人事領域における AI 導入とアルゴリズム嫌悪
：社会規範の役割に関する実証研究

神内 郁哉 [a]　武川 菜那 [b]　西浦 智顕 [c]

要約

本研究は、人事評価領域における AI 支援システム導入に伴うアルゴリズム嫌悪を，社会規範フレーミングによって軽減できるか実験を用いて検証した．アルゴリズム嫌悪を誘発させた後，「AI が多くの企業で導入されている」という社会規範を提示する実験群と，提示しない統制群に分け、AI からの支援を受けた人々の評価行動を比較した．結果，両群の評価スコアや AI への信頼度に有意差は見られず，当初の仮説は棄却された．この結果は，抽象的なフレーミングが具体的な AI の失敗経験による強い嫌悪感を覆すには不十分であった可能性，あるいは AI の普及によってアルゴリズム嫌悪そのものがすでに形骸化しつつある可能性を示唆している．本研究は，実験デザインの欠陥という限界を認めつつ，今後の研究でより効果的な介入方法の検証や，AI に対する世代間の信頼度の差を比較する必要があることを提言する．

JEL 分類番号： M50，D87，O33
キーワード：人事評価，アルゴリズム嫌悪，社会規範，AI 支援システム，認知バイアス

---

*なお，本論文に関して，開示すべき利益相反関連事項はない．
[a] 神内 郁哉 同志社大学 cgfj0143@mail3.doshisha.ac.jp
[b] 武川 菜那 同志社大学 cgfj0547@mail3.doshisha.ac.jp
[c] 西浦 智顕 同志社大学 cgfj0395@mail.3.doshisha.ac.jp

1. イントロダクション

現代のビジネス環境において，DX（デジタルトランスフォーメンション）は企業の成長に不可欠な戦略であり，AI はその中核技術である。特に、膨大なデータを取り扱う人事（HR）部門は AI との親和性が高く，その業務変革に大きな期待が寄せられている．AI はデータ分析とパターン認識において卓越した能力を発揮し，新たなビジネス価値の創出を可能にする．しかし，採用や人事評価といった重要な意思決定領域では，AI の導入に対して心理的抵抗や懸念が存在する．人々は AI が感情や文脈を理解できないこと，または評価バイアスを含むのではないかと懸念している．この AI に対する不信感は，最終決定者が AI の推奨を軽視する原因となり，結果として AI の優れた分析結果が意思決定に十分に反映されない，アルゴリズム嫌悪（Dietvorst et al., 2015）を引き起こす．本研究は，このアルゴリズム嫌悪の軽減を目的とし，AI に対する心理的抵抗を払拭してその活用を促進する．人間と AI が協調し，質の高い人事評価や採用判断を実現するための理論的・実践的な枠組みを構築する．これにより，AI が HR 部門の変革を加速させ，組織全体の持続的な成長に貢献する未来を目指す.

2. サーベイ実験と仮説

2.1.実験期間と参加人数

実験実施期間は 2025 年 9 月 4 日から 2025 年 9 月 5 日であり，いずれかの企業で過去もしくは現在において人事評価の経験がある方（人事部に所属した経験がある，所属している方，もしくは人事部所属でなくても，部署で部下等の評価を行った経験がある，行っている方）に対して行った．最終的な有効回答数は，20 歳から 75 歳（平均年齢 48.7 歳）の 148 名（男性 105 名，女性 37 名，その他 6 名）となった.

2.2.実験デザイン

本実験はクラウドソーシングサイトの Yahoo!クラウドソーシングを用いて，参加者には仮想シナリオのもとで，企業の人事担当者の役割として人事評価を行ってもらった．まず，先行研究をもとに参加者にアルゴリズム嫌悪を誘発させるべく，アルゴリズムが不完全であり，間違った評価をしたという過去の事実を示した（Dietvorst et al., 2015）． その後，参加者を以下の 2 つの群にランダムに割り振った．統制群：被評価者に対する AI のポジティブなフィードバックを提示する．実験群：被評価者に対する AI のポジティブなフィードバックを提示する．加えて，AI が多数の企業で導入されているという主旨の社会規範を強調したフレーミングを提示する．実験群では AI にかかる社会規範フレーミングを提示することにより，参加者に対して AI 嫌悪の抑制を図った．そして，参加者には

社員プロフィールを提示した後，目標と自己統制による管理である MBO（Drucker, 1954）と，知識やスキルといった目に見える能力だけでなく，動機や特性などの見えない部分を含むコンピテンシー（McClelland，1973） の項目に分け，これらの評価を【1：非常に劣る 2：劣る 3：やや劣る 4：やや優秀 5：優秀 6：非常に優秀 】の 6 段階で尋ねた．また，参加者の基本的な個人属性（年齢，性別）や本実験における AI に対する信頼度，AI 全般への信頼度，AI の使用頻度，AI が人間の判断より正確であると思うかなどの情報を収集した．

2.3.仮説

「実験群の評価者は統制群の評価者に比べ，AI によるフィードバックをより強く信頼し，寛大な評価をする.」仮説の根拠として，先行研究ではアルゴリズムの使用が規範であると提示された状況では，人々は人間よりもアルゴリズムによる判断を好む傾向を示している（Bogard and Shu，2022）．そのため本実験では，社会規範フレーミングがなされる実験群において，AI からのポジティブなフィードバックに強く影響され，評価者は寛大な評価をすると考えた．

3. 結果

表 1　人事評価セッションにおける評価項目ごとの t 検定結果

評価項目	統制群（n=74）	実験群（n=74）		
	M (SD)	M (SD)	t 値	p 値
目標管理意識（MBO）	4.91 (0.78)	4.84 (0.76)	0.53	0.594
問題解決能力	4.23 (0.84)	4.19 (0.77)	0.31	0.76
協調性	4.45 (1.00)	4.24 (0.82)	1.35	0.179
責任感	4.54 (0.97)	4.47 (0.91)	0.44	0.662
コミュニケーション能力	4.11 (0.88)	4.14 (0.82)	-0.19	0.847

注. M=平均値, SD＝標準偏差

表 2　AI への信頼度の t 検定結果

評価項目	統制群（n=74）	実験群（n=74）		
	M (SD)	M (SD)	t 値	p 値
Confidence1	4.35(0.77)	4.26(0.74)	0.76	0.447
Confidence2	4.24(0.76)	4.19(0.73)	0.44	0.66
Accuracy	3.92(0.92)	3.80(0.81)	0.85	0.394

注. Confidence1：本サーベイにおける、人事評価における AI の支援システムの信頼度
Confidence2：本サーベイに関わらず、人事評価における AI の支援システムの信頼度
Accuracy：人工知能（AI）は、人間の判断よりも正確であると思うか

3.1.本分析

統制群及び実験群間で t 検定を実施し，有意差を測定した．全項目において $p>0.05$ となり，有意差は見られなかった．実験後に行った AI への信頼度に関するアンケートにおいても同様に有意差は見られず，実験を通した意識的な変化はなかったと考えられる．

3.2.補助分析

表 3　MBO 評価と AI への信頼度の相関係数

評価項目	Group	Confidence1	Confidence2	Accuracy
MBO	統制群	0.24 |	0.203	0.181
	実験群	0.367	0.326	0.436

評価の寛大さと AI への信頼度に関する相関分析をおこなった．表 3 のように，両群共に相関係数 0.5 以上の相関関係はみられなかったが，実験群において相対的に強い相関が見られた．特に MBO 評価において，統制群では AI への正確性について問う項目（Accuracy）との相関係数がわずか 0.181 だったのに対し，実験群では 0.436 という中程度の正の相関関係が生じた．

4. 考察

本研究は，人事評価において AI 支援システムを利用する際，アルゴリズム嫌悪を軽減するための「社会規範フレーミング」が，評価者の評価行動や AI への信頼度に与える影響を検証することを目的とした．分析の結果，社会規範のフレーミング介入を受けた実験群と AI 支援のみの統制群との間で，評価スコアや AI への信頼度の平均値に統計的な有意差は認められなかった．この結果から，当初の「実験群の評価者は，統制群に比べて寛大な評価をする」という仮説は棄却された．この事実は，本研究で用いた社会規範フレーミングが，評価者の AI への信頼度を直接的に向上させたり，評価の寛大さを変化させたりする上で，限定的な効果しか持たなかったことを示唆している．したがって，当初の「実験群の評価者は，統制群に比べて寛大な評価をする」という仮説は棄却された．

一方，平均値に差はなかったが，相関分析からは一つの発見が見出された．社会規範フレーミングを受けた実験群では，統制群には見られなかった「MBO 評価」と AI への信頼度を測る各種項目（Confidence1, Confidence2, Accuracy）との間に若干の正の相関が確認された．これは，社会規範フレーミングが参加者の心理に以下のようなプロセスを引き起こしたことを示唆するのではないだろうか．フレーミングは，それ自体が AI への信頼

度を直接上げる効果はなかったが，参加者に「AI の利用は社会的に許容されている」という新たな規範を意識させた．この意識は，「AI を使って評価を下す」という自らの行動を正当化するための強力な拠り所となる．Bogard（2022）らの規範理論が示すように，人々は非規範的な選択によって悪い結果が生じた場合により強い後悔を感じるため，意思決定において規範を重視する．その結果，実験群の参加者は「社会的に認められている AI を使って高い評価を付けたのだから，この AI は信頼できるし，そもそも AI は正確なはずだ」という形で，自らの行動（MBO 評価）と信念（AI への信頼）を一貫させようとする自己正当化の心理が働いた可能性がある．

当初の仮説が統計的に支持されなかった原因は，主に「実験デザインの欠陥」と「前提の 見誤り」の２つの可能性が考えられる．まず「実験デザインの欠陥」についてだが，第１に，アルゴリズム嫌悪の誘発と社会フレーミングによる軽減のパワーバランスを見誤った可能性がある．本研究では全参加者に対して，まず「AI の予測と人間の評価が乖離した失敗事例」を提示した．これは，Dietvorst（2015）らが繰り返し実証したように，人々がアルゴリズムの誤りを一度でも目撃すると，人間の誤りよりも急激かつ過度に信頼を失うという「アルゴリズム嫌悪」の最も強力な誘発要因である．この具体的でインパクトの強いネガティブな体験に対し，実験群にのみ提示された「多くの企業で導入されている」という社会規範フレーミングは，抽象的なテキスト情報に過ぎなかったと考えられる．人事評価という，従来人間が行うことが強い規範として存在する領域において，この程度の弱い介入で，具体的な失敗事例によって強固に喚起されたアルゴリズム嫌悪を覆すことは困難であった可能性が極めて高い．つまり，強力な「嫌悪の誘発」を，微力な「嫌悪の軽減策」で打ち消すことができなかったのではないだろうか．第２に，AI によるポジティブな支援情報の影響力が強すぎた可能性がある．本研究の AI は中立的な情報提示者ではなく，「『控えめ』といった印象に引きずられず成果に基づいて評価してください」と，参加者にポジティブな評価を促す強力な支援者として機能していた．この AI による強いポジティブな働きかけが，統制群の参加者に対しても，嫌悪感を覆すような強い影響を与え，評価を寛大化させていた可能性がある．しかし，本研究では「AI の支援がない群」を設けていないため，この AI 支援情報が持つ本来の影響力を測定することができない．もし AI の支援自体が評価を大きく引き上げていたとすれば，それに加えてフレーミングがもたらす僅かな追加効果を検出するのは統計的に困難となる．つまり，AI の強い影響によって両群の評価が既に天井に近づいており，介入の効果が見えにくくなったと考えられる．

もう一つの可能性として考えられるのが，「前提の見誤り」である．本研究は「人事評価領域にはアルゴリズム嫌悪が存在する」という前提に立っている．しかし，現代のビジ

ネス環境，特にデジタルネイティブ世代が意思決定層になりつつある状況下では，アルゴリズム嫌悪という現象自体が既に形骸化， あるいは解消されつつあるという視点が考えられる．Bogard（2022）らの先行研究では，アルゴリズムへの態度はその利用が「規範的か」に強く依存すると論じられている． 例えば，スマートフォンのナビゲーションアプリやオンラインのマッチングサービスのように，アルゴリズムの利用が完全に社会規範となった領域では，もはやアルゴリズム嫌悪はほとんど見られない．近年，ChatGPT に代表される生成 AI の爆発的な普及により，多くのビジネスパーソンが日常的に AI の支援を受けるようになった．これにより，ビジネスにおける意思決定プロセスにおいても，AI の利用が急速に「新たな規範」として受容されつつある可能性は否定できない．

5. 本研究の限界と今後の展望

本研究は，人事評価における AI 支援システム導入時のアルゴリズム嫌悪を，社会規範フレーミングで軽減できるかを検証したが，いくつかの限界がある．第１に，実験デザインの不均衡性である．AI の失敗事例による強力な嫌悪誘発に対し，社会規範フレーミングは微弱であり，介入効果を検出困難にした可能性がある．また，AI 支援自体の効果を測定するための，社会規範に関するフレーミングのない純粋な統制群が不在だった．第２に，現代のデジタル環境におけるアルゴリズム嫌悪が克服されている可能性がある．生成 AI 普及により，AI 利用が新たな規範となり，嫌悪感自体が薄れていると考えることもできる．今後の展望として，今回用いた社会規範以外の，より強力かつ多様な介入手法の探索と，AI 支援なしの統制群を含む厳密な実験デザインが求められる．また，AI への態度を世代間比較するなど，時代的・世代的要因を考慮した研究により，アルゴリズム嫌悪の現代的側面を深く探求することが必要である．

6. 引用文献

・Dietvorst, B. J., Simmons, J. P., and Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144, 114-126.

・Drucker, P. F., 1954. The Practice Of Management. Harper & Brothers.

・McClelland, D. C., 1973. Testing for competence rather than for 'intelligence'. American Psychologist 28, 1-14.

・Bogard, J. and Shu, S., 2022. Algorithm aversion and the aversion to counter normative decision procedures. Journal of Experimental Psychology: General.