや製品マネージャーへの AI の倫理問題に関する**教育**
- AI 開発の推進を監督し、AI 倫理上の懸念事項に対する意見交換の場を提供する AI 倫理審査委員会の**創設**
- 新たな AI 技術をデプロイする前に AI 倫理審査が行われることを確実にするため、サービス開発プロセスにおいて AI 影響評価を**義務付け**
- AI のデプロイメント後のネガティブな影響を確実にフィードバックし、改善するためのプロセスを**開発**

当社では、社内外を問わず AI に関する懸念事項の受付と対応に特化したプロセスを整備し、適切な場合には是正措置を講じることを保証します。Firefly には、アドビ社内の AI 倫理審査委員会に加え、ユーザーが機能の結果にバイアスや不正確さなどを感じた場合に報告できるビルトインのフィードバック機能を備えています。このようなユーザーコミュニティとのフィードバックループは、アドビのツールが危害を最小限に抑え、説明責任を果たし続けるための重要な方法です。

アドビの説明責任への取り組みについて詳細を見る

社会的責任

アドビでは、自社の影響と世界に提供するイノベーションに対する責任を重視しています。アドビは、AI の開発と利用に関する思慮深いガードレールを設けることで、潜在的な悪影響を緩和しつつ、社会に利益をもたらす AI の可能性を最大限に実現することにつながると信じています。責任ある AI 開発に対する私たちのアプローチは、以下を包含しています。

- AI システムを思慮深く**設計**
- エンドユーザーとのインタラクション方法を**評価**
- 望ましくない有害なバイアスを軽減するためにデューディリジェンスを**実施**
- AI 技術が人に与える影響を**評価**

有害性とバイアスの軽減

AI の品質はそのトレーニングに使用される学習データの品質に等しいことを踏まえ、アドビでは、あらゆるデータにバイアスがあることを認識しています。AI の責任ある開発には、潜在的なバイアスとそれによる危害を予測し、そのような懸念を軽減するための予防措置を講じ、危害の発生頻度を測定および文書化し、予期せぬ有害な結果を監視および対処するシステムを確立することが必要です。私たちは包摂性を念頭に設計していますが、AI システムが人間の属性（例：人種、性別、所得、宗教、障害など）に関連する有害なバイアスや固定観念を拡散させる可能性のある 、不公平 、差別的 、不正確な結果をもたらすことのないよう、その人的影響を継続的に評価しています。アドビの責任と倫理的取り組みが確実に履行されるよう、Firefly の生成 AI の影響を継続的に評価するための専用リソースと標準化されたプロセスを実装しました。このプロセスには以下が含まれます。

- AI 倫理委員会 – 開発チームと協力し、開発のあらゆるステージで AI 倫理原則を遵守するとともに、潜在的に有害な影響やバイアスを積極的に緩和します
- AI 倫理審査委員会 – AI を活用したアドビのすべての機能と製品を、リリース前に厳格に審査します

透明性

アドビは透明性が信頼を築くとも考えており、お客様、パートナー、業界関係者の皆様が当社製品に寄せられる信頼を真摯に受け止めております。アドビは、AI によって生成された出力にも透明性を拡大することを約束し、Adobe Firefly で生成されるすべての画像に Content Credential を適用します。Centent Credential は、Content Authenticity Initiative (CAI) を通じて利用可能な技術エコシステムの一部であり、アドビはその創設メンバーです。Content Credential は、改ざん防止機能を備えた新たなメタデータであり、信頼性と信用を構築すると同時に、オンライン上の誤情報の拡散を防ぐ役割を果たします。これは、クリエイターが作品に対して評価を得られることを保証するだけでなく、画像に生成 AI が使用された場合にそれを示すことで、他のユーザーや消費者がその制作方法を理解できるようにするものです。

アドビの透明性の原則について詳細を見る。

アドビが、革新、学習、成長を続ける中で、AI テクノロジーの責任ある開発にどのように取り組んでいるかについては、担当者にお問い合わせください。

Adobe