1E1-OS-24a-4

人工知能技術の導入判断にかかる意思決定者のバイアスとその解決に向けて

Toward resolving decision-making biases for deploying artificial intelligence technology

大岩 秀和 Hidekazu Oiwa　数原 良彦 Yoshihiko Suhara　淡島 英輝 Hideki Awashima

Recruit Institute of Technology

It is important to consider how to appropriately deploy artificial intelligence technologies to real-world problems along with the advancement of machine learning and computer science in recent years. However, it's hard for humans to evaluate the performance of artificial intelligence technologies precisely, and determine when and how to deploy these technologies. In this paper, we introduce the notion of decision-making biases induced from artificial intelligence technologies, which may prevent the efficient use of artificial intelligence technologies. We categorize these biases into three representative classes and introduce problems induced from these biases. Then, we discuss the approach toward resolving decision-making biases for deploying artificial intelligence technologies.

1. はじめに

近年の計算機性能の向上や機械学習をはじめとする人工知能技術の発展は目覚ましく，これまで人手で行ってきた作業の自動化や生産性の向上を目指した人工知能技術の利用は今後さらに進むと期待されている [The White House 16]．特に，機械学習などを用いた認識・予測タスクに対する技術進歩は飛躍的で，医療診断 [Dawes 79] や画像からの文字・物体の認識 [He 15] など，多くのタスクで人工知能技術の性能が人間を凌駕する結果が得られている．また，自動運転やコールセンターの問い合わせ対応など，これまで人手で処理されてきた作業の一部を人工知能技術によって模倣し，一部作業の自動化や効率化を目指す例も知られている．このように，計算機性能の向上や機械学習技術の発展に伴い，人工知能技術の適用可能性が高まっている．

一方で，人工知能技術の導入を進める上では，人工知能技術によって誘発される特有の問題にも注意しなければならない．[Sculley 15] では，人工知能技術を用いたサービス開発や運用時に発生する問題点を整理し紹介している．本先行研究では，人工知能技術を用いないソフトウェア開発や人手でのサービス運用では生じにくい様々な技術的負債に着目し，データ分析などの人工知能技術を利用する際に誘発される問題点を指摘している．このように人工知能技術によって誘発されるコストを導入判断時に見落とすと，人工知能技術の過大評価が生じる事となる．その結果，人工知能技術の導入が本来の目的にとって最適でなかったとしても，人工知能技術の導入が推進される危険性がある．適切な人工知能技術の導入を推進するためには，人工知能技術により誘発されるコストや問題点の正確な把握が必須となる．また，人工知能技術の導入判断時には，人工知能技術の評価や人工知能技術を用いない代替手段との比較が適切になされ，目的に最も合致する手段が選ばれなければならない．本研究では，人工知能技術の適切な導入判断に必要な要素として，以下の三条件を取り扱う．

- 人工知能技術の性能やコスト面での正確な評価
- 人工知能技術を用いない代替案との正確な比較検証と，その結果としての目的に最も適した手段の選択
- 導入判断に必要な情報の最小コストでの収集

人工知能技術を効果的に利用するためには，適切な人工知能技術の導入判断が重要な課題となる．これら三条件を満たすような，適切な人工知能技術の導入判断を支援する仕組み作りが求められている．

しかし，人間が人工知能技術の導入判断を行う際に，適切な人工知能技術の導入判断を阻害する様々なバイアスが存在する事が複数の先行研究で示唆されている．本稿では，意思決定者による人工知能技術の導入判断時に発生するバイアスと誘発される問題点について議論する．ここで意思決定者とは，採用する手段を最終決定する一人の人間あるいは複数の人間で構成されるグループと本稿では定義する．現状，人工知能技術の最終的な導入判断は意思決定者によって行われる事が多い．事前に定められた基準にもとづいて機械的に導入判断が行われる事は少なく，人工知能技術の評価などは意思決定者によって下されるケースも多い．しかし，意思決定者が人工知能技術の導入判断を行う場合，意思決定者がとらわれがちなバイアスの影響で不適切な導入判断が誘発される事が先行研究によって示唆されている．近年の急速な人工知能関連技術の進歩は，人工知能技術の適切な評価と導入判断をますます困難なものにしている．適切な導入判断を推進するためには，意思決定者によるバイアスの影響を軽減する仕組み作りが必要となる．本稿では，人工知能技術の導入判断時に意思決定者が陥りがちなバイアスを整理し，三種類の代表的なバイアスを紹介する．各バイアスによって誘発される問題点について紹介したのち，適切な導入判断と意思決定を支援する仕組みづくりについて考察する．本稿では，これら三種のバイアスをまとめて以下では導入判断バイアスと表記する．

2. 人工知能技術の導入判断にかかるバイアス

人工知能技術を効果的に利用するためには，人工知能技術が目的に対して最も適した手段となる場合に，適切な導入判断が意思決定者によって下されなければならない．人工知能技術がある一定の要求水準を超えた精度を達成したり，人間を上回る性能を示すのみでは不十分である．本稿では，適切な導入判断に必要な三条件を 1) 人工知能技術の正確な評価， 2) 人工知能技術を用いない代替案との正確な比較検証， 3) 必要な情報の最小コストでの収集，と定義した．上記の三条件を満たさ

連絡先: 大岩秀和, oiwa@r.recruit.co.jp

表 1: 導入判断バイアスと誘発される問題

	過大・過小評価	不正確な比較評価	不必要な作業要求
先入観バイアス	√	√	
失望バイアス	√	√	
解釈性バイアス		√	√

ない導入判断は，目的に不適な手段選択や意思決定にかかる不必要なコスト増大などの問題を誘発する．しかし，人工知能技術の導入判断時には，正確な評価検証等を阻害する導入判断バイアスが誘発される事が先行研究によって示唆されている．これらの導入判断バイアスは適切な導入判断を阻害し，

- 人工知能技術の過大なあるいは過小な評価
- 人工知能技術を用いない代替案との不正確な比較検証と，目的に不適な手段の選択
- 正確な評価や公平な比較に不必要な作業要求

などの問題を誘発しうる．

本章では，人工知能技術の導入判断時に発生する代表的な導入判断バイアスを

1. 先入観バイアス
2. 失望バイアス
3. 解釈性バイアス

の三種類に分類する．各バイアスの概要と誘発される問題について以下の章で詳細を述べる．表 1 では，各バイアスによって誘発される問題点を整理している．

2.1 先入観バイアス

人工知能技術の適切な評価と人工知能技術を用いない代替案との適切な比較評価を実現するためには，意思決定者が特定の手段に対して偏った先入観を持たず，各手段の正確な評価と比較が必要となる．しかし，人工知能技術を利用したアルゴリズムと代替案との比較検討を行う際，公平な比較を阻害する先入観バイアスが誘発されることが示唆されている．

統計学や人工知能技術を利用したアルゴリズムの性能が様々なタスクで人間を上回る結果 [Dawes 79, Dzindolet 02] を示すと同時に，人間のアルゴリズムに対する信頼度を実証実験を通して測る研究が行われてきた [Dawes 79, Eastwood 12, Logg 16]．これらの先行研究では，人間のアルゴリズムへの信頼度と，代替手段となる自身や他の人間に対する信頼度とを比較する実験を行っている．人間に対して，アルゴリズムと人間のどちらがより信頼できる手段であるか選択してもらう形式でそれぞれの手段への信頼度を測る実証実験となっている．これらの先行研究によって得られた洞察として，人工知能技術や統計学を利用したアルゴリズムが人間の判断よりも正確な場合にも，人間はアルゴリズムを信用しにくく利用を忌避する傾向があることが示されている [Dawes 79, Dzindolet 02]．アルゴリズムへの信頼性が低くなるタスクの例として，医療診断行為 [Dawes 79, Eastwood 12] や大学・会社の採用判断 [Dietvorst 15, 潜道 16] などが先行研究で挙げられている．逆に，アルゴリズムへの信頼度が人間を上回るタスクも存在する事が知られている [Logg 16]．このように先入観に起因するバイアスに意思決定者が影響を受ける時，適切な導入判断が阻害され人工知能技術の過大あるいは過小な評価につながる．その結果，人工知能技術を用いない代替案との正確な比較検証が困難となり目的に適さない手段を選択してしまうリスクが発生する．

[Logg 16] では，包括的な先入観バイアスの実証実験を行っている．この先行研究では，代替案やタスクの種類などの各種条件を変更した時の先入観バイアスの変化を明らかにしている．結論として，以下の考察を与えている．

- 写真に写る人間の体重予測など客観性の高いタスクでは，人間はアルゴリズムに対して高い信頼を置く傾向にある．
- 映画の推薦など主観性の高いタスクでは，アルゴリズムへの信頼度は下がる傾向にある．
- 一般的な他人ではなく，自身の判断やエキスパートの判断がアルゴリズムの比較対象となる場合，アルゴリズムに対する信頼度は下がる傾向にある．

このように，各種条件や時勢に応じて，先入観バイアスの性質は変化しうる．先入観バイアスによって人工知能技術と人間への信頼度に歪みが生じている状態では，人工知能技術の適切な導入判断は困難となる．意思決定者が人工知能技術の評価検証や人工知能技術を用いない代替案との比較を適切に行うためには，先入観バイアスの影響を軽減する仕組みが必要となる．

2.2 失望バイアス

先入観バイアス以外にも，人工知能技術の導入判断時に意思決定者が陥りがちなバイアスが存在する．人工知能技術の評価検証を行う際にも，導出された結果が誤っている事実を観察する過程において，信頼度の過度な低下が生じるバイアスの存在が示唆されている．

人工知能技術の導入判断時には，正確な有用性評価と人工知能技術を用いない代替案との比較検証のため，導出された結果を正解と照らし合わせながら評価検証を行うことがある．導出された結果の検証過程において，人間やアルゴリズムへの信頼度の変化度合いを分析した先行研究が存在する [Dietvorst 15]．本先行研究から得られた洞察として，検証作業中にアルゴリズムから導出された結果が誤りを含む事を確認すると，人間はアルゴリズムへの信頼度を著しく損なう傾向にあることが実証されている．さらに，検証作業中のアルゴリズムへの信頼度の低下度合は，人間が導出した結果の誤りによる低下度合よりも大きい事が示されている．これら実証実験から得られた洞察として，人間はアルゴリズムの失敗に対してより厳しく評価すること，アルゴリズムが誤った時には人間が誤った場合よりも大きく信頼度が下がる傾向にあることが示されている．たとえアルゴリズムのパフォーマンスが人間より高く先入観バイアスの影響も存在していなかったとしても，検証作業中にアルゴリズムによって導出された結果が誤りを含んでいることを確認すると，人間はアルゴリズムへの信頼度を大幅に低下させ，実システムへの導入を忌避しがちになるリスクが存在する事が本先行研究において示唆されている．

人工知能技術を利用したアルゴリズムは完璧な結果のみを導出することは困難な事が多い．たとえ人工知能技術が代替案を性能で上回っていたとしても，人工知能技術を用いたアルゴリズムは誤った結果を導出することがある．ある目的において人工知能技術の導入が最適であったとしても，性能の検証を繰り返し行ううちに人工知能技術への信頼度が代替案と比較して過度に低下してしまうと，人工知能技術の導入が忌避されるリスクが存在する．このような失望バイアスを軽減する仕組みづくりも適切な導入判断のためには必要となる．

2.3 解釈性バイアス

人工知能技術の導入判断時には，アルゴリズムの挙動の透明性や導出された結果の解釈性も重要な役割を果たす．人工知能技術などを用いた自動的な意思決定をサービス等に導入する場合，ユーザへの影響が大きい場合にはアルゴリズムが導出した結果の根拠をユーザが知ることができる権利を要求する規制が EU では成立している [Goodman 16]．また，人工知能技術の導入判断を人間が行う際にも人工知能技術を用いたアルゴリズムの解釈性が求められるケースは多い．データ分析に従事する複数のデータサイエンティストに日常の業務内容などについてインタビューを行った調査結果が先行研究において報告されている [Kim 16]．インタビューの結果として，データサイエンティストの多くが上司や意思決定者への報告を行う際に分析結果の解釈性を高めるための業務を行っていると回答している．対象としているタスクでの専門用語を利用してモデルの挙動や得られた洞察を表現することで意思決定者に対する解釈性を高めていることが報告されている．別の先行研究では，人工知能技術の適切な評価や利用・運用時に求められる解釈性の詳細な分類と分析を行っている [Lipton 16]．本先行研究では人工知能技術の解釈性が必要となる理由を，1) 運用者・利用者からの信頼性向上，2) 意思決定において有益な情報の提供，などに分類し，各目的に求められる解釈性の内訳とそれぞれの向上手段について整理している．予期せぬ出力やそれに伴うエラーを防ぐため，または正確な性能評価や代替案との比較検証のコストを低減するためにも，人工知能技術を用いたアルゴリズムの透明性や解釈性は，アルゴリズムそのものの性能に加えて重要な評価要因となっている [Lipton 16]．

一方で，人工知能技術の解釈性は高いほどよいものでは必ずしもなく，解釈性を過度に高めすぎる事のリスクも先行研究では指摘されている [Lipton 16]．本先行研究では医療診断行為への人工知能技術導入判断を例に，人工知能技術を用いたアルゴリズムの解釈性と性能がトレードオフの関係になるリスクが示唆されている．意思決定者が人工知能技術に解釈性を過度に求めると，解釈性は高いが診断行為の予測能力が低い手段を選択するリスクが存在する．導入判断において解釈は重要な要因ではあるものの，解釈性を重視するあまり本来の目的から逸脱した手段を採用するリスクをはらむ．さらに，正確な比較評価に不必要な作業コストを発生させるリスクも存在する．導入判断において解釈性が過度に求められている時，意思決定者の解釈に過剰に適合した手段が導入されるリスクが高まる事にも注意しなければならない．採用活動や入学審査などのタスクにおいて意思決定者が過去に自身で作業を行った経験がある場合，過去の経験を模倣するアルゴリズムを意思決定者は選択しがちになるリスクが存在する．この際，過去の経験に過剰に適合するような歪んだ最適化が開発時に行われるリスクも存在する．人工知能技術の導入判断時には，これらの解釈性バイアスの影響にも注意しなければならない．

人工知能技術を用いたアルゴリズムの透明性が低い場合や下された判断の根拠が解釈困難な場合，意思決定者の人工知能技術への忌避感が強くなる事がある．しかし，意思決定者が人工知能技術の解釈性を要求する場合，その要求内容は導入判断に強い影響を及ぼす．解釈性の向上はアルゴリズムの信頼性保証や予期せぬ挙動の防止に重要な要素であるものの，解釈性を過剰に重要視するあまり適切な導入判断を妨げることのないよう注意する必要がある．

3. 導入判断バイアスの解決に向けて

2. 章では，意思決定者による人工知能技術の導入判断時に発生する三種の代表的なバイアスを紹介し，それぞれの問題点について整理した．導入判断バイアスは適切な導入判断を阻害し，人工知能技術の非効率的な利用が誘発されるリスクが存在する．適切な導入判断を実現するためには導入判断バイアスの影響を軽減する仕組みの構築が求められる．導入判断バイアスの解決に向けた仕組みづくりとして，本稿では以下の二種類のアプローチについて考察する．

1. 事前の導入判断基準設定による機械的な意思決定
2. 人工知能技術によって導出された結果の修正システム導入

導入判断バイアスを回避するためのアプローチの一つは，事前に導入判断基準となる項目を設定し，導入判断時には作成された基準に従って機械的に意思決定を行う方法である．事前に定めたルールや基準項目に従った自動的な導入判断を実現することで，先入観バイアスなどの不正確な評価・比較を誘発するリスクを導入判断プロセスから取り除く事が可能になる．先行研究では，人工知能技術を利用したシステム品質を機械的に評価する方法論として，人工知能技術評価のためのルーブリックを作成している [Breck 16]．事前に導入判断基準項目を過不足なく列挙することは困難であるものの，人工知能技術の評価基準や必要十分な可視化項目を事前に定めることで導入判断バイアスのリスク低減が期待される．また，事前に設定した基準項目に従った機械的な導入判断を実現するためには問題定義やタスクの切り分けも重要となる．人工知能技術の導入検討がプロジェクトの意思決定に含まれる場合には，人工知能技術の評価や代替案との比較を行う際に，正解率などの数値指標や解釈性の向上に必要な要求項目など機械的にチェック可能な基準項目から最終的な導入判断が自動的に可能となるようにタスクを切り出すことが重要となる．自動的な導入判断の運用が不可能な場合，先入観バイアスや検証事後の失望バイアスの影響を回避することは困難となる．例として，人工知能技術による医療画像診断を導入検討する際に，意思決定者やその他の人間のアルゴリズムへの信頼度を導入判断に利用する場合，導入判断バイアスが誘発されるリスクを回避することが困難となる．人工知能技術による医療画像診断の導入判断を自動的に行うためには，人工知能技術による診断性能が一定の基準を上回るかどうかなど，機械的に判断可能な項目のみを用いて導入判断が可能となるように事前に導入判断のためのルールを設定することが重要となる．このように，事前の判断基準項目設定による機械的な導入判断を実現するための取り組みは，導入判断バイアスの影響を軽減し人工知能技術の適切な導入判断を推進するアプローチとして有力である．

事前の基準項目設定による導入判断の自動化は導入判断バイアスを回避するための有力なアプローチであるものの，人工知能技術の導入判断基準を事前に過不足なく整備する事は非常に困難である．導入判断バイアスの影響を軽減するために補完的な別のアプローチとして，人工知能技術によって導出された結果の修正を可能にするシステム構築が挙げられる．

[Dietvorst 16] では，失望バイアスに代表される導入判断バイアスの影響を軽減する方法として，アルゴリズムが導出した結果を修正する手段を意思決定者に与えることの重要性を主張している．本先行研究では様々な実証実験を通して得られた洞察として，アルゴリズムが導出した結果を修正する手段を人間に与えることで，アルゴリズムに対する忌避感を軽減可能なことを実証している．アルゴリズムが導出した結果をそのまま利用する場合，先入観バイアスや失望バイアスの影響によりアルゴリズムに対する信頼度が低くなるリスクが存在する．一方で，たとえ人間が修正可能な数が僅かだとしても修正する手段を与えることで人間はアルゴリズムへの信頼度を高め導入判断バイアスの影響を軽減できることが示されている．この時，人間が修正可能になる結果の数の多寡は大きな影響を及ぼさない．このように，意思決定者のアルゴリズムに対する忌避感を防ぐ手段として，人工知能技術によって導出された結果を修正する手段を与えるシステム構築が重要となる事が示唆されている．さらに，人間がアルゴリズムの結果を容易かつ効果的に修正可能なように，各予測結果の解釈や統計量の可視化を含めた修正支援システムを構築することが，導入判断バイアスを低減するための補完的で有力なアプローチとなる．

4. まとめ

本稿では，意思決定者による人工知能技術の導入判断時に発生する三種類の代表的なバイアスと各バイアスが引き起こす問題点について整理を行った．人工知能技術を効果的に利用するためには適切な導入判断が必要であり，導入判断バイアスの影響を軽減する仕組みづくりが必要となる．本稿では，導入判断バイアスを軽減するための仕組みとして，1) 事前の導入判断基準設定による機械的な意思決定，2) 人工知能技術が導出した結果の修正支援システム導入，の二つが有効的な手段であることを紹介した．ただし，これらの手段は導入判断バイアスの影響を完全に取り除くには未成熟であり，人工知能技術の導入判断時に誘発されるバイアスなどの問題点をより効果的に解決するための取り組みが今後さらに重要となる．

参考文献

[Breck 16] Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D.: What' s your ML Test Score? A rubric for ML production systems, in *Reliable Machine Learning in the Wild - NIPS 2016 Workshop* (2016)

[Dawes 79] Dawes, R. M.: The robust beauty of improper linear models in decision making., *American Psychologist*, Vol. 34, No. 7, pp. 571–582 (1979)

[Dietvorst 15] Dietvorst, B. J., Simmons, J. P., and Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err, *Journal of Experimental Psychology: General*, Vol. 144, No. 1, pp. 114–126 (2015)

[Dietvorst 16] Dietvorst, B. J., Simmons, J. P., and Massey, C.: Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them, *Management Science*, Vol. 58, No. 12, p. mnsc.2016.2643 (2016)

[Dzindolet 02] Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. a.: The Perceived Utility of Human and Automated Aids in a Visual Detection Task, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 44, No. 1, pp. 79–94 (2002)

[Eastwood 12] Eastwood, J., Snook, B., and Luther, K.: What People Want From Their Professionals: Attitudes Toward Decision-making Strategies, *Journal of Behavioral Decision Making*, Vol. 25, No. 5, pp. 458–468 (2012)

[Goodman 16] Goodman, B. and Flaxman, S.: EU regulations on algorithmic decision-making and a "right to explanation", in *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, pp. 1–6 (2016)

[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in *2015 IEEE International Conference on Computer Vision (ICCV)*, Vol. abs/1502.0, pp. 1026–1034, IEEE (2015)

[Kim 16] Kim, M., Zimmermann, T., DeLine, R., and Begel, A.: The emerging role of data scientists on software development teams, in *Proceedings of the 38th International Conference on Software Engineering - ICSE '16*, pp. 96–107, New York, New York, USA (2016), ACM Press

[Lipton 16] Lipton, Z. C.: The Mythos of Model Interpretability, in *ICML Workshop on Human Interpretability of Machine Learning* (2016)

[Logg 16] Logg, J. M.: *When do people rely on algorithms?*, PhD thesis, University of California, Berkeley (2016)

[Sculley 15] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D.: Hidden Technical Debt in Machine Learning Systems, *Advances in Neural Information Processing Systems*, pp. 1–9 (2015)

[The White House 16] The White House, : Artificial Intelligence, Automation, and the Economy (2016)

[潜道 16] 潜道 隆 : ビッグデータでどのような価値創造が可能か？, 経営情報学会全国研究発表大会要旨集 2016 年秋季全国研究発表大会, pp. 2–4, 一般社団法人経営情報学会 (2016)