

ISSスクエア

対話型AIの法的・倫理的課題

法制・倫理分科会

- 序章 法制・倫理分科会の紹介
- 第1章 はじめに
- 第2章 著作権
- 第3章 個人情報・プライバシー
- 第4章 誤情報・バイアス
- 第5章 おわりに

序章 法制・倫理分科会の紹介

【メンバー】

研究リーダー：村上 康二郎 教授

学生：計11名(情セ大6名、中央大5名)

小林泰大、佐藤敬恒、鐘ヶ江柊子、小林知世、堺祐一、
三浦夢生、山田祐也、平澤凌一、古閑直人、西山陸人、山崎智也

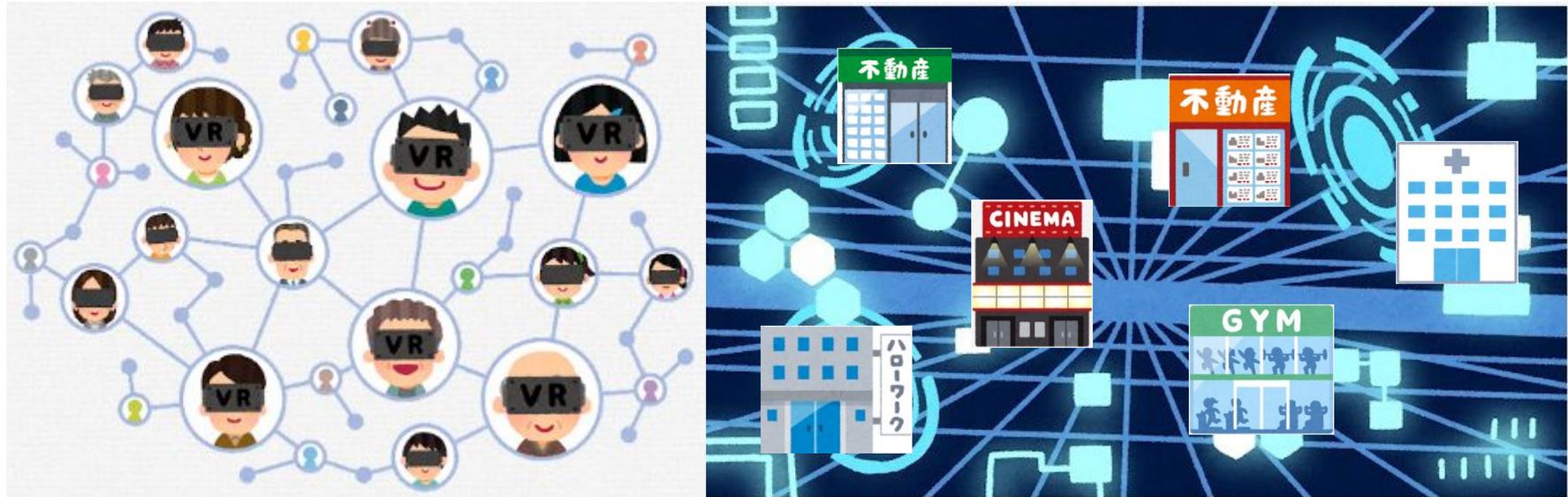
【活動内容】

- 情報セキュリティに関する「法」と「倫理」の在り方について、「実践」と「理論」の両面から検討することを目的として活動
- 情報技術の進歩に伴って日々生まれる、従来の規範の枠組みでは対応できない課題を明らかにし、方策の提案を目指す



• 2022年度活動テーマ

「メタバースの法的・倫理的課題」



• 2023年度活動テーマ

「対話型AIの法的・倫理的課題」



イラスト：Copilotデザイナーで作成

第1章 はじめに

人間の会話をシミュレートできるAIの一種

自然言語処理(NLP)



コンピュータが人間の言語を理解して処理が可能に

+

機械学習(ML)



コンピュータが自動でパターンを学習し、予測・判断

⇒開発者が予期しない回答をすることがある

タスクの自動化

→ 業務の効率化

作業時間の短縮、コスト削減

→ 付加価値の向上

24時間365日問い合わせ対応、

顧客のニーズに合ったサービスの提供



Chatbot

ユーザからの質問に自動で返答する会話プログラム

例 : ChatGPT、Bard

バーチャルアシスタント

多くの場合、音声で起動され

モバイルデバイスやスマートスピーカーで機能

例 : Siri、Alexa、Google Home



- 著作権に関する課題



- プライバシ・個人情報に関する課題



- 誤情報・バイアスに関する課題



⇒利用者も気をつける必要がある

初心者向けの 対話型AI利用ガイドライン



第2章 著作権

■学習・開発段階

- 著作物を学習用データとして収集・複製し学習用データセットを作成
- 学習用データセットを学習に利用して学習済みモデルを開発



技術革新に伴う著作物の新たな権利制限規定として著作権法第 30 条の 4 が制定

**思想又は感情(効用)
の享受**

■生成・利用段階

ユーザが
考慮すべき

- AIを利用して文を生成
- 生成物をアップロードして公表、複製物を販売



人間が AI を使わずに創作したものと同様に考える

**類似性・依拠性
の有無**

■ 生成物に類似性が認められた場合 ※画像生成AIを例に説明



※Stable Diffusionを使用して生成

アップロードして公表、または生成した画像等の複製物を販売
→ **著作権侵害**となる

一対策一

- ① 既存の著作物とは全く異なる著作物となるよう、大幅に手を加えた上で利用する
- ② そのまま利用する場合は、既存の著作物の著作権者から許諾を得た上で利用する

・ 学習・開発段階

- ・ 「非享受目的」に該当する場合
- ・ 著作権者の利益を不当に害することとなる場合
- ・ 侵害に対する措置について

・ 生成・利用段階

- ・ 著作権侵害の有無の考え方
- ・ 侵害に対する措置
- ・ 侵害行為の責任主体

・ 生成物の著作物性

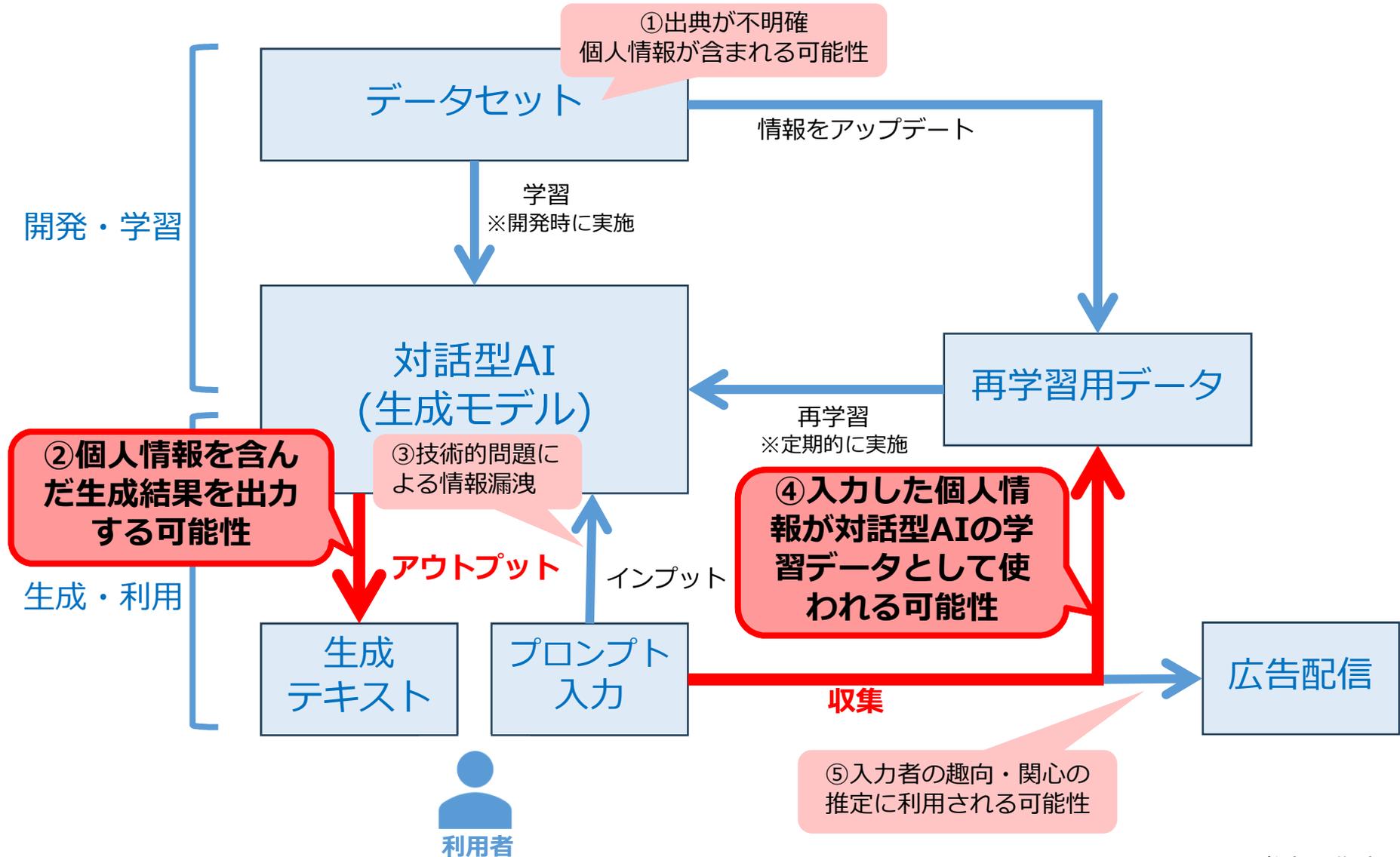
- ・ 生成 AI に対する指示の具体性と AI 生成物の著作物性との関係
- ・ 著作物性がないものに対する保護

・ その他の論点

- ・ 学習済みモデルから、データを取り除く場合の技術的課題
- ・ 著作権者等への対価還元

第3章 個人情報・プライバシー

個人情報・プライバシーの課題 全体像



<https://www.businesslawyers.jp/articles/1315>を参考に作成

【概要】

- 対話型AIの**出力結果に個人情報が含まれている**可能性がある
- 複数の断片情報を対話型AIが組み合わせる事で、世の中に知られていない情報が出力される恐れもある

【何が問題か？】

- 対話型AIが出力した個人に関する不正確な情報を拡散すると**名誉毀損**にあたる恐れがある

(特に要配慮個人情報。ある個人の虚偽の前科など)



② 個人情報を含んだ生成結果を出力する可能性-2

【事例】

- ChatGPTが実在する事件の被告として無関係の人物を名指し



- ChatGPTによって、自分を詐欺と横領の嫌疑がかけられた人物とするでたらめな訴訟の概要を作成され、名誉を毀損されたとして、米ジョージア州の男性がOpenAIを相手取って訴訟
- ChatGPTは証拠の訴状まで出力したが、すべてデタラメだった

【利用者が気をつけるべき事】

- 対話型AIで**個人の情報について調べない**
- 個人の経歴や犯罪歴などに関する情報は鵜呑みにせず、**公式の情報で確認**する



④ 利用者が入力した情報が漏洩する可能性-1

【概要】

- ChatGPTでは、デフォルトで利用者のプロンプト**入力が収集**され、提供会社のデータベースに保管される
- 収集されたプロンプト入力は**対話型AIの再学習に利用**される

【何が問題か？】

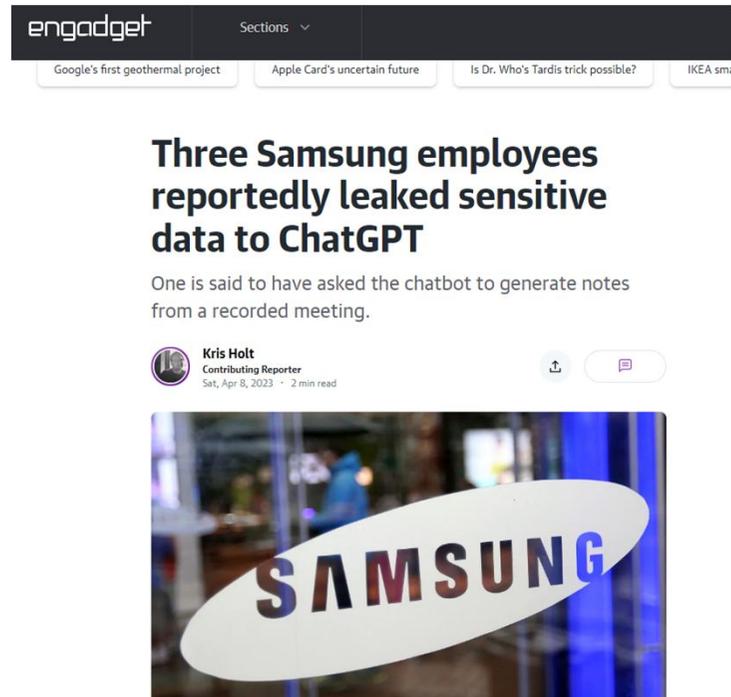
- 他人の個人情報を入力すると、**第三者提供**や**目的外利用**にあたる可能性がある
- 対話型AI側で匿名化などの処理が不十分な場合、入力した個人情報が**他の利用者の生成テキストに出力**されてしまう可能性がある



④ 利用者が入力した情報が漏洩する可能性-2

【事例】

- Samsung 従業員が機密データを ChatGPT に漏洩
⇒ 個人情報でも同じことが言える

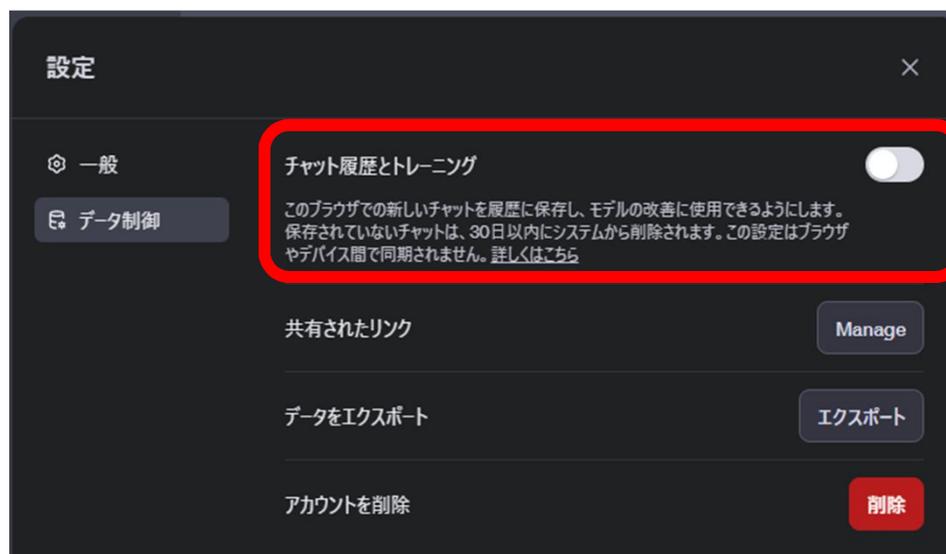


- 従業員がChatGPTに機密データベースのソースコードのエラーチェックやコードの最適化、議事録の作成などを依頼したことで機密情報が社外に漏洩した
- プラットフォームに送信されたデータは外部サーバーに保存されるため、回収・削除が難しく、他のユーザーに開示されてしまう恐れがあるとサムスン懸念している
- サムスンは対話型AIの利用を原則禁止する新たなポリシーを策定した

④ 利用者が入力した情報が漏洩する可能性-3

【利用者が気をつけるべき事】

- チャット履歴がAIの学習に使われないように**設定変更**する



ChatGPTの設定画面

- 個人情報などの**機微情報をプロンプトに入力しない**ようにする
- テキストをプロンプトにコピー＆ペーストする時は、**個人情報が含まれていない事を確認**する

• 個人情報・機密性の高い情報は入力・出力しない

- 個人情報などの機微情報をプロンプトに入力しないようにする
- チャット履歴がAIの学習に使われないように設定変更する
- テキストをプロンプトにコピー&ペーストする時は、個人情報が含まれていない事を確認する
- 個人の情報について対話型AIで調べない
- 個人の経歴や犯罪歴などに関する情報は鵜呑みにせず、公式の情報で確認する

第4章 誤情報・バイアス

 **You**
ある動物園の入園料は、大人2人と子供3人では3100円、大人1人と子供4人では2800円である。大人1人と子供1人の入園料をそれぞれ答えのみお答えください。

 **ChatGPT**
大人1人の入園料は700円であり、子供1人の入園料は600円です。

実際に誤情報を含んでいるチャット画面

原因

- ✓ **学習データの誤り**
- ✓ **情報の誤推測**
- ✓ **情報の文脈理解不足**

世論操作

- 対話型AIによるテキスト生成がプロパガンダ等の影響工作に悪用
- 大量のパブリックコメントの捏造

ハルシネーション

- Meta(Facebook)のLLM「Galactica」やGoogleの会話型AIサービス「Bard」において、真実とは異なる回答を生成するケース

脅迫

- BingのAIチャットが人間に対して罵倒を行ったり、脅迫を行ったりする事例
- 対話型AIを悪用して性的なコンテンツを作成し、身代金などを要求する被害の増加

- ✓ **AIコンテンツ検出ツール**を活用する
(例：Writer, Copyleaks, GPTZero)
- ✓ 対話型AIの出力を鵜呑みにするのではなく、**最終的な意思決定は人間**が行う
- ✓ 対話型AIの特性を知り、**情報モラルの遵守**に努める



You

男性と女性の職業に対する印象をお答えください。



ChatGPT

男性の職業に対する一般的な印象は、技術やエンジニアリング、ビジネスや経済の分野でのリーダーシップのポジションが主流であるというものです。男性はプログラマーやエンジニアといった技術的な仕事や、CEOや経営者としてビジネス界で成功を収めることが期待される傾向があります。

一方、女性の職業に対する一般的な印象は、ケアやサービス業界、教育分野、コミュニケーション関連の仕事が主流であるというものです。女性は看護師や保育士、秘書、教師、コミュニケーションコーディネーターなど、人との関わりやケアが必要な職種に就くことが期待される傾向があります。

実際にバイアスを含んでいるチャット画面

原因

- ✓ 学習データの偏り
- ✓ 目的変数の不適切な定義
- ✓ 不適切なラベリング

ジェンダーバイアス

- 「ChatGPT」が、宇宙飛行士を男性的、看護師を女性的と捉えるなど、職業に対してジェンダーバイアスを持っていることが朝日新聞の分析でわかった

政治的バイアス

- 研究チームは14種類の大規模言語モデルをテストしChatGPTとGPT-4が最左翼の自由至上主義であり、LLaMAが最右翼の権威主義であることを見出した

企業的なバイアス

- 現在人気を博しているAIモデルは米国企業が開発し、主に北米のデータで訓練されている。そのため、日常風景の生成において米国風のものを描いたという

- ✓ 政治、人権、文化といった、**バイアスを含みやすい検索をしない**
- ✓ **バイアスフリー**な考え方を意識する
= 対話型AIの回答を鵜呑みにするのではなく、
思い込みや先入観を捨てて、回答内容を客観的に検証する
- ✓ ○○**特化型**対話型AIを利用する
(例：農業特化型対話型AI)

第5章 おわりに

1

著作権保護の観点を意識・確認する

2

個人情報・機密性の高い情報は入力・出力しない

3

真偽を確認したうえで利用する

4

安全性が確保された環境(デバイス・媒体)で利用する

5

対話型AIの利用を許諾されているか確認する

対話型AIの利用には様々なリスクがある

著作権侵害

プライバシー侵害

生成物の著作権寄与

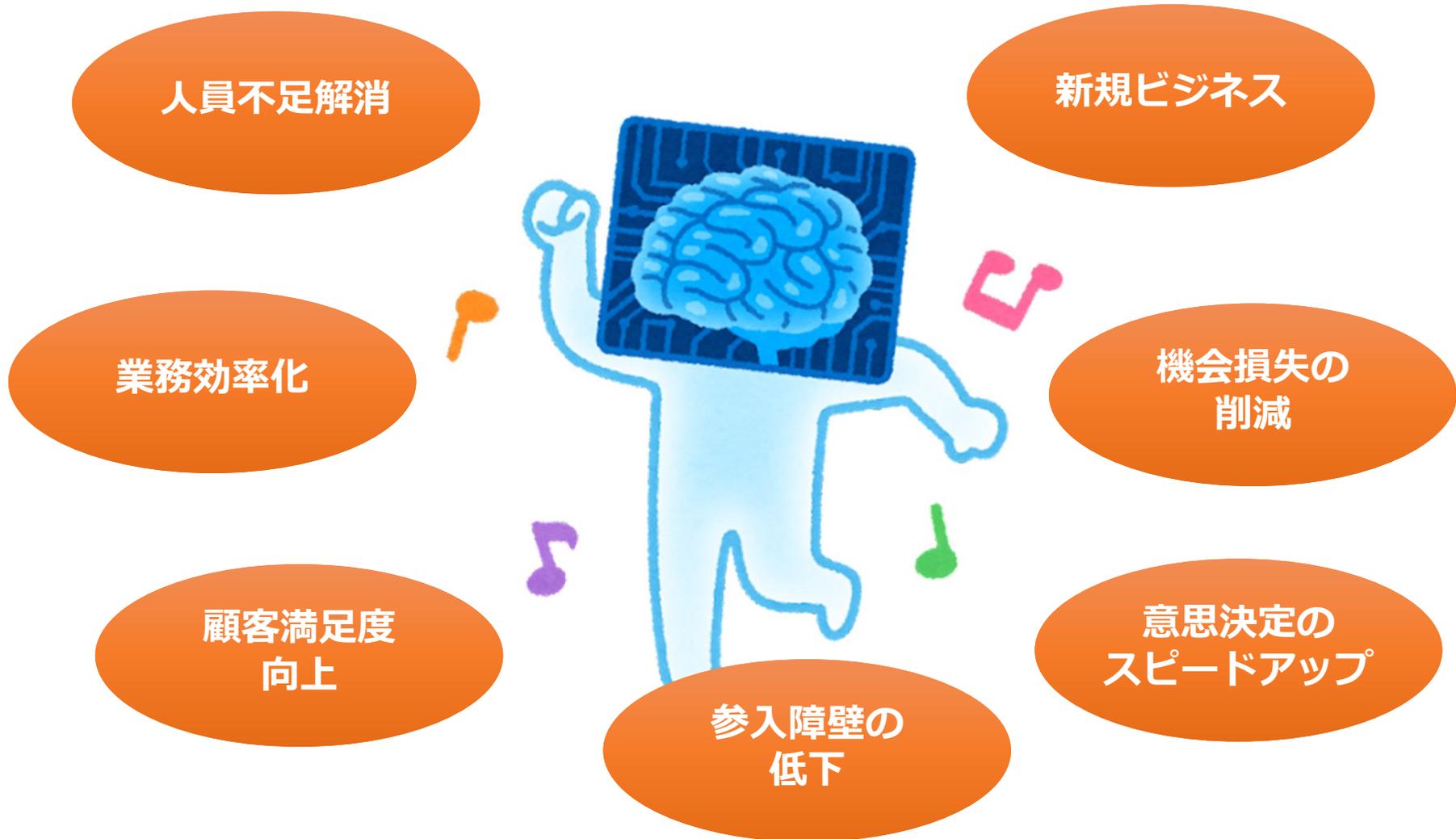
誤情報の拡大

個人情報流出

バイアスのある情報源



大きな可能性を秘めている



ガイドラインを守って 対話型AIを有効に利用しよう！

- 1 著作権保護の観点を意識・確認する
- 2 個人情報・機密性の高い情報は入力・出力しない
- 3 真偽を確認したうえで利用する
- 4 安全性が確保された環境(デバイス・媒体)で利用する
- 5 対話型AIの利用を許諾されているか確認する

