第2回AI品質マネジメントシンポジウム

Ad-Sol Nissin

デジタル社会の“あした”をリードするイノベーションカンパニー

社会的視点からのAI品質

～　AIの公平性実現　～

アドソル日進株式会社

AI研究所　浜谷千波

2023年10月31日

信頼されるAI

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

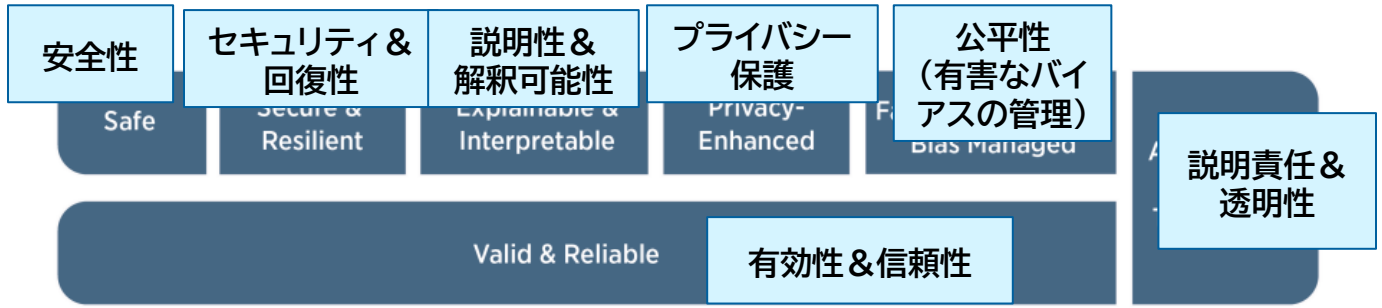
AI技術が進化し普及・社会展開が進むにつれて、 影響力が増大、扱う課題も複雑化
ブラックボックス性、 倫理面や法律面での課題が重要に

「Trustworthy AI 」 （信頼されるAI）

2023年6月 EU AI Act が採択
一般原則（議会修正案第4条a）:

- 人間の主体的な関与と監視
- 技術的な堅牢性と安全性
- プライバシーとデータガバナンス
- 透明性
- 多様性、差別がないこと、公平性
- 社会的、環境的な健全性

2023年1月 米国NIST AI RMF v1.0
（AI リスク マネジメント フレームワーク）



Characteristics of trustworthy AI systems.
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

性能面は当然として、社会的な側面が重要に

顕在化が進む　AIの公平性問題

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

AIが実社会で活用が進む中、重要性が増す社会的側面のひとつ

「公平性」

COMPAS　事件　（米国の予測的犯罪分析AI）

◆ AI公平性の様々な議論を呼んだ有名事案

米国大手通販　　AI人事採用システム　利用中止

◆ 女性に不利なジャッジ

米国大手IT企業　顔認証AI事業撤退

◆ 人種間で精度が明らかに差

三つの課題

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

公平性要求の「多様性」

不公平なAIをもたらす、「様々な要因」

「要配慮属性」、「観測できない属性」の扱い

課題1　公平性要求の多様性

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

「同等に扱われている こと」を、**説明する視点は実に様々**

◆問われているのは、「取り扱い」なのか 「結果」なのか？

例えば　入試における男女差や人種差

- 「結果の人数を同じにする」 ？
- 「応募者数に対しての合格者比を同じにする」？
- 「成績に対して判定ロジックを同じにする」 ？

2023.9
「都立高校の男女別定員撤廃へ」

女生徒はより高い点を取らないと合格出来なかった…

◆積極的差別是正措置(”Affirmative Action”）のこと

例えば　米国のハーバード大学の公平な入試訴訟

2023.6　最高裁判決　Affirmative Action(意図的にバイアスを入れ込む措置)は憲法違反

◆両立は困難な、「取り扱い」と 「結果」

- 現実にある不平等を持ち越してはいけない、という前提があるか？
- 関連する法令の要請はあるのか？

開発に入る前に**社会的要求の注意深い明確化**が不可欠

課題2　不公平なAIをもたらす様々な要因

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

社会、システム

既に人間のシステムに存在する、不平等、偏り、歴史的・制度的な偏見がもたらした結果

- 医療費予測に関するAI、　現在医療がゆきわたってない地域があると,,,
- 入試判定AI、 人種による収入　格差が学生の成績に影響をおよぼしていると…
- 求人広告を出すAI、　これまで男性がその職業に多いとすると…

人の思考

人間の、様々な「認知バイアス」

- ストリートライトエフェクト
- アンカリング
- マクナマラの誤謬
- …

統計的・技術的　なもの

データ収集～機械学習の進め方における誤謬・不適切さ

- 典型例：「シンプソンのパラドックス」

現実社会のデータのバイアス

訓練用データセットやAIアルゴリズムのバイアス

課題3　要配慮属性や、観測できない属性

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

要配慮属性（センシティブ属性）

歴史的・社会的に、「その違いによる不公平はダメだ」と一般的に認識されているもので、
現実社会データには深く入り込んでいる　　性別、人種、年齢、障害状況、e.t.c

それだけを注意しても…

- 他の属性から間接的に学習されてしまいかねない

訓練用データセットから要配慮属性をとってしまうと…

- 精度が落ちる可能性、結果の公平性実現は困難

真に欲しい属性は何か？それは観測できるものか？

本来評価・予測した属性が測定不可の場合、代替え手段でアプローチせざるえない…

たとえば　雇用シーン向けAIを開発したい場合

- 本当に評価（予測）したいのは「雇用適正」
- 代わりに用いた情報：「転職回数」「前職の勤続年数」といった個人の特定属性、
  「現在の役職」や「勤務継続年数」　など、その会社社員の実績

観測不可、
Ground Truthも存在しない！

公平性AIの有名事例

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

COMPAS （Correctional Offender Management Profiling for Alternative Sanctions）

米国の再犯予測AI　　犯罪の再犯リスクを評価し、量刑判断、執行猶予,保釈などの裁判所決定を支援

- 前科や違反歴のほか、家族や友人の犯罪歴、宗教、年齢、性別、教育、趣味、居住地域など137個の項目をもとに「再犯リスク」を評価
- 人種は直接の項目には入っていないが、黒人に不利な評価をするなどの問題提起

問題提起側の主張	ツール開発側の主張
偽陽性が大きく違う！	リスク評価に対しての正答率は同じだ！

- 刑事事件において被告から違憲性の訴訟も発生

➡ 2016年7月、ウィスコンシン州最高裁判決
https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690

合憲性は肯定、ただし、**使用するうえでの制限や注意事項**を提示

「AIの結果をconsiderしてもいいがrelyは許されない」

COMPAS事件が示唆すること

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

◆直接要配慮属性を使わなくても、不公平さは生じうる

◆「公平性の評価」定義は難しい

- 偽陽性 （真実に対して正しい評価かどうか）
- 正答率 （評価に対して、事実がどれぐらいあっているか）

➡ どちらがより適切かの判断は裁判所でも困難…

◆社会に埋め込まれている不公平の存在

- 黒人のほうが再犯率が高い、という事実そのものが、歴史的・社会的な不公平さを反映…

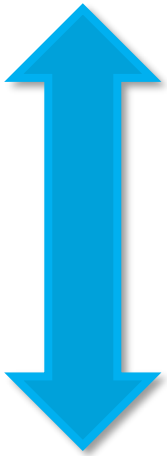
◆公平性視点からのAI活用、一つの指針

- 人より正しいといいきれなくとも、活用することで、人がより公平な判断をする助けになり得る

公平性実現プロセス

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

本質的に**定性的・抽象的な上位要求**
「不公平に取り扱われない」、「平等に取り扱われる」、e.t.c…



このギャップの、適切な解消が大事！

開発で品質管理のために必要な**「公平性メトリクス」**は、
偏りの度合いの**数値指標**

留意すべきポイント

Ad-Sol Nissin
デジタル社会の“あした”をリードするイノベーションカンパニー

◆定量的を急ぎすぎない、妄信しすぎない

- いったん「定量化」されると、その数値のほうにとらわれやすい（マクナマラの誤謬）
- 定量化するときに、どうしてもこのなかで、「解釈」「価値観」が入り込む
  ➡ 定量化の論拠は残しのちに、「見直し」せるようにすること

◆マルチステークホルダー、ダイバーシティ　の重要性、特に上流

- 認知バイアスには、かなり効果的
- AIの「監視」や「テスト」は、ドメイン専門家にはたいていの場合難しい

◆パフォーマンスやプライバシーといった他属性とのバランス判断

- とりわけ「結果の公平性」を実現しようとする場合、他の視点とぶつかることも

ご興味を持たれた方は、
「機械学習品質マネジメントガイドライン」にて、
詳細をお読み頂ければ幸いです

https://www.digiarc.aist.go.jp/publication/aiqm/