



AI倫理とAIガバナンス

作成者：辻 大貴

AI倫理とAIガバナンス

・法と倫理

- AIを活用していく上で法の遵守・倫理の遵守が重要である
→法律を遵守することは当然であり、ある程度理解しやすい
- 倫理は、地域・時代・文化などによって大きく変わるため、
環境に合わせた対応が必要になるため、理解が難しい
- 人によって倫理が変わることも多い

AI倫理とAIガバナンス

・価値原則

AI開発などを適切に行っていくためには指針が重要になる

→指針は、人々の様々な価値観のもと原則として定められていく

→人の価値観が変われば、原則（指針）を変えていく必要がある

→ガイドラインは指針の1つである

AI倫理とAIガバナンス

・ソフトローとハードロー

- ガイドラインや自主規制などのように法的な強制力が弱い規則のことをソフトローという
→法律などのように法的な拘束力のある規則のことをハードロー
- ソフトローは、ハードローより柔軟に変更しやすい特徴がある
AIに関する規則はソフトローによるものが多い

| AI倫理とAIガバナンス

・ソフトローとハードロー

EUはソフトロー、 ハードローでAIに関して規則を設ける方向で議論が進んでいる

→日本やアメリカなどは、 基本的にはソフトローでAIに関して規則を設ける方向で議論が進んでいる

| AI倫理とAIガバナンス

・AIガバナンス

AIの開発などを、一定の法律や倫理などの規則に基づいて、
管理・統治していく仕組み・考え方のこと

- 規則を遵守していない場合、規則がないのと同じである
どのように規則を遵守させるのか、**仕組み**を作ることが大切
- 第三者機関によるAIに対する監査などもAIガバナンスの1つ

AI倫理とAIガバナンス

・リスクベースアプローチ

組織や運用しているAIによってリスクは異なっている

→画一的な規則を全ての組織に適応することは**不合理**である

ゲームで使用するAI、医療用のAIではリスクが異なる

→組織などのリスクに合わせて、適用する規則・対応を変更する

アプローチ（リスクベースアプローチ）を取ることが**合理的**



国内外のガイドライン

| 国内外のガイドライン

- ・日本におけるガイドライン例
 - ・国際的な議論のためのAI開発ガイドライン案（総務省）
 - ・AI利活用ガイドライン（総務省）
 - ・AI原則実践のためのガバナンス・ガイドライン（経済産業省）
 - ・人間中心のAI社会原則（内閣府）

| 国内外のガイドライン

- EUにおけるガイドライン例

- 信頼性を備えたAIのための倫理ガイドライン

欧州委員会がAIに関する倫理ガイドラインとして発表(2019)

- AI法

2024年に可決した人工知能の規則に関する法律

→2021年に欧州委員会によって提案された法案である

| 国内外のガイドライン

- ・アメリカにおけるガイドライン例

- ・AI Risk Management Framework

米国国立標準技術研究所(NIST)が公表

→商務省（日本における経済産業省）の傘下の研究所である

- ・Blueprint for an AI Bill of Rights

ホワイトハウスが公表、Bill of Rightsは権利章典という意味

| 国内外のガイドライン

- OECDにおけるガイドライン例

- OECD（経済協力開発機構）

ヨーロッパ諸国を中心とした先進国が加盟する国際機関

→日本、アメリカなども加盟国である

- AIに関する理事会勧告

OECDが公表したAIに関するガイドライン

| 国内外のガイドライン

- ・ その他のガイドライン例

- ・ アシロマAI原則

NPO法人である**Future of Life Institute**(生命の未来研究所)が
公表した倫理に関する**23の原則**をまとめたガイドライン

- ・ 倫理的に調和された設計

IEEE (米国電気電子学会) が公表

| 国内外のガイドライン

- ・ その他のガイドライン例

- ・ 信条

AmazonやGoogleなどアメリカのIT企業を中心に組織された
Partnership on AIが公表したAIに関するガイドライン

→安全にAIが開発・研究されるように、
世界中でAI開発に関する議論が行われている

プライバシー



| プライバシー

・プライバシー権

他人から干渉を受けない権利、情報を開示されない権利

→最近では、自分の情報をコントロールする権利も含まれる

→ネットやスマホの普及により、多くの情報（画像・動画など）が

ネット上で公開され、プライバシー権を侵害するものもある

→プライバシー権を侵害しないようにAIを開発することが重要

| プライバシー

・データ収集

学習目的、分析目的などとして多くのデータが収集されている
→プライバシーに関するデータも収集されている可能性が高い
炎上につながる可能性があるため、慎重に対応する必要

→事前にどのようなデータを収集しているのかを伝えたり、
プライバシーが保護されるようにデータを加工したりする

プライバシー

・データ収集

名前	年齢	購入商品
山田 太郎	24	商品A



加工

名前	年齢	購入商品
＊＊＊	20代	商品A

| プライバシー

- ・カメラ画像利活用ガイドブック

カメラの低価格により、様々な場所にカメラが設置されている
→カメラの高精度化により、個人を特定することも可能

→経済産業省・総務省は、配慮すべきポイントなどが記載された
　　カメラ画像利活用ガイドブックを策定し、公表した

| プライバシー

- 推論

推論により、プライバシー権が侵害されてしまうことがある

- 様々な情報から知られたくない情報を推論されてしまったり、
生成AIにより誤った情報が提供されてしまったり、
自分の声や画像などが詐欺で使用されてしまうことがある
- 自分の情報をコントロールする権利が大きく侵害されている

| プライバシー

・プライバシー・バイ・デザイン

システムの設計段階からプライバシーに配慮する考え方のこと
→あらかじめプライバシーに配慮したシステムを開発することで
利用者から信頼を得やすく、コスト面でメリットが多い

→プライバシーに配慮したシステムを途中から作ろうとすると
作り直しの可能性があり、コストがかかることが多い

公平性



| 公平性

・公平性・説明責任・透明性

AIを提供する企業は、人種、宗教などによって差別されない、
公平で透明性のある結果を出力するモデルを提供する責任と
その結果を説明する責任が求められている

→米国コンピューター学会（ACM）が主催する「ACM FAT」が
FATに関する会議として有名である

| 公平性

- 公平性・説明責任・透明性

AIを活用していく上で公平性・説明責任・透明性についても議論
→公平性、説明責任、透明性をFATと表現する

- 公平性：Fairness
- 説明責任：Accountability
- 透明性：Transparency

| 公平性

・公平性の定義

AIには公平性が求められているが、何をもって公平と
考えるかによってAIに求められるものは変わる
→自分に似合う化粧品を見つけるためのAIを開発する場合、
大量の女性のデータは収集するが、男性のデータが少ないと
AIが男性に対しては適切な結果を返すことができないとき、
これは公平性に欠けるのではないか など

| 公平性

・公平性の定義

公平性を評価する基準が変わればAIの評価も変わってしまう
→人には様々な認知バイアス（先入観・偏見）が存在するため、
AIに歪みが生じてしまうことがある
→観察者バイアスも認知バイアスによるものが大きい

| 公平性

・観察者バイアス

観察者が期待する結果を求めているとき、
分析結果を期待する結果に合うように解釈したり、
求めていない結果が出たときはその結果を
軽視したりしてしまうことで発生するバイアスのこと

→期待した結果がでなかつたとき特別な値だと解釈する など

| 公平性

- データの偏り

抽出したデータに偏りがあるサンプリングバイアスや
偏ったデータを学習したこと、モデルが偏った結果を
出力してしまうアルゴリズムバイアスなどがある

- サンプリングバイアスの例

→男子高校生の体重を調べたいとき、相撲部所属の学生を多く抽出

| 公平性

- ・アルゴリズムバイアスの例

現場には男性が多く、男性の方が好ましいとAIが学習したことで
履歴書評価AIが女性に対して不利な結果が表示してしまう

→特定の職業の人が平均的に年収が低いことからローン審査AIが、
特定の職業の人に対して、低い信用度を与えててしまう など

| 公平性

- ・センシティブ属性

公平性の観点から、性別、国籍、職業など

繊細に扱った方が良いと考えられる属性のこと

→学習データからセンシティブ属性を排除するのも1つ

→病名の診断AIの場合などは性別データは重要なデータである

状況によってセンシティブ属性の扱い方を変えていく必要がある

| 公平性

・代理変数

- 他の変数と関連性が高い変数で、他の変数の推定が可能な変数
→性別をデータから削除したとしても、
身長のデータなどから性別を推定できてしまう
- 間接的にセンシティブ属性を学習してしまう
公平性の観点から、代理変数について考慮していく必要がある

| 公平性

・開発後の公平性

公開時点では偏りがないAIを開発したとしても、

公開後、学習するについて偏りのあるAIになる可能性がある

→Microsoftが開発した会話ボットTayが、複数ユーザーによる

不適切な学習により、差別的な発言を投稿したことで問題

→モデルを修正するときもデータの偏りに気をつける必要がある

說明責任・透明性



説明責任・透明性

・公平性・説明責任・透明性

AIを活用していく上で公平性・説明責任・透明性についても議論
→公平性、説明責任、透明性をFATと表現する

- ・公平性：Fairness
- ・説明責任：Accountability
- ・透明性：Transparency

説明責任・透明性

・説明責任

AI開発者側には、開発の流れ、パラメータ、出力結果などを利害関係者に可能な限り説明する責任が求められている
→バイアスが発生しないようにどのように対策を講じたのか
実際にアルゴリズムにバイアスが発生していないのか など
→全ての情報を説明することは技術的、ビジネス的に不可能

説明責任・透明性

・説明責任

AIがどのようなプロセスで結果を出力しているのかを
説明することも求められている

→説明可能なAI(XAI)の研究が進んでいる(説明可能性の研究)

→XAIの研究は、アメリカ国防総省の機関である

DARPA(アメリカ国防高等研究計画局)が発端と言われている

説明責任・透明性

・透明性

開示できる情報は、適切に利害関係者に開示し、

透明性を高めていくことが求められている

→AIで使用されているアルゴリム、評価方法、評価の結果 など

→学習データなどをどのように入手し、加工したかなどの

データの来歴を開示することも透明性を高める上で大切

説明責任・透明性

- 透明性レポートの公開

利用者のデータなどをどのように活用しているのか、

政府などに情報提供した件数などをまとめたレポート

→GoogleやAmazonなど多くの個人情報を扱う企業が

Webなどで公開し、適切に個人情報を扱っていることを説明

→プライバシー対策やセキュリティ対策についても公開している

安全性とセキュリティ



| 安全性とセキュリティ

・ 安全性

- AI開発・提供するときは**安全性**についても考慮する必要がある
→自動運転車など人に危害を与える可能性が高い
- AIについては**品質管理**を徹底的に行うことが求められる
- リスクをゼロにすることは不可能であるため、
問題が起きたときの対処法について考えておくことも必要

| 安全性とセキュリティ

・セキュリティ

モデルを開発し、実際に運用が始まると様々な攻撃を受ける
→ハッキングされ、システムを止められる、データを消される、
個人情報が盗まれる、モデルが書き換えられる など

→開発時点で想定していなかった攻撃を受けることもある
セキュリティを高め、悪意のある攻撃を防ぐことが求められる

| 安全性とセキュリティ

・想定外の攻撃と対策

アクセスを制限し、開発者以外がモデルを
書き換えることができないようにする（不正アクセスへの対応）
→システムにログインするパスワードなどを複雑にする

→当たり前のことを行っていくことが
セキュリティ対策では一番大切である

| 安全性とセキュリティ

- AIに関する攻撃
 - 敵対的な攻撃
 - データ汚染攻撃
 - モデル汚染攻撃
 - データ窃取攻撃
 - モデル窃取攻撃

| 安全性とセキュリティ

・敵対的な攻撃

入力する画像などに人間が認識できない程度のノイズを加えて、
モデルに意図的に誤った出力をさせるように仕掛ける攻撃のこと
→自動運転車のカメラにノイズシールを貼る など

→敵対的な例を学習データに混ぜてモデルを学習（敵対的学习）
させることでモデルの頑健性を高めることができる

| 安全性とセキュリティ

・データ汚染攻撃

学習データにノイズを加えたデータや誤ったデータを混入させ、
モデルの性能を低下させる攻撃のこと

→誤ったデータを大量に与えることで誤情報を出力させる など

→信頼できる学習データの利用、

検知システムなどを活用して汚染データを事前に排除する

| 安全性とセキュリティ

・ モデル汚染攻撃

攻撃者が悪意のあるモデルを作成し、配布し、
他者にモデルを使用させ、問題を起こさせる攻撃のこと
→モデルを実行するとデータを破壊するコードを生成したり、
誤情報に基づく情報を出力させたりすることが可能
→信頼できる事前学習モデルを使用する

| 安全性とセキュリティ

・データ窃取攻撃

モデルから出力されるラベル・確信度と入力データを分析することで、近似した学習データを推測すること
→過学習モデルの場合、学習データに近い値を入力すれば、出力される確信度が極めて高い値になってしまう
→過学習の抑制、ラベルのみ出力、アクセス制限 などが対策

| 安全性とセキュリティ

・ モデル窃取攻撃

モデルから出力されるラベル・確信度と入力データを分析することで、モデルのパラメータの推測を行う攻撃のこと
→似たモデルを作成されてしまい、
模倣サービスを展開されてしまう可能性がある
→アクセス制限、模倣モデルを早期に見つけ対応 など

| 安全性とセキュリティ

・セキュリティ・バイ・デザイン

システムの設計段階からセキュリティに配慮する考え方のこと
→開発段階で高めるようとすると作り直しの可能性もある

→途中からセキュリティを高めることはコストがかかる
最初からセキュリティを高めることを念頭に
設計をしていくことが社会的に、コスト的にメリットが大きい

悪用と民主主義



| 悪用と民主主義

・ 悪用

- 開発したモデルが想定外の方法で使用されることがある
→フェイクニュースを作るために自社のサービスが利用されるなど
- ディープフェイクはディープラーニング技術を悪用した例である
- ディープラーニング技術を使用して、画像や動画に
別の顔や声を合成したりして**偽の動画・画像**などを作り出すこと

| 悪用と民主主義

・ディープフェイク

大統領の顔・声で過激な発言をしている

動画などが作られたりするなど社会問題にもなっている

→技術力の向上により本物と偽物を見極めるのが困難になっている

→アニメや漫画のキャラクターが無断で使用されて、

生成AIによって作られた作品が販売される

| 悪用と民主主義

- ・ディープフェイク

悪用されていることを確認したら、悪用できないように
対策を講じていく必要がある（運営者としての責任）

→運営者が悪用している事業者等を発見した場合、
その事業者等へのサービスを停止する（利用規約）

| 悪用と民主主義

- ・ディープフェイクへの対抗策
 - ・偽情報を見極めるAIの開発・提供
- ・生成コンテンツの表示

→YouTubeでは生成コンテンツ（一部）を投稿する場合は、
生成コンテンツであることを表示する義務がある

| 悪用と民主主義

・ 民主主義

AIを悪用することで、人民の選挙行動などを操作できる可能性があり、**民主主義の根幹**を揺るがす問題である
→AI開発者が人民の選挙行動を操作し、特定の党に誘導するなど
→政治家の顔・声で過激な発言をしている動画を生成することで
その政治家に票を入れさせないようにすることが可能である

| 悪用と民主主義

・ 民主主義

AIが利用者の趣味嗜好を学習することで、
支持している政党の情報しか入ってこないという問題が発生する
→異なる意見に触れる機会が少なくなり、視野が狭くなる可能性
様々な意見を知り、自分で考え、行動することが大切
→フィルターバブル現象：趣味嗜好に合った情報のみ表示される現象

| 悪用と民主主義

・ 民主主義

自分と似た意見を持つ人が集まる場所に行ったり、
SNSなどで自分の意見に近い人をフォローしたりすることで、
自分の意見や思想が増幅する現象をエコチェンバー現象という

→エコチェンバーは自分と似た人をフォローすることで発生する
フィルターバブル現象はアルゴリズムの影響によって起きる



環境保全と労働政策

環境保全と労働政策

・AIと電力

モデルを学習させたり、AIを使用したりするとき、
大量の計算が必要になるため、多くの電力を消費してしまう
→GPT-3の学習に必要な電力量は約1,300MWhと言われている

→原子力発電1基の約1.5時間分の電力量である
3~6万世帯の1日分の電力に相当する

環境保全と労働政策

・AIと水

計算を行う過程で多くの熱が発生してしまう

冷却のために大量に水を消費していると言われている

→一般的に高度なAIほど使用電力量や水は増えるため、

気候変動に何かしらの影響を与えると考えられる

→環境面から省電力のAIを開発していくことが求められる

| 環境保全と労働政策

• AIとの協働

人とAIが協力して、問題などを解決していくこと

→AIの高精度化により、**特定の分野**では人よりも
高い精度でタスクをこなすことができるようになっている

→特定のタスクは人からAIが行っていくと予想される

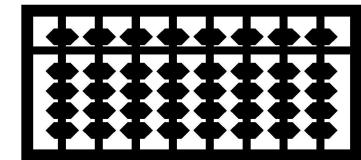
人類史では似たようなことは何度もあった（産業革命など）

環境保全と労働政策

・AIとの協働

スキルが必要だったタスクがAIに変わることで、

スキルの喪失が考えられる



→企業に電卓などが導入される前などは、

そろばんを使用して、経理処理を行っていた

→電卓が登場したことで、そろばんスキルを持った人が減った

環境保全と労働政策

・AIとの協働

現状は仕事には多くのタスクがあるため、
数年間で完全にAIに代替されることはないと考えられている

- 今まで時間がかかっていたタスクをAIに置き換えることで、
業務効率が高まることが期待されている
- 労働者不足**の解消・緩和に繋がると期待されている



その他の重要な価値

| その他の重要な価値

- ・ダイバーシティ（多様性）

人々が持つ様々な違い（性別、年齢など）を認め尊重すること
→ダイバーシティを無視したAIは炎上リスクがある

- ・インクルージョン（包括性）

ダイバーシティを活かせるような環境・状態のこと
→AIにおいても様々な人々が利用できるように改良し続ける

| その他の重要な価値

- 有事に対応する

世界中でAIと安全保障・軍事技術に関する議論がなされている
→2019年、非人道的な兵器を規制する
特定通常兵器使用禁止制限条約（CCW）の締約国は、
兵器による攻撃の判断には人間が
関与しなければならないとする国際指針を採択した

| その他の重要な価値

- CCW（特定通常兵器使用禁止制限条約）

過剰な傷害または無差別の効果を発生させると
認定される通常兵器の使用を禁止または
制限する多国間条約のことである

→自律型兵器（AWS）や自律型致死兵器システム（LAWS）の
研究開発を禁止すべきだと世界中で議論されている

| その他の重要な価値

- **自律型兵器 (AWS)**

人間の介在なしで、人工知能などが自動的に攻撃対象を決めて攻撃する兵器のこと

- **自律型致死兵器システム (LAWS)**

殺傷能力等を有する自律型兵器 (AWS) のこと

| その他の重要な価値

• 自律型兵器 (AWS)

自律型兵器の開発が報じられると問題になる

→2018年に韓国の国立大学**KAIST**（韓国科学技術院）は
自律型兵器などを含む研究を行っていると報じられた
世界中のAI研究者から「**KAIST**が自律型兵器の研究を行う限り、
KAISTとの協同研究を行わない」という宣言が出された
→**KAIST**は自律型兵器の研究を行う予定ないと表明した

| その他の重要な価値

・死者への敬意

死者の音声データをもとに合成音声で音声データを出力したり、動画データを生成したりすることは、倫理的に許される行動なのかを考えていく必要がある

→死者に対する人権（プライバシー権など）をどの程度守っていくのかという議論を行っていく必要がある

| その他の重要な価値

・人間の自律性

AIによって自分にとって興味がある情報が流れていく場合、他人の干渉を受けずに、**自己の判断で行動できているのか**
→AIによって思考・行動を操作させられていると考えられる

→検索した情報が嘘（フェイク）の可能性もある
嘘の情報を鵜呑みにした場合、**選択する行動は変化する**

AIガバナンス



| AIガバナンス

- AIガバナンス

AIの開発などを、一定の法律や倫理などの規則に基づいて、
管理・統治していく仕組み・考え方のこと

- AIポリシー

企業などが策定したAIの開発や利用に関する指針のこと
→社内・社外のステークホルダーに指針を伝えることが大切

| AIガバナンス

・倫理アセスメント

プロジェクトなどが倫理的な観点から適切か評価するプロセス
→チェックリストなどを使用して網羅的に評価していく

→定期的に倫理的な観点から適切か評価するプロセスを
実施することで、プライバシー侵害、公平性の欠如などの
早期発見に繋がり、適宜修正していくことができる

| AIガバナンス

・責任の所在

- 責任の所在を決め、適切にAIプロジェクトを行うことが重要
→責任の所在を決めずにAIを開発すると無責任な開発になりやすい
誰かが確認しているだろうと考えて開発を進めてしまう
- 社内組織や第三者機関によるAIに対する監査を強化し、
AIガバナンスを高めていくことが求められる

| AIガバナンス

・人間の関与

重要度が高いものについては、人間の関与も大切になってくる
→AIの出力結果を参考にして、人が物事を判断していく など

→定期的にモニタリングを行い、適切に推論等が
行われているのかどうかを確認していくことも大切である
→不適切な推論等が行われているときはAIを停止するなど

| AIガバナンス

- **再現性**

同じ入力をしたら、似たような出力になることが求められている
→入力に対して、大きく異なる出力がなされているAIは
一定の精度を担保することが難しい

→1回目は精度が良かったが、2回目は精度が悪かった など

| AIガバナンス

・トレーサビリティと文書化

不測の事態に備えて、学習データ、使用しているモデルなど
追跡できることが求められている

→バイアス等を発見しても、追跡できなければ原因が分からない

→適切に追跡できるようにするためには、

日常的にAI開発のプロセス等を文書化していく必要がある



クライシス・マネジメント

| クライシス・マネジメント

・トラブルの予防と対策

トラブルが起きないように、トラブルが発生しても
被害が最小になるように体制を整えることが大切
→トラブルが起きやすい箇所、炎上しやすい箇所を学ぶことで
事前にトラブル・炎上などを回避することが可能
→危機管理やリスク管理について解説をしていく

| クライシス・マネジメント

- ・ クライシス・マネジメント（危機管理）

危機は発生するという前提に基づいて、危機が発生したときに被害を最小限に抑えるための対応について管理すること

- ・ リスク・マネジメント（リスク管理）

問題が起きないように事前に策を講じること

| クライシス・マネジメント

- コーポレートガバナンス（企業統治）

企業経営が適切に行われているか監視する仕組みのこと
→社外取締役、委員会などを設置して経営を監視する

- 内部統制

経営目標を達成するために社内で守るべきルール・仕組み
→社内ルールを設置することで不祥事などの発生を抑える

| クライシス・マネジメント

- 内部統制の更新

目標達成のために、不祥事が起きにくくするために
ルール・仕組みを更新していくことが大切

- 内部統制の注意点

ルールや仕組みを現場の人々に理解してもらう必要
→ルール・仕組みが整っていても守らないと意味がない

| クライシス・マネジメント

・炎上対策とダイバーシティ

データセットの偏りなどから炎上することがある

→性別、人種などに偏りがあると炎上リスクが高くなる

→炎上しないためには、ダイバーシティ（多様性）、

インクルージョン（包括性）を理解し、偏りを減らすこと

→偏りがある場合は丁寧な説明などが必要になる

| クライシス・マネジメント

AI開発ではELSIについて考える必要がある

→倫理的にどうなのか、法的にどうなのか、社会的にどうなのか

法律をクリアしたとしても、社会的に悪影響があるならば中止

- ELSI (Ethical, Legal and Social Issues)

倫理的・法的・社会的な問題のこと

→もともと生命科学の分野から生まれた用語

| クライシス・マネジメント

システムの設計段階からプライバシー、セキュリティなどを考慮したAIを開発していくことが大切である

→後からシステムを変更することは大変であり、コストがかかる

→プライバシーやセキュリティを無視したシステムは炎上の原因になってしまふため、事前に対策して炎上リスクを低く抑えることが重要である

| クライシス・マネジメント

・バリュー・センシティブ・デザイン

システムの設計段階からプライバシーやセキュリティなど
価値全般に配慮する考え方のこと

→データを暗号化して、情報が漏洩しても
大きな問題が発生しないようにする など

| クライシス・マネジメント

・シリアル・ゲーム

社会問題をゲームという形で体験し、解決するもの

→天災などにより飢饉が発生するシミュレーションゲームを通して飢饉が発生する仕組み・対策法を学ぶ など

→ゲームとしてエンターテイメント性も存在する

子供や大人が社会問題をゲームを通して学ぶことができる