生成AIの著作権侵害等のリスクとその低減技術動向 シンポジウム

生成AIの著作権侵害等のリスクとその低減技術動向調査結果

みずほリサーチ&テクノロジーズ
情報通信研究部

2024年2月26日

ともに挑む。ともに実る。
MIZUHO

目次

1. 調査目的・背景

1-1. 目的
1-2. 背景

1-1. 目的

AIと著作権に関する技術動向調査

- ✓ 委託元: 国立研究開発法人 新エネルギー・産業技術総合開発機構（NEDO）
- ✓ 委託先: みずほリサーチ&テクノロジーズ株式会社

◆ 目的

- ✓ 主に日本国内の事業者が生成AIの事業化の際に検討するべき、著作権等の権利侵害をはじめとした生成AIにおけるリスクとそのリスク低減技術の動向調査を行う
- ✓ リスクを低減するため今後必要と考えられる技術開発の検討・提案を行う

本シンポジウムでは・・・・・・

- ➢ 生成AIの事業化の際に検討するべきリスクについて
- ➢ 生成AIにおけるリスクの低減技術について

の調査結果の公表を行う

1-2. 背景

◆ 生成AI技術の急激な発展

- ✓ 画像生成、文章作成等の分野で、人間が作成したものと見分けがつかないほど精巧なコンテンツを生成する能力を秘めた強力な生成AIが、相次ぎ公表され、急速に普及
- ✓ 多様な分野で活用されはじめており、人々の生産性や創造性の向上が期待されている

◆ 増大する生成AIにおけるリスク

- ✓ 生成AIには、他者の権利を侵害するリスクが指摘されており、その活用には注意が必要である
  - ➢ 学習に利用するためのデータの収集・複製や生成AIによる生成物の利用などが、著作権やプライバシーなどの権利侵害になる恐れ
- ✓ また近年生成AI事業者は、生成AIにより生成した画像やテキストなどのコンテンツに対する社会的責任を負うことも求められている

→ 権利侵害のリスクおよび社会的責任に関するリスクに対するリスク低減技術の動向調査を実施

2. 生成AIにおける権利侵害等に関するリスクについて

2-1. 生成AI事業
2-2. 生成AIに関する法令・ガイドライン
2-3. 本調査で取り上げるリスク

2-1. 生成AI事業

◆ 生成AI事業の動向

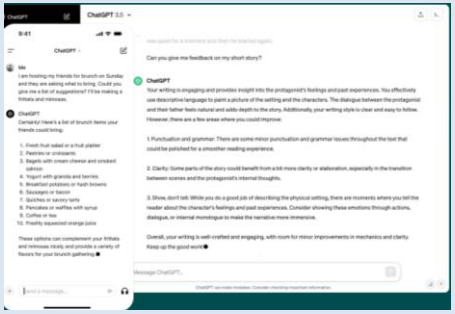
✓ 2020年以降、画像およびテキストを生成するAI技術が急速に発展。まるで人間が作成したような自然な文章や写実的な画像など、精巧なコンテンツを作成可能なAIツールが数多く公開され普及が進む

生成AIサービス例

出力形式に着目して分類

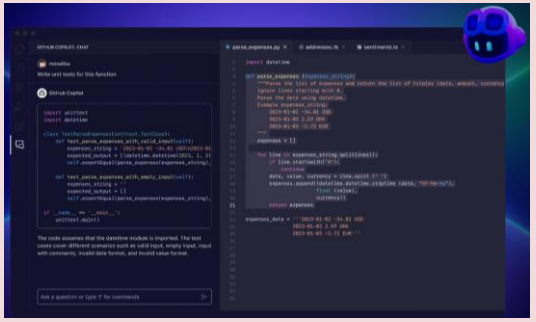
種類	サービス例	事業者
テキスト 生成AI	Bard	Google
	ChatGPT	OpenAI
プログラムコード 生成AI	GitHub Copilot	GitHub
	Codey	Google
画像/動画 生成AI	Stable Diffusion	Stability AI
	Adobe Firefly	Adobe
	Bing Image Creator	Microsoft
音声/音楽 生成AI	Voicebox	Meta
	Stable Audio	Stability AI
その他 生成AI (アバター、3次元モデルなど)	SenseAvatar	Sensetime

チャットでの文章生成(ChatGPT)



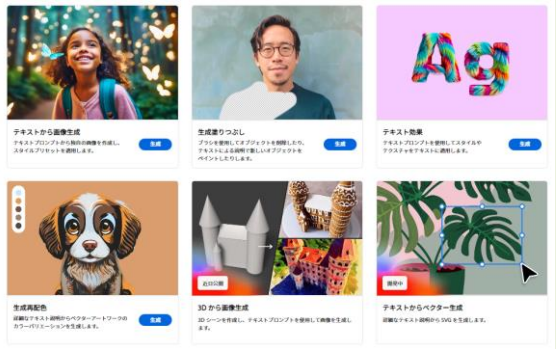
出典: OpenAI ChatGPT サイト(アクセス: 2023.12.06)
https://openai.com/chatgpt

プログラムのコード生成・作業支援(GitHub Copilot)



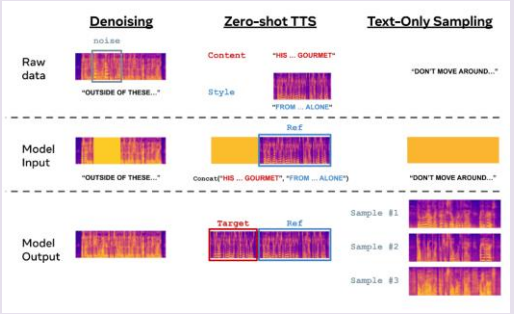
出典: Github 'Github Coplilot'サイト (アクセス: 2023.12.06)
https://github.com/features/copilot

文章からの画像生成や補完(Adobe Firefly)



出典: Adobe 'Adobe Firefly'サイト (アクセス: 2023.12.06)
出典: https://firefly.adobe.com/

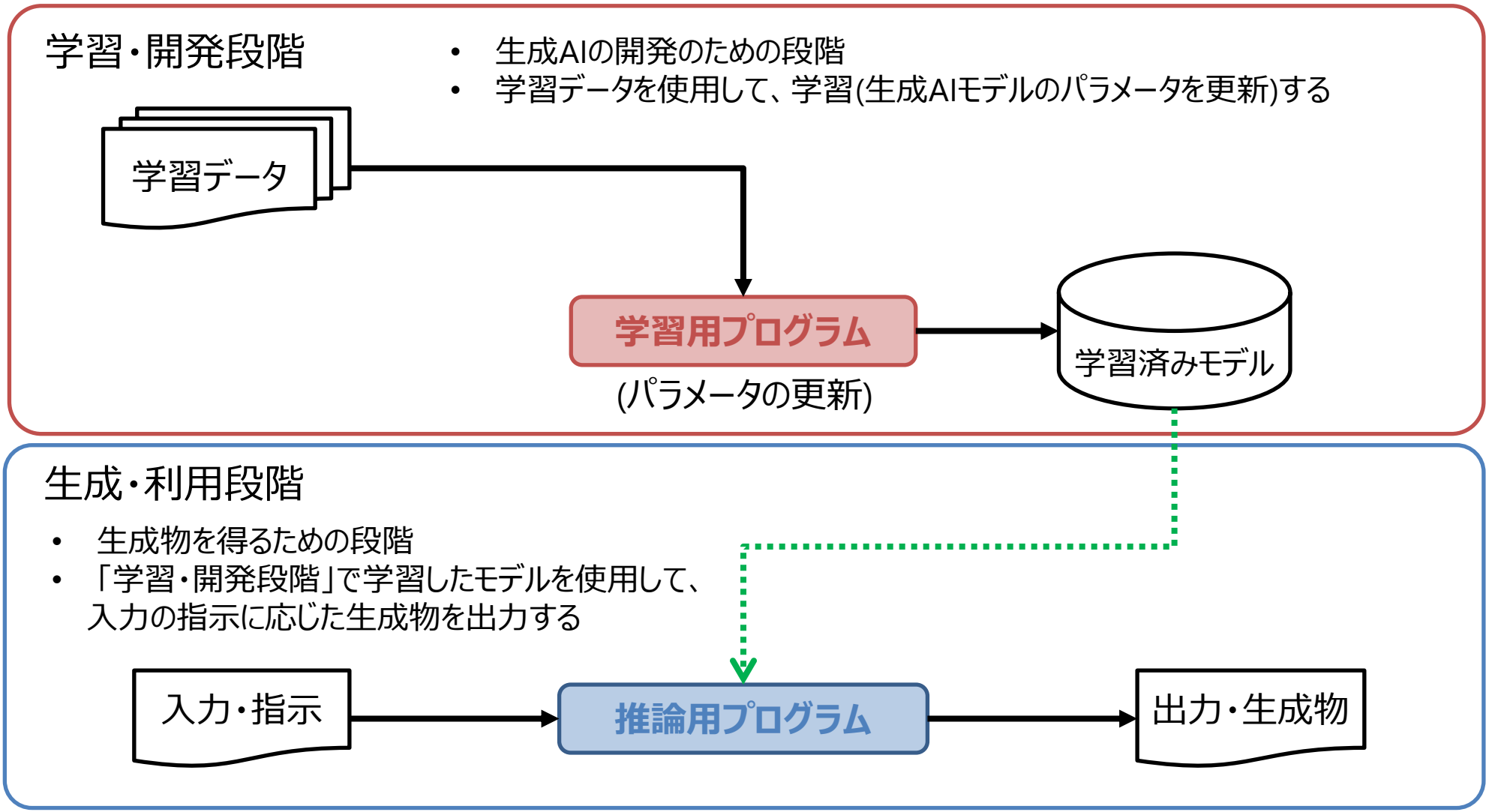
音声の変換やテキストからの音声生成(Voicebox)



出典: Meta Voiceboxサイト (アクセス: 2023.12.06)
https://voicebox.metademolab.com/

2-1. 生成AI事業

◆ 生成AIにおける開発・利用の流れ



2-2. 生成AIに関する法令・ガイドライン

◆ 日本の生成AIに関連する法令・ガイドライン

- ✓ 「著作権法」や「個人情報の保護に関する法律」を考える際には、「開発・学習段階」と「生成・利用段階」で、行われている利用行為が異なり、関係する法律の条文も異なるため、リスクの検討の際には、分けて考える必要がある
- ✓ 日本では、AIの利活用を直接規律する法律はなく、ガイドラインなどにより指針などが示されている

著作権	個人情報の保護	生成AIに関するガイドライン
著作権法 →日本の著作権に関する法律 （文化審議会においてAIと著作権に関する考え方の整理に向けた議論が行われている） (学習・開発段階) ・「著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない」場合には、その必要と認められる限度において著作物の利用が認められている。ただし、「著作権者の利益を不当に害することとなる場合」を除く(30条の4) (生成・利用段階) ・著作権侵害となるか否かは、(AIを利用しない場合の著作権侵害と同様に)「類似性」および「依拠性」により判断される	**個人情報の保護に関する法律(個人情報保護法)** → 日本の個人情報保護に関する規則 ・氏名や識別番号などのテキストデータや、顔写真などの画像データなど、個人を識別可能な情報である個人情報の取り扱いに関する法律 (学習・開発段階) ・個人情報を取得するにあたっては、利用目的の通知または公表が必要。また要配慮個人情報の取得時には、本人の同意も必要となる (生成・利用段階) ・生成されたデータに個人情報が含まれる場合、個人情報保護法などにおける第三者提供に関する規定との関係に留意が必要となる ・個人データ(データベース化された個人情報)を第三者に提供する場合、本人の同意を得るか、個人情報保護委員会に届け出た上で本人に異議がある場合には提供を中止するオプトアウト方式の実施が必要	**「高度な AI システムを開発する組織向けの広島プロセス国際指針/行動規範 (広島ＡＩプロセスに関するＧ７首脳声明)」(2023.10)** →基盤モデルや生成AIを含む高度なAIシステムに関するG7による首脳声明 ・ 日本が議長国となり、AIシステムに関する国際的な指針及び行動規範を発表 ・ AIシステムに対するリスク評価および適切な措置の導入や透明性の確保、セキュリティ対策、コンテンツの認証、有害な偏見バイアスを軽減するためのデータの質の管理や個人データおよび知的財産の保護などを求める 他

2-2. 生成AIに関する法令・ガイドライン

生成AIの事業者は、著作権やプライバシーなどに関する法令を遵守するだけでなく、生成AIによる生成物に対する社会的責任への配慮が必要となる

◆ 法令の遵守

✓ 生成AIの学習・開発、生成・利用の各段階において、**著作権**や**プライバシー**などに関する法令を遵守し、権利侵害へ対処することが求められる

◆ 社会的責任

✓ 生成AIの事業者は以下のような社会的責任を持つ

- 生成物に対する責任
  - **生成内容に関する倫理・公平性**
  - **生成AIの使用についての明示責任**
- 使用するデータ・アルゴリズムの透明性確保
- セキュリティの担保
- 適切なリスク評価

等

2-3. 本調査で取り上げるリスク

本調査においては、技術調査の対象として、以下のリスクを取り扱うこととした

◆ 権利侵害に関して法令に抵触するリスク

✓ 著作権侵害のリスク

- 学習データの収集や、生成AIモデルにより生成されたデータの公開・配信により著作権侵害となるリスク

✓ プライバシー侵害のリスク

- 学習データの収集や、生成AIモデルにより生成されたデータの公開・配信により個人情報保護の規制に違反するリスク

◆ 生成AIの生成物に対する社会的責任に関するリスク

✓ 生成内容の倫理・公平性に関するリスク

- 生成AIによって、非倫理的な内容や、差別などの特定のバイアスを含んだ内容が生成されるリスク

✓ 生成AI使用有無の不透明性に関するリスク

- コンテンツに対して生成AIの使用有無が不透明になることにより、生成AIによる生成物が現実のデータと区別がつかなくなるリスク

2-3. 本調査で取り上げるリスク

◆ リスクの発生段階による整理

※リスクの発生段階(×リスク対策を行う段階) により6つのリスクに整理

リスク項目	学習・開発段階	生成・利用段階
著作権侵害のリスク (略記:「著」)	基盤モデルおよび生成AIモデルの開発のため、学習データの収集等を行う際に、著作権侵害となるリスク ① ※日本は一定の条件のもとで学習のための著作物の複製等を許諾なく行うことが認められており、一部のケースにおいてリスクとなりうる	生成AIモデルにより生成されたデータが、既存の著作物に類似している場合に、著作権侵害となるリスク ③
プライバシー侵害のリスク (略記:「プ」)	基盤モデルおよび生成AIモデルの開発のため、学習データの収集等を行う際に、個人情報保護の規制に違反するリスク ②	生成AIモデルにより生成されたデータに、個人情報が含まれてしまい、プライバシーを侵害するリスク ④
生成内容の倫理・公平性に関するリスク (略記:「倫」)	生成AIにより出力された生成物に対して発生するリスクのため、除外	生成AIによって、非倫理的な内容や、差別などの特定のバイアスを含んだ内容が生成されるリスク ⑤
生成AI使用有無の不透明性に関するリスク (略記:「透」)		コンテンツに対して生成AIの使用有無が不透明になることにより、生成AIによる生成物が現実のデータと区別がつかなくなるリスク ⑥

✓ 各リスクの内容について次ページ以降で詳細説明

2-3. 本調査で取り上げるリスク

① 学習・開発段階の著作権侵害のリスク

✓ 基盤モデルおよび生成AIモデルの開発のため、学習データの収集等を行う際に、著作権侵害となるリスク

✓ 学習・開発段階においては、日本は一定の条件のもとで学習のための著作物の複製等を許諾なく行うことが認められている

➢ 著作権法30条の4の規定により、「著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない」場合には、その必要と認められる限度においてAI開発のような情報解析等に対して著作物の利用が認められている。ただし、以下の場合を除く

① 非享受目的と享受目的が併存する場合

② 著作権者の利益を不当に害することとなる場合

└ 例えば、情報解析を行う者の用に供するために作成されたデータベースの著作物を情報解析のために複製等する場合は、30条の4の但し書きに当てはまり、著作権侵害になってしまう可能性がある

➢ (参考)海外での著作物の利用に関する規定

– EU、イギリスなど：一定の条件のもと、データマイニングにおける著作物の利用が認められている

2-3. 本調査で取り上げるリスク

② 学習・開発段階のプライバシー侵害のリスク

- ✓ 基盤モデルおよび生成AIモデルの開発のため、学習データの収集等を行う際に、個人情報保護の規制に違反するリスク

- ✓ 日本は学習データ収集時に、利用目的の通知または公表が必要(個人情報保護法)
  - ➢ また要配慮個人情報の取得時には、本人の同意が必要

- ✓ (参考)EU: 管理者の身元や連絡先、処理の目的、同意の撤回権など、日本より多くの情報の提供が必要。また「本人の同意」または「正当利益」などを含む適法化の根拠が必要(一般データ保護規則)

2-3. 本調査で取り上げるリスク

③ 生成・利用段階の著作権侵害のリスク

- ✓ 生成AIモデルにより生成されたデータが、既存の著作物に類似している場合に、著作権侵害となるリスク

- ✓ 生成された著作物を公開・配信などした際に著作権侵害となるかは、その「類似性」および「依拠性」により判断される
  - ➢ 人がAIを利用せずに作成したものと同様の判断が行われる
  - ➢ 「類似性」「依拠性」に関しては様々な見解がある
    - – 類似性: どの程度類似していれば著作権侵害となるか
      - » 既存の著作物の「表現上の本質的特徴部分」を感得できる程度に類似しているか？
    - – 依拠性: AIの依拠性をどのように考えるべきか
      - » 見解(一例)
        - 「元の著作物がAIの学習に用いられていれば依拠性を認める？」
        - 「AI生成物が、学習に用いられた元の著作物の表現と類似していれば、依拠性ありと推定?」

        など

2-3. 本調査で取り上げるリスク

④ 生成・利用段階のプライバシー侵害のリスク

- ✓ 生成AIモデルにより生成されたデータに、個人情報が含まれてしまい、プライバシーを侵害するリスク

- ✓ 生成されたデータに個人情報が含まれる場合、個人情報保護法などにおける第三者提供に関する規定との関係に留意が必要となる

2-3. 本調査で取り上げるリスク

⑤ 生成内容の倫理・公平性に関するリスク

- ✓ 生成AIによって、
  - ➢ 倫理に反する内容
  - ➢ 性差別や人種差別など、特定のバイアスを含み公平性に欠けた内容

  などが生成されるリスク

- ✓ 多くの生成AIに関する法令やガイドラインで、倫理・公平性に関する責任を持つことが求められている
  - ➢ 「高度な AI システムを開発する組織向けの広島プロセス国際指針」(広島ＡＩプロセスに関するＧ７首脳声明)
  - ➢ 連邦取引委員会によるガイドライン(アメリカ)
  - ➢ 人工知能サービス管理暫定弁法(中国)
    - – 等

2-3. 本調査で取り上げるリスク

⑥ 生成AIの使用有無に関する不透明性リスク

- ✓ コンテンツに対して生成AIの使用有無が不透明になることにより、生成AIによる生成物が現実のデータと区別がつかなくなることに関するリスク

- ✓ ディープフェイク・AIフェイクといわれるような、生成AIにより生成された偽情報、偽動画、偽画像などの悪用の懸念

- ✓ 多くの法令(法令案含む)およびガイドラインで生成AIによる生成物であることの明示について記載がされている
  - ➢ 「高度な AI システムを開発する組織向けの広島プロセス国際指針」(広島ＡＩプロセスに関するＧ７首脳声明)
  - ➢ 人工知能の安全・安心・信頼できる開発と利用に関する大統領令(アメリカ)
  - ➢ AI規制法案(EU)
  - ➢ インターネット情報サービスディープフェイク管理規定(中国)　　等

3. 生成AIにおける権利侵害等に関するリスクの低減技術について

3-1. リスク低減技術についての調査結果概要
3-2. 学習・開発段階に適用可能なリスク低減技術
3-3. 生成・利用段階に適用可能なリスク低減技術

3-1. リスク低減技術についての調査結果概要

◆ 生成AIに対するリスク低減技術を調査

- ✓ 既存の生成AI事業における取り組み (公開情報の文献調査やヒアリング)
- ✓ 研究事例調査 (論文調査)

※調査期間: ～2023年12月

◆ 調査結果

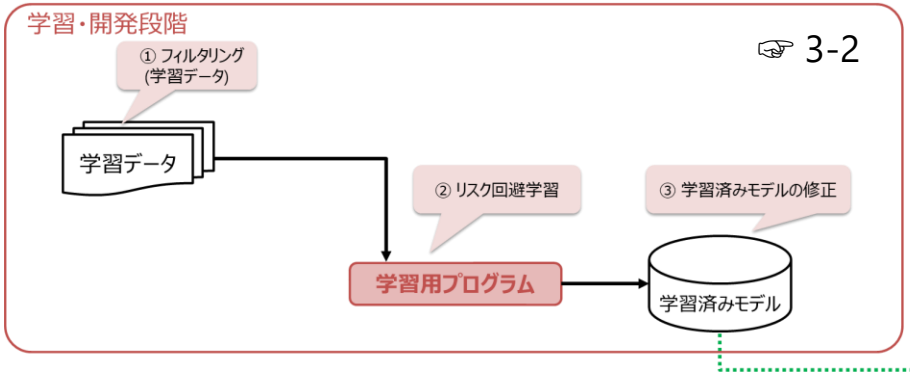
- ✓ リスク低減に使用できる可能性のある手法について、生成AI事業者の行っている取り組みだけでなく研究段階の手法も含め調査を行い、各手法の適用できる段階(学習・開発段階/生成・利用段階)とアプローチの仕方に応じて整理 (☞次ページに一覧表示)

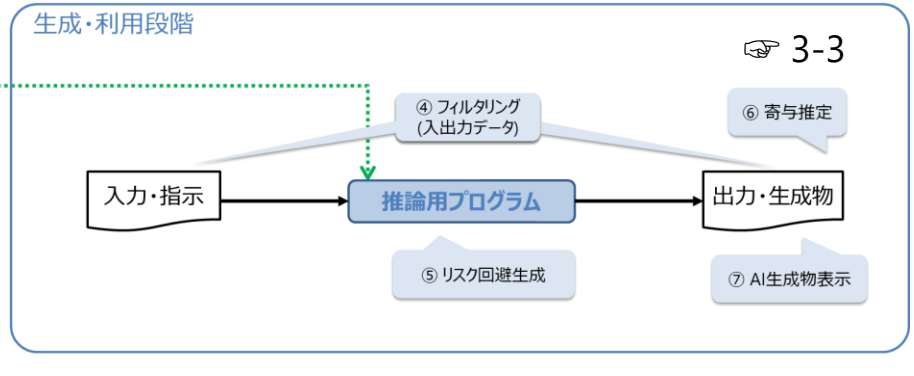
- ✓ 現状、生成AI事業者は、データの管理や利用規約など非技術的な施策を含め、複数のリスク低減のための取り組みを組み合わせて採用している

3-1. リスク低減技術についての調査結果概要

◆ 調査結果 リスク低減技術一覧

各手法を学習・開発段階に適用可能な手法と生成・利用段階に適用可能な手法に分け、合計7つのリスク対策の区分に整理





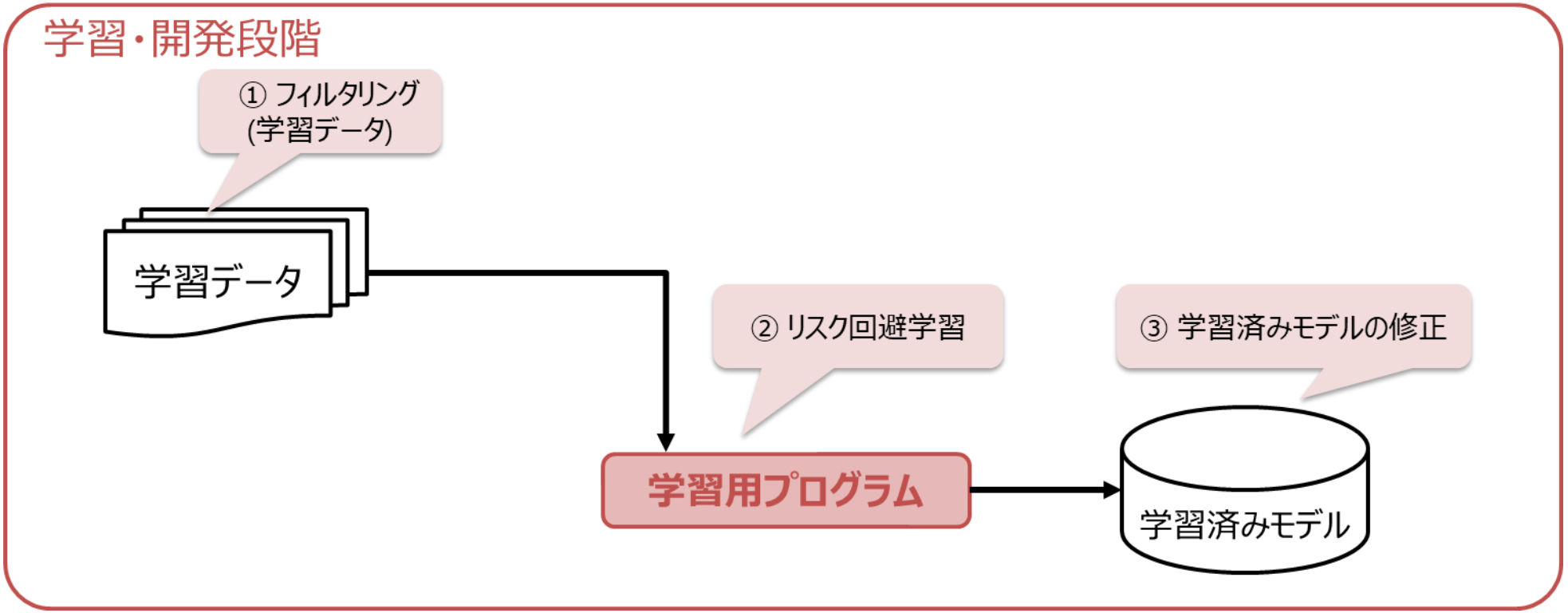
段階	リスク対策の区分	名称	概要	対応リスク					
				学習・開発		生成・利用			
				著	プ	著	プ	倫	透
学習・開発	フィルタリング(学習データ)	除外リストフィルタ	除外リスト(事前に用意する、権利侵害などの恐れがある除外候補のデータのリスト)と類似しているデータを除外する。	△		○		○	
		類似物排除フィルタ	学習データに複数の類似データが含まれると該当データに類似した出力が生成されやすくなるため、学習データ内の類似物を除外する。	△		○			
		個人情報除外フィルタ	特定の個人を識別可能な情報(PII)を検出して、加工または学習データから除外する。		○		○		
		倫理評価フィルタ	倫理的に問題のある内容を含む学習データを検出し、該当するデータを学習に使用しない。					○	
	リスク回避学習	学習データの複製生成抑制	学習に使用されたデータと類似性の高いデータが生成されにくくなるように学習する。	△		○			
		人工データでの学習	著作物・個人データを含まない人工的に生成されたデータをもとに学習を行う。主に基盤モデルの学習において利用可能な技術。	△	○	○	○		
		差分プライバシー	プライバシー情報の区別ができないようにノイズを加えて学習する。				○		
		Human in the Loop学習	生成AIの学習などを行う際に、生成AIによる生成物を人間が評価しそこから新たな学習データを作成するなど、生成AIの学習へ人間の評価のフィードバックを行う手法。			○	○	○	
	学習済みモデルの修正	アンラーニング	学習済みの生成AIモデルから特定概念を除去する。			○	○	○	
		アンバイアシング	学習済みの生成AIモデルがもつバイアスを軽減する。					○	

段階	リスク対策の区分	名称	概要	対応リスク			
				生成・利用			
				著	プ	倫	透
生成・利用	フィルタリング(入力・出力データ)	除外リストフィルタ	除外リスト(事前に用意する、権利侵害などの恐れがある除外候補のデータのリスト)と類似しているデータを検出して、生成AIへの入出力から取り除く。	○	○	○	
		個人情報除外フィルタ	特定の個人を識別可能な情報(PII)を検出して、生成AIへの入出力から取り除く/加工する。		○		
		倫理評価フィルタ	生成AIの入出力データに含まれるバイアスや差別的内容を検出し、取り除く/加工する。			○	
	リスク回避生成	プロンプトエンジニアリング	主にテキスト形式の入力(プロンプト)を受け付ける生成AIに対して、入力プロンプトの修正・改良や、複数回の実行を繰り返すなど、生成AIの利用方法を工夫することで所望の出力を得る技術。	○	○	○	
	寄与推定	寄与推定	生成された出力データに対して、その生成に寄与した学習データを推定する。	○			
	AI生成物表示	コンテンツ認証	データのメタデータに対してハッシュ値やデジタル署名を埋め込むことで、来歴情報（情報の発信者や流通経路）の検証や改ざん検知を可能にする。				○
		不可視的透かし	人間に知覚できない形でデータに情報(生成に使用したサービス名など)を埋め込み、所定の方法に従うことでその情報を抽出できるようにする。				○
		可視的透かし/明示的表示	ユーザが生成AIであることに気づかずに意図せずに生成物を利用することを防ぐために、生成AIによる生成物であることを明示する。				○

今回は、時間の都合上、一部の手法に絞って紹介

3-2. 学習・開発段階に適用可能なリスク低減技術

✓ 学習・開発段階に適用可能なリスク低減技術を、以下の3つの区分に整理



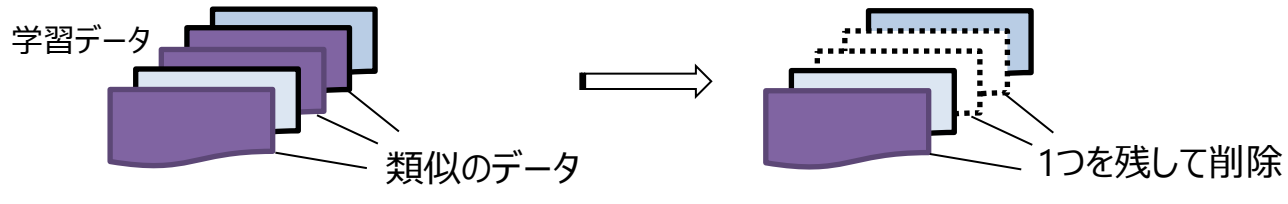
3-2. 学習・開発段階に適用可能なリスク低減技術

① フィルタリング(学習データ)

- ✓ 学習データのフィルタリング(絞り込み)を行うことで、著作権やプライバシー、倫理・公平性に関するリスクを低減
  - ➢ 除外リストフィルタ
  - ➢ 類似物排除フィルタ
  - ➢ 個人情報除外フィルタ
  - ➢ 倫理評価フィルタ

◆ 類似物排除フィルタ

- ✓ **学習データの中の類似したデータを取り除く技術**
  - ➢ 類似データが複数含まれている学習データを使用して学習を行うと、該当の学習データそのものが生成されやすくなることが知られている
  - ➢ 学習に使用するデータから、事前に似たデータを削除しておくことで、学習データがそのまま出力されてしまい、著作権等の侵害になるリスクを低減することが可能



  - ➢ 画像生成AI Dall-E 2 (OpenAI)などの開発においても使用されている

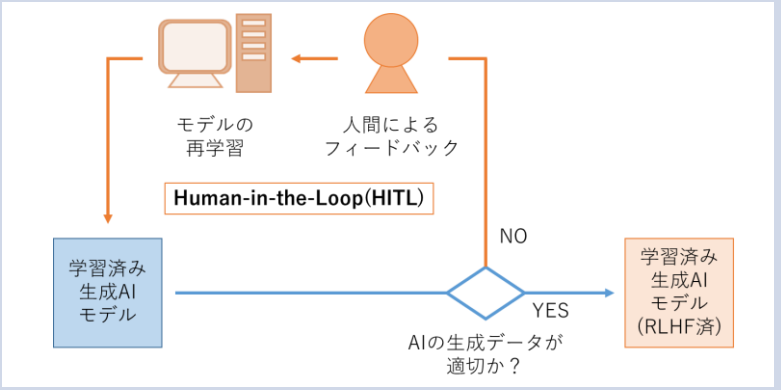
3-2. 学習・開発段階に適用可能なリスク低減技術

② リスク回避学習

- ✓ 人工的に生成されたデータを使用した学習や人間によるフィードバックを反映など、学習方法を工夫することで、著作権やプライバシー、倫理・公平性に関するリスクを低減
  - ➢ 学習データの複製生成抑制
  - ➢ 人工データでの学習
  - ➢ 差分プライバシー
  - ➢ Human in the Loop学習

◆ Human in the Loop学習

- ✓ **生成AIの学習に人間の評価をフィードバックする技術**
  - ➢ 例えば、学習済み生成AIによる生成物を人間が評価し不適切な場合には、入力に対する適切な生成物を作成して学習データに追加し、再度学習を行うといった方法がある
  - ➢ ChatGPT (OpenAI)の学習などにも利用されている



3-2. 学習・開発段階に適用可能なリスク低減技術

③ 学習済みモデルの修正

✓ 生成AIモデルに対して、特定の情報の消去やバイアスの調整などが必要となった際に、ゼロから再学習することなく、効率的に学習済みモデルを修正

- アンラーニング
- アンバイアシング

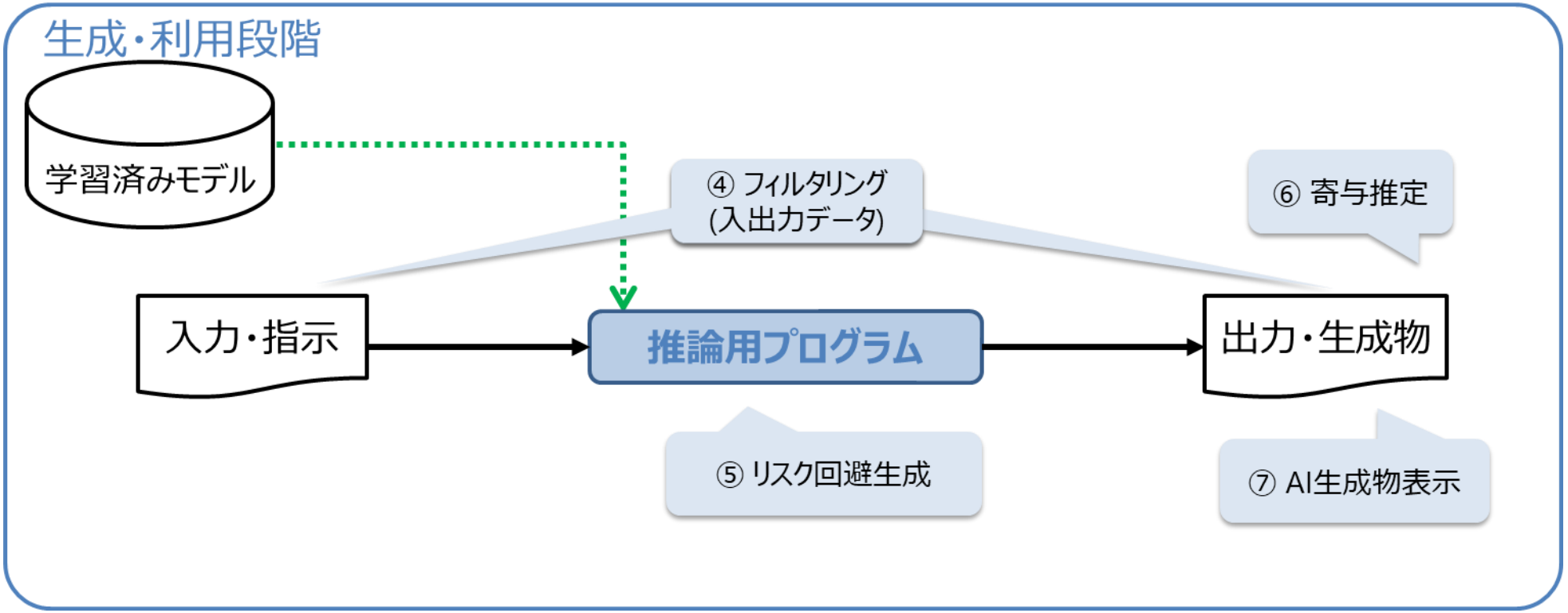
◆ アンラーニング

✓ **学習済みの生成AIモデルから「ものを忘れさせる」技術**

- 生成AIモデルの学習には巨大な計算資源と時間コストが必要となる
- 特定のデータや概念を生成AIモデルから効率よく削除するため、モデルの調整や再学習などを行う手法が研究されている
- 例えば、Microsoftの論文などでは、既存の大規模言語モデルLlama2-7bに対して、最初から再学習を行うことなくハリーポッターの関連情報を忘れさせることに成功している

3-3. 生成・利用段階に適用可能なリスク低減技術

✓ 生成・利用段階に適用可能なリスク低減技術を、以下の4つの区分に整理



3-3. 生成・利用段階に適用可能なリスク低減技術

④ フィルタリング(入力・出力データ)

- ✓ 生成AIモデルの入出力をフィルタリングすることで、著作権やプライバシー、倫理・公平性に関するリスクを低減
- ✓ フィルタリングの方法としては学習・開発段階における「フィルタリング(学習データ)」と共通
  - ➢ 除外リストフィルタ
  - ➢ 個人情報除外フィルタ
  - ➢ 倫理評価フィルタ

◆ 除外リストフィルタ

- ✓ **事前に、権利侵害などの恐れがある除外候補のデータ(=除外リスト)を用意しておき、除外リスト内のデータと類似しているデータを検出する技術**
  - ➢ 例えば、国立研究開発法人情報通信研究機構（NICT）はテキスト生成AIが学習データと類似するテキストを出力した場合に警告を出す、著作権侵害チェック支援ツールを開発している

3-3. 生成・利用段階に適用可能なリスク低減技術

⑤ リスク回避生成

- ✓ 生成AIモデルの使用方法を工夫することで、著作権やプライバシー、倫理・公平性に関するリスクを低減
  - ➢ プロンプトエンジニアリング

◆ プロンプトエンジニアリング

- ✓ **主にテキスト形式の入力データ(プロンプト)を受け付ける生成AIに対して、所望の出力を得られるように使用方法を工夫する技術**
  - ➢ 以下のようなさまざまな方法が検討および実施されている
    - – プロンプトへの介入:
      - » プロンプトの前後等に追加の指示文を追加
      - » プロンプト自体の修正・加工
    - – 実行方法の工夫:
      - » 同一の入力に対して複数回生成を実行し、一番よい結果を使用
      - » 実行して得られた生成物を生成AI自身に再度入力し、評価・修正を行うなど、繰り返し生成AIを使用

3-3. 生成・利用段階に適用可能なリスク低減技術

⑥ 寄与推定

- ✓ 生成AIによる生成物に対して、各学習データがどの程度寄与したか(生成物に影響を与えたか)を予測して、特定の学習データの寄与が大きいデータを出力しないなどの方法で権利侵害のリスクを低減
  - ➢ 寄与推定

◆ 寄与推定

- ✓ **生成物に与えた学習データの影響を予測する技術**
  - ➢ 元は分類モデルなどにおいて、推論過程の可視化・分析のために使用される手法
  - ➢ 著作権侵害の「依拠性」の判断根拠の1つとして、利用できる可能性がある
  - ➢ 生成AIへの応用については研究の初期段階
  - ➢ 例えば、ある学習データに対して、該当データを使用して学習した生成AI と 使用せずに学習した生成AI によって得られる生成物を比較することで、該当データを学習データに使用することの影響を評価するといった手法が存在する
    - – 実際には、生成AIを再学習するのには大きなコストがかかるため、生成AI自体を再学習するのではなく、学習データに使用しなかった場合に得られる生成物を予測することで評価を行う

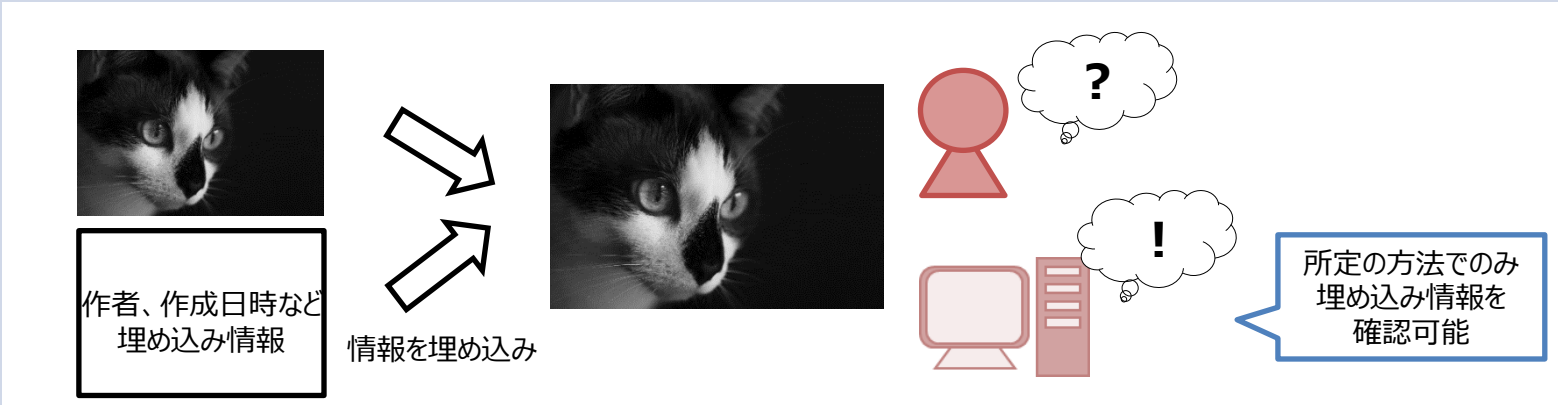
3-3. 生成・利用段階に適用可能なリスク低減技術

⑦ AI生成物表示

- ✓ 生成AIにより生成された生成物に対して情報を付与し、生成AI使用を明確にする
  - ➢ コンテンツ認証
  - ➢ 不可視的透かし
  - ➢ 可視的透かし/明示的表示

◆ 不可視的透かし

- ✓ **画像や音声などのデジタルデータに対して人間に知覚できない形でデータに関連した情報を埋め込んでおき、所定の方法に従うことでその情報を抽出することを可能にする技術**
  - ➢ 電子透かし、またはデジタルウォーターマークと呼ばれる
  - ➢ 技術自体は古くから研究され他の用途では実用化されているものも多い
  - ➢ 生成AIの生成物に対して適用することで、AIによる生成物であることの証明が可能となる



4. まとめ

まとめ

- ✓ 生成AI事業者が検討するべきリスクを調査し、
  - ➢ 権利侵害に関して法令へ抵触するリスク
  - ➢ 生成AIの生成物に対する社会的責任に関するリスク

  から6つのリスクを取り上げた

- ✓ 各リスクに対して生成AI事業者がおこなっている取り組みや研究事例について調査を行い、整理を行った

段階	リスク対策の区分	名称	対応リスク					
			学習・開発		生成・利用			
			著	プ	著	プ	倫	透
学習・開発	フィルタリング(学習データ)	除外リストフィルタ	△		○		○	
		類似物排除フィルタ	△		○			
		個人情報除外フィルタ		○		○		
		倫理評価フィルタ					○	
	リスク回避学習	学習データの複製生成抑制	△		○			
		人工データでの学習	△	○	○	○		
		差分プライバシー				○		
		Human in the Loop学習			○	○	○	
	学習済みモデルの修正	アンラーニング			○	○	○	
		アンバイアシング					○	

段階	リスク対策の区分	名称	対応リスク			
			生成・利用			
			著	プ	倫	透
生成・利用	フィルタリング(入力・出力データ)	除外リストフィルタ	○	○	○	
		個人情報除外フィルタ		○		
		倫理評価フィルタ			○	
	リスク回避生成	プロンプトエンジニアリング	○	○	○	
	寄与推定	寄与推定	○			
	AI生成物表示	コンテンツ認証				○
		不可視的透かし				○
		可視的透かし/明示的表示				○

まとめ

- ✓ 生成AIの急速な進展により、生成AIにおいてリスクを低減するための技術に関してはまだ検討が始まったばかりの状況である。**今後生成AIの発展とともに、リスク低減技術についても成熟が期待される**

- ✓ 生成AIの事業化の際には、最新のリスク低減技術の動向にも注意を払い、非技術的な対応も含めて既存の技術や対応を応用・活用しつつ、**複数の技術・対応を組み合わせてリスク低減を図る**ことが望ましいと考えられる

詳細な調査結果は後日NEDOホームページで公開予定