様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ４ 月 ２１ 日現在

機関番号：３２６９０
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２０６５１

研究課題名（和文）アルゴリズム的偏見を回避する人工知能原則の探求

研究課題名（英文）Quest for artificial intelligence principles to avoid algorithmic bias

研究代表者
岡田　勇（OKADA, ISAMU）

創価大学・経営学部・准教授

研究者番号：６０３２３８８８

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：本研究では、アルゴリズム的偏見という、既存の偏見がＡＩの利用により強化・固定化される現象について検討し、具体的には主に以下の４つの研究を行った。１．研究動向調査、２．進化ゲーム理論を用いた理論的分析、３．被験者実験を用いた実証的分析、４．人工知能原則の抽出と一般への還元。つまり、アルゴリズム的偏見を回避するための人工知能原則について、進化ゲーム理論を用いた理論的分析から規範的知見として抽出するとともに、被験者実験を用いた実証的分析から実現可能性について検討した。またこの両面を踏まえて一般原則の抽出に挑戦した。その成果については哲学者と議論することで、誰にでも分かりやすい考え方として提示した。

研究成果の学術的意義や社会的意義
アルゴリズム的偏見は、人工知能の発達に伴い増大されるため、これからの社会にとって潜在的な脅威である。これをどのように回避するのかというのは、人工知能の開発の現状を理解しつつも、哲学や認知心理学といった他分野からのアプローチが求められる。そこで、本研究では、進化ゲーム理論を用いた数理的・認知論的なアプローチによって、望ましい規範の特徴について抽出し、その有効性について被験者実験で確認した。また広く国民に分かりやすい原則として提示するために、哲学者と協力して一般向けの著作物を作成し、出版を計画中である。

研究成果の概要（英文）：In this project, I have extracted normative knowledge about artificial intelligence principles for avoiding algorithmic bias from theoretical analysis using evolutionary game theory, and have examined the feasibility from empirical analysis using subject experiments. I also tried to extract general principles based on those two approaches. By discussing the results with a philosopher, I have presented a way of thinking that is easy for anyone to understand. Specifically, I mainly conducted the following four studies. 1. Research trend survey, 2. Theoretical analysis using evolutionary game theory, 3. Empirical analysis using subject experiments, and 4. Extraction of artificial intelligence principles and reduction to the public.

研究分野：計算社会科学

キーワード：アルゴリズム的偏見

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

深層学習をはじめとする人工知能技術は、様々な驚くべき革新をもたらすと同時に倫理的課題も生起させている。その中でもアルゴリズム的偏見は、マイノリティへの差別や社会の分断を助長しかねない深刻な問題である。これは、既存の偏見がＡＩの利用により強化・固定化される現象であり、学習データそのものに偏見が投影されていることに原因がある。例えば、大手 IT 企業の人事採用の際に開発されたアルゴリズムは、男性を採用しやすくする差別的判定をして問題となった。このようにデータドリブンで価値判断を求める際に生じるアルゴリズム的偏見の回避は、高度科学技術社会の到来に伴って必然的に生じる重大なテーマである。これまでは、公平性マイニングやジェンダーニュートラル・ジェンダースワッピングといった技術的改良が提案されている。しかし、アルゴリズム的偏見が人工知能技術に内在化する問題である以上、こういった技術的解決とは別に、運用する側がそのような偏見を十分考慮したうえで、どのように意思決定を行うべきかに関する社会的合意や哲学的議論が必要となるにも関わらず、そのような研究プロジェクトはほとんどない状況であった。

２．研究の目的

本研究では、**アルゴリズム的偏見を回避するための人工知能原則**について、進化ゲーム理論を用いた理論的分析から規範的知見として抽出するとともに、被験者実験を用いた実証的分析から実現可能性について検討する。またこの両面を踏まえて一般原則の抽出に挑戦する。その成果については哲学者と議論することで、誰にでも分かりやすい考え方を提示することを目的とする。

３．研究の方法

主に４つの研究を行った。それぞれの詳細と成果は次節に詳述する。
- 研究動向調査
- 進化ゲーム理論を用いた理論的分析
- 被験者実験を用いた実証的分析
- 人工知能原則の抽出と一般への還元

４．研究成果

（１）研究動向調査

アルゴリズム的偏見に対応する技術的動向を調査した。

(1-1)　深層学習は翻訳・生態認証・自動運転・ゲームなど多様な応用分野を有している。これを可能にしたのが利用可能データの増加である。IDC によれば 2010 年に世界のデータ空間に蓄積されているデータ総量は 1ZB(=10^{12}GB)であったが、2020 年には 50ZB となり、2025 年には 175ZB と予測されている。人工知能におけるデータの比重が増加するにつれ、データに内在するノイズとバイアスが問題視されるようになる。後者のうち最大の関心はジェンダーバイアスと人種バイアスである。

(1-2)　バイアスの発生時点における分類と類型化

データ内在バイアス	アルゴリズム内在バイアス	人間の評価時のバイアス
歴史的バイアス 表現バイアス 測定バイアス 一時的バイアス 略された価値のバイアス	アルゴリズムバイアス 評価バイアス 累積バイアス 人気バイアス ランキングバイアス 緊急バイアス リンクバイアス	行動バイアス 表現バイアス 文脈生成的バイアス・ 社会的バイアス

(1-3)　データバイアスを回避する公平性の検討

データバイアス発生因は３つの理由に類型化できる。例：COMPAS(犯罪リスク算定システム)

１．偏ったデータ：特定の人口統計のデータが多いとトレーニングデータにバイアスがかかる

２．外部要因：人種ではなく、周囲の環境が犯罪に影響を及ぼす。疑似相関

３．公平性の基準誤り：公平性の尺度がそもそも現実を正確に反映していない可能性がある

ニューラルネットワークを用いた公平性の測定手法としては Demographic Parity [Feldman et

al., 2015] や Equalized Odds [Hardt et al., 2016]、Predictive Parity [Chouldechova, 2017] などが提案されているが、それを新たな公平性基準[Barocas et al., 2019] として独立性・分離可能性・十分性の観点から評価した。その結果、データバイアスを回避するには以下の３つの時点からの修正が重要であることを指摘した。
1. 学習データそのものを修正するプロセス(元データの評価に基づく)
2. データベースから生成されるプロセス(人工的な小規模非バイアスデータとの比較)
3. 取得するデータそのものを選択するプロセス(データ探索時からデータ数の均等化を考慮)

（２）進化ゲーム理論を用いた理論的分析
万人が万人に対し善悪の２値からなる印象を持っているとして、社会的ジレンマーゲームを行ったときに、どのような規範（行動原則）を有している場合に、社会的に協力行動で安定するかについて網羅的検討を行い、その成果を Scientific Reports に発表[6]し、国際会議[4,12,14]、国内会議[11,15]で報告した。また、同研究をネットワーク構造の持つ公共財ゲームに拡張することで、より現実を意識した分析を行い、その成果を Scientific Reports に発表[8]した。
間接互恵性について、ダイナミクスの精緻なシミュレーション分析を行い、その成果を Frontiers in Physics に発表[10]した。また、その統合モデルについての分析を行い、成果を国内会議[21-25]で報告した。
分析手法である社会シミュレーションについて、新たに勃興しつつある計算社会科学の文脈からまとめて、同学問における国内初の教科書[1]として執筆し、関連講義[3,5]を行った。

（３） 被験者実験を用いた実証的分析を
進化ゲーム理論を用いた理論的分析の結果明らかになった、いくつかの行動原則について実際に被験者実験をすることでその正当性を明らかにした。具体的には間接互恵性に関する実験的検討を行い、その成果を『社会心理学研究』に発表[7]し、国内会議[16-20]で報告した。また公共財ゲームにおける誘因と結び付けた実験的検討を行い、その成果を PLoS ONE に発表[9]し、国際会議[13]で報告した。

（４）人工知能原則の抽出と一般への還元
これらの研究成果を踏まえていくつかの人工知能原則を抽出した。具体的には「AI が様々な自動判断の適用範囲を拡大させている現状」に焦点を当て、応用分野（結婚や人事採用・自動運転車など）での、直近の技術的動向を踏まえたうえでの、AI と人間の距離感に関する現実的な解を探索した。とくに個別具体的分野について議論することは、一般読者にとっても重要であると考えた。これらのテーマを深堀することで、全体として人間がどこまで決定し、AI にどこまで決定権を渡すべきかについての技術と哲学のバランスの取れた議論になったと考える。
本研究をまとめるにあたって、蝶名林亮博士(哲学・メタ倫理学)を対話者とした対談を行った。章立ては下記のとおりである。最終的にこの内容を書籍化[2]する予定である。

１．はじめに
自動決定は至る所にある・便利だと実用化される・AI はその理由を説明できない・AI と人間のどちらが決定するべきか
２．AI は結婚にどこまで関与してよいのか
AI 婚活の現状と問題提起・結婚は誰が決めるべきと言っているか：道徳的証言論による答え・AI アプリは悪用されないか・AI アプリの予測と結婚がもたらす幸福は無関係？・結局、どこまで AI は介入すべきか・ブラックボックス問題はどう考えるか
３．就職
４．信用スコア
５．自動運転車
６．アルゴリズム的偏見

【著書】
[1] 岡田勇,山本仁志.「社会シミュレーション」『計算社会科学入門』, 丸善出版. pp.189-212, 2021(1)
[2] 岡田勇, 蝶名林亮『(仮題) 人工知能はどこまで決定すべきか:AI をめぐる哲学者と計算社会科学者との対談』(出版準備中)

【招待講演】
[3] 岡田勇. 社会シミュレーション、神戸大学計算社会科学研究センターCCSS『計算社会科学入

門』, 2021(2)
[4] Isamu Okada. Social dilemma, scoring dilemma, and punishment dilemma in indirect reciprocity. A mini-symposium "Evolutionary Game Theory under Uncertainty" at the 2021 Annual Meeting of the Society for Mathematical Biology, 2021(6)
[5] 岡田勇、『計算社会科学入門第８章社会シミュレーション』講義、オンライン教材,2022(3)

【査読あり論文】
[6] Okada I. Two ways to overcome the three social dilemmas of indirect reciprocity. Scientific Reports 10, 16799. 2020(10)
[7] 梅谷凌平、後藤晶、岡田勇、山本仁志、公正世界信念がアップストリーム互恵的協力に与える影響の検討、社会心理学研究 36(2), pp. 31-38. 2020(12)
[8] Okada I, Yamamoto H, Akiyama E, Toriumi F, Cooperation in spatial public good games depends on the locality effects of game, adaptation, and punishment, Scientific Reports 11, 7642. 2021(4)
[9] Hackel J, Yamamoto H, Okada I, Goto A, Taudes A. Asymmetric effects of social and economic incentives on cooperation in real effort based public goods games. PLoS ONE 16(4): e0249217. 2021(4)
[10] Yamamoto H, Okada I, Uchida S, Sasaki T. Exploring norms indispensable for both emergence and maintenance of cooperation in indirect reciprocity. Frontiers in Physics 10:1019422. 2022(9)

【査読なし論文】
[11] 岡田勇. ポストコロナ社会における協力の進化　第 12 回横幹連合コンファレンス 企画セッション「ポストコロナ社会に計算社会科学はいかに貢献するか」. 2021(12)

【国際会議発表論文】
[12] Okada, Isamu. Private assessment of indirect reciprocity changes the landscape of cooperation. A minisymposium "Evolutionary Game Theory under Uncertainty", MMEE Mathematical Models in Ecology and Evolution (MMEE), 2022(7)
[13] Yamamoto H, Hackel J, Okada I, Goto A, Taudes A. Effect of two types of incentives on cooperation: a real effort based public goods game, 19th International Conference on Social Dilemmas, 2022(7)
[14] Isamu Okada, Hannelore De Silva, Kryztoph Paruch, Sending spies as insurance against Bitcoin pool mining block withholding attacks. International Workshop on Distributed Ledgers and Related Technologies (DLRT 2022) on the 33rd DEXA Conferences and Workshops (DEXA2022), 2022(8)

【国内会議発表論文】
[15] 岡田勇. その裏切りには理由があるはずだ―私的観察系間接互恵性の進化ゲームによる網羅的分析―.数理社会学会第 69 回大会, 2020(9)
[16] 山本仁志, 鈴木貴久, 岡田勇, 梅谷凌平. 間接互恵状況に評判ダイナミクスの分析：理論・実験・統合. データ指向構成マイニングとシミュレーション研究会(DOCMAS), 2021(3)
[17] 梅谷凌平, 後藤晶, 岡田勇, 山本仁志. アップストリーム互恵性の規定因　搾取という側面からの検討. 日本社会心理学会第 62 回大会予稿集 2021(8)
[18] 梅谷凌平、山本仁志、後藤晶、岡田勇. 搾取がアップストリーム互恵的協力に与える影響. 社会情報システム学研究会第２８回シンポジウム学術講演論文集, 2022(1)
[19] 梅谷凌平, 山本仁志, 後藤晶, 岡田勇, 秋山英三. 被験者実験によるネガティブアップストリーム互恵性に関する検討. 2022 年社会情報学会大会予稿集, 2022(9)
[20] 梅谷凌平, 山本仁志, 後藤晶, 岡田勇, 秋山英三. 搾取という選択肢がアップストリーム互恵性に与える影響. 日本社会心理学会第 63 回大会予稿集, 2022(9)
[21] 佐々木達矢, 内田智士, 岡田勇, 山本仁志.アップストリーム型とダウンストリーム型間接互恵性の統合モデルのダイナミクス分析. 第 15 回日本人間行動進化学会大会, 2022(12)
[22] 佐々木達矢, 内田智士, 岡田勇, 山本仁志. アップストリーム型とダウンストリーム型間接互恵性の統合モデルのダイナミクス分析. 社会情報システム学研究会第２９回シンポジウム学術講演論文集, 2023(1)
[23] Sasaki T, Uchida S, Okada I, Yamamoto H. Integrated indirect reciprocity and the evolution of cooperation, 第 6 回人工生命研究会, 2023(2)
[24] Sasaki T, Uchida S, Okada I, Yamamoto H. An integrated model of upstream and

downstream reciprocity，ゲーム理論ワーショップ 2023，2023(3)
[25] 佐々木達矢，内田智士，岡田勇，山本仁志．間接互恵性の進化における互恵戦略とフリーライダーの安定共存．数理社会学会第 74 回大会，2023(3)

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　5件／うち国際共著　1件／うちオープンアクセス　7件）

１．著者名 Okada Isamu、Yamamoto Hitoshi、Akiyama Eizo、Toriumi Fujio	４．巻 11
２．論文標題 Cooperation in spatial public good games depends on the locality effects of game, adaptation, and punishment	５．発行年 2021年
３．雑誌名 Scientific Reports	６．最初と最後の頁 7642
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-021-86668-3	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Okada Isamu、Yamamoto Hitoshi、Akiyama Eizo、Toriumi Fujio	４．巻 11
２．論文標題 Cooperation in spatial public good games depends on the locality effects of game, adaptation, and punishment	５．発行年 2021年
３．雑誌名 Scientific Reports	６．最初と最後の頁 7642
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-021-86668-3	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hackel Jakob、Yamamoto Hitoshi、Okada Isamu、Goto Akira、Taudes Alfred	４．巻 16
２．論文標題 Asymmetric effects of social and economic incentives on cooperation in real effort based public goods games	５．発行年 2021年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0249217
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0249217	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Okada Isamu、Okano Nozomi、Ishii Akira	４．巻 10
２．論文標題 Spatial opinion dynamics incorporating both positive and negative influence in small-world networks	５．発行年 2022年
３．雑誌名 Frontiers in Physics	６．最初と最後の頁 953184
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fphy.2022.953184	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yamamoto Hitoshi、Okada Isamu、Uchida Satoshi、Sasaki Tatsuya	4．巻 10
2．論文標題 Exploring norms indispensable for both emergence and maintenance of cooperation in indirect reciprocity	5．発行年 2022年
3．雑誌名 Frontiers in Physics	6．最初と最後の頁 1019422
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fphy.2022.1019422	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 岡田 勇	4．巻 2021
2．論文標題 ポストコロナ社会における協力の進化	5．発行年 2021年
3．雑誌名 横幹連合コンファレンス予稿集	6．最初と最後の頁 A-4-2
掲載論文のDOI（デジタルオブジェクト識別子） 10.11487/oukan.2021.0_A-4-2	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Okada Isamu	4．巻 2020
2．論文標題 Evolution of cooperative study	5．発行年 2020年
3．雑誌名 Impact	6．最初と最後の頁 76～78
掲載論文のDOI（デジタルオブジェクト識別子） 10.21820/23987073.2020.8.76	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　2件／うち国際学会　2件）

1．発表者名 Isamu Okada
2．発表標題 Social dilemma, scoring dilemma, and punishment dilemma in indirect reciprocity
3．学会等名 A mini-symposium "Evolutionary Game Theory under Uncertainty” at the 2021 Annual Meeting of the Society for Mathematical Biology（招待講演）（国際学会）
4．発表年 2021年

１．発表者名 Isamu Okada
２．発表標題 Towards a soft landing for a smart social credit system. Keynote of the symposium
３．学会等名 Towards a soft landing for a smart social credit system. in HICSS-56（招待講演）（国際学会）
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	中井　豊 (NAKAI YUTAKA) (00348905)	関西大学・ソシオネットワーク戦略研究機構・非常勤研究員 (34416)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
Towards a soft landing for a smart social credit system. in HICSS-56	2023年～2023年

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オーストリア	ウィーン経済経営大学			