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　 ６ 月 １９ 日現在

機関番号：３２６８７
研究種目：挑戦的研究（萌芽）
研究期間：2019～2022
課題番号：１９Ｋ２１５７０

研究課題名（和文）人・AI混在環境における人の道徳判断基準の進化の解明

研究課題名（英文）Evolution of human moral judgement criteria in mixed human-AI environments

研究代表者
山本　仁志（Yamamoto, Hitoshi）

立正大学・経営学部・教授

研究者番号：７０３２８５７４

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）：本研究課題では人間の行動メカニズムを明らかにする被験者実験と、人々の行動モデルの理論的妥当性や異なる環境下での振る舞いを予測するためのエージェントシミュレーションを併せて用いることで、理論・実践の両面から総合的に検討をおこなった。本研究課題の最大の目的である人・AI共存下での道徳判断基準の進化についてはこの状況に着目し、様々な状況を想定したシナリオ実験を中心に人・AI共存環境における人の同党判断基準ならびにAIの判断の受容過程について分析した。

研究成果の学術的意義や社会的意義
本研究課題の学術的・社会的意義は課題の速報的重要性と学術的知見の蓄積が未整備であるという現状である。人の道徳判断基準や向社会的行動の要因を探る研究は古くから社会心理学分野を中心に研究が蓄積されてきた。しかし人間より多くの情報を高度に処理する主体が存在するという環境は近年になって人類が初めて直面したものである。そこで本研究課題では人が判断を保留ないし迷う状況下で下される善悪判断が、AIと人間の混在環境下でどのように進化しうるかをシミュレーション、被験者実験を通じて明らかにした。

研究成果の概要（英文）：In the research project, we conducted a comprehensive study from both theoretical and practical perspectives, using subject experiments to clarify the mechanisms of human behavior and agent simulations to predict the theoretical validity of human behavior models and behavior under different environments. The evolution of moral judgment criteria under human-AI coexistence, which is the main objective of this research project, was focused on this situation, and the process of acceptance of human party judgment criteria and AI judgment in human-AI coexistence environments were analyzed, focusing on scenario experiments assuming a variety of situations.

研究分野：社会情報学

キーワード：協力の進化　間接互恵　社会的ジレンマ　社会規範

１．研究開始当初の背景

人と AI が共存することで人間の行動戦略が変化することは経済市場を模した実験においても観察されている[1]。つまり限定合理的な人間だけが存在する社会における行動と、人から見てはるかに合理的な AI が存在する時の行動は異なることが予測できる。急速な技術の発展により経済的判断に限らず、道徳的判断が必要な場面に活用される日は遠くないことが予見できる。申請者は人のみが存在する環境における道徳判断基準の実験[2]を実施する中で AI 混在環境における人の振舞の分析が重要であることに思い至った。

また、AI が社会に普及する過程で人の持つ価値観や評価基準がどのように変容していくのかは社会的要請の大きい喫緊の課題である。一例を挙げると 2018 年現在ではチェスのみならず囲碁・将棋においてもソフトウェアが人間の能力を上回ることは広く受容されているが、2013 年に初めて現役のプロ将棋棋士がソフトウェアに敗北した時にその棋士は非常に強い非難に晒された。これは知的な活動において AI が人の能力を超えていく過程を社会が経験していなかったための拒否反応ともいえる。こうした混乱は AI の利活用が一般社会に広がるにつれ広範かつ深刻に発生することが懸念される。AI が存在する社会において人々が AI の判断のみならず他者に対して下す評価や行動についての知見を学術的に探究する必要性は非常に高いと言える。AI の利活用が進展する過程で生じる混乱と対応について学術的な知見がほぼ存在せず、かつ目前に迫った課題であると考え本研究課題の構想に至った。

２．研究の目的

本研究は人と人工知能(AI)が混在した環境において人の道徳判断基準はいかに進化しうるのかを明らかにする。AI 関連技術の急速な発展により人と AI の共存は世界的に大きな課題となっている。メディアにおいても人間の労働を代替するのかといった議論が多く交わされている。しかしその多くは人の知的処理能力を超える AI と人間の能力を如何に共存させるかという能力的な共存関係が中心的な課題である。AI と道徳に関する研究も一部では進展しているが、例えば MIT で行われているモラルマシンプロジェクト も「AI(自動運転車)の道徳的判断基準はどうあるべきか」という疑問に留まっている。本研究では人と AI が混在する環境で人が持つ道徳判断基準がどのように変化するのかについて議論する。なぜなら人類より高い情報処理能力を持つ存在が下した判断を人がどのように受け入れるのかについて学術的知見が完全に未整備だからである。人のみの環境において相互に他者を評価する際には、人は非常に単純な評価基準を採用することが観察されている 。しかし現代の情報化社会においては人々の過去の行動履歴や人間関係は広範に観測可能となり、それらのデータは利用可能な形で日々蓄積されている。こうした大量のデータとそれを高度に処理する機能が AI として実装される社会で、人々は他者をどう評価するのか、AI やシステムから提供される情報をどのように利用するのか、またどのように行動が変容するのかについては、既存の学問的積み上げからは推論・検討することが困難であり、探索的に幅広く研究を進める必要がある。

本研究課題は AI 時代の善きサマリア人の法を新たに探究するものといえる。善きサマリア人の法とは、善意により誠実に行動した結果の失敗・損失は責任を問われないという考えである。これは人間の限定合理性や情報処理能力等の制約を前提としている。では大量な情報に対して高い処理能力が実装された AI が存在する今後の情報化社会においてどのような「新たな善きサマリア人の法」が求められるのであろうか？本研究ではこの問いに人・AI 混在環境における人

の道徳判断基準の進化というアプローチで答えようとするものである。

協力の進化に関する近年の研究で、これまでの理論的な予測とは異なり人々は正当化される非協力に対しては良い・悪いの判断を避け中立的な態度をとることがわかった。他方で「悪い人を助ける（正当化されない協力）」については良いと判断することがわかった。つまり、正当化される非協力については人々は自身での判断を避けて正当化も不当化もしない。では、正当化される非協力を他者が判断する際には人々はその判断をどのように評価するのであろうか。例えば組織内であれば、上司にあたる人間がある人の行動の善悪を判断し評価する場面が該当しよう。さらにはこうした人の行動に対する判断に AI が導入された場合、受容の度合いは人間が下す判断とどのように異なるのであろうか。この疑問に答えることは、AI が広く実装される社会で、人々は他者をどう評価するのか、AI やシステムから提供される情報をどのように利用するのか、またどのように行動が変容するのかについて理解するために重要な課題である。

３．研究の方法

本研究の目的を達成するために次の 3 つのサブ課題について被験者実験を中心に Web 調査、シミュレーション、ソーシャルメディア分析を援用し探索的に研究する。

1. 人・AI 混在環境において、人はどのような道徳判断基準をもつのか。それは従来の人間のみが存在するときの道徳判断基準とは異なるのか。
2. 人・AI 混在環境において、人々は AI が提供する情報をどのように利用し、人々の行動はどのように変わるのか。

これまでの社会的ジレンマに関する被験者実験の多くは人同士の集団、もしくは人同士と教示するが背後では単純なプログラムが動いているなど、被験者に人間同士の相互作用であることを意識させるものが主流であった。本課題ではそこに AI が混在することを明示することで行動戦略や判断基準がどのように変化するかを分析する(サブ課題 1)。また、AI が用いる戦略や情報を操作し様々なタイプの AI が混在する環境で人々の振舞いがどのように変化するのかを分析する(サブ課題 2)。

４．研究成果

本研究課題では人間の行動メカニズムを明らかにする被験者実験と、人々の行動モデルの理論的妥当性や異なる環境下での振る舞いを予測するためのエージェントシミュレーションを併せて用いることで、理論・実践の両面から総合的に検討をおこなった。

理論モデルの構築においては「規範混在系」のエージェントシミュレーションをおこなった。複数の規範が混在する環境をモデル化し協力と規範の共進化メカニズムを分析した。その結果、協力の進化に必要な規範・協力の維持に必要な規範を明らかにすることができた[1,2]。また、囚人のジレンマにおいてゲームに参加しないという行動選択を可能にした場合の支配戦略を明らかにすることに成功した[3]。

被験者実験においては、人が持つバイアスの特徴が向社会的行動を規定する要因となりうるのかを明らかにするための実験[4,5]や、間接互恵状況下において特定の条件がそろうと人々が他者の行動に対する評価を保留することを明らかにした[6]。本研究課題の最大の目的である人・AI 共存下での道徳判断基準の進化についてはこの状況に着目し、様々な状況を想定したシナリオ実験を中心に人・AI 共存環境における人の同党判断基準ならびに AI の判断の受容過程について分析した[7,8,9]。

社会における実証面では、新型コロナウィルス感染拡大時の様々な行動自粛に関して社会的ジレンマの枠組みで 2020 年から継続的にパネル調査を実施し、他罰的な規範を持つ人々の特徴やメディア接触の影響を明らかにした[10,11]。

[1] 山本仁志. (2019). レギュラーネットワーク上の規範と協力の共進化ダイナミクス. 社会情報学, 8(2), 35-46. https://doi.org/10.14836/ssi.8.2_35
[2] Yamamoto, H., Okada, I., Uchida, S., & Sasaki, T. (2022). Exploring norms indispensable for both emergence and maintenance of cooperation in indirect reciprocity. Frontiers in Physics, 10(September), 1-9. https://doi.org/10.3389/fphy.2022.1019422
[3] Yamamoto, H., Okada, I., Taguchi, T., & Muto, M. (2019). Effect of voluntary participation on an alternating and a simultaneous prisoner's dilemma. Physical Review E, 100(3), 032304. https://doi.org/10.1103/PhysRevE.100.032304
[4] 梅谷凌平, 後藤晶, 岡田勇, & 山本仁志. (2020). 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), 31-38. https://doi.org/10.14966/jssp.1912
[5] Hackel, J., Yamamoto, H., Okada, I., Goto, A., & Taudes, A. (2021). Asymmetric effects of social and economic incentives on cooperation in real effort based public goods games. PLOS ONE, 16(4), e0249217. https://doi.org/10.1371/journal.pone.0249217
[6] Yamamoto, H., Suzuki, T., & Umetani, R. (2020). Justified defection is neither justified nor unjustified in indirect reciprocity. PLOS ONE, 15(6), e0235137. https://doi.org/10.1371/journal.pone.0235137
[7] 山本仁志, 鈴木貴久, AI と人の判断はどちらが受け入れられるのか：間接互恵場面を用いた分析, 日本人間行動進化学会第 15 回大会, 2022
[8] 山本仁志,鈴木貴久, 間接互恵における AI と人間の判断に対する受容の違い, 2022 年度社会情報学会（SSI）学会大会, 2022
[9] 山本仁志, 鈴木貴久, 判断に迷う状況において AI と人のどちらの判断が受容されるか：間接互恵場面を用いた分析, 社会システムと情報技術研究ウィーク (WSSIT2023), 2023
[10] Suzuki, T., Yamamoto, H., Ogawa, and Umetani, R. (2023). Effects of media on preventive behaviour during the COVID-19 pandemic. Humanities and social sciences communications, 10, 58. https://doi.org/10.1057/s41599-023-01554-9
[11] Yamamoto, H., Suzuki, T., Ogawa, and Umetani, R. (2023). The effects of psychological attitudes on voluntary cooperation against COVID-19: an analysis using a social dilemma framework. Journal of Socio-Informatics (In printing)

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　1件／うちオープンアクセス　8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 -
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 -
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 梅谷 凌平、後藤 晶、岡田 勇、山本 仁志	４．巻 36
２．論文標題 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	５．発行年 2020年
３．雑誌名 社会心理学研究	６．最初と最後の頁 31～38
掲載論文のDOI（デジタルオブジェクト識別子） 10.14966/jssp.1912	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、Suzuki Takahisa、Umetani Ryohei	４．巻 15
２．論文標題 Justified defection is neither justified nor unjustified in indirect reciprocity	５．発行年 2020年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0235137
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0235137	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yamamoto H.、Okada I.、Taguchi T.、Muto M.	4．巻 100
2．論文標題 Effect of voluntary participation on an alternating and a simultaneous prisoner's dilemma	5．発行年 2019年
3．雑誌名 Physical Review E	6．最初と最後の頁 32304
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevE.100.032304	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 Toriumi, F., Yamamoto, H., & Okada, I.	4．巻 3
2．論文標題 A belief in rewards accelerates cooperation on consumer-generated media	5．発行年 2019年
3．雑誌名 Journal of Computational Social Science	6．最初と最後の頁 19-31
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42001-019-00049-5	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 山本 仁志	4．巻 8
2．論文標題 レギュラーネットワーク上の規範と協力の共進化ダイナミクス	5．発行年 2019年
3．雑誌名 社会情報学	6．最初と最後の頁 35-46
掲載論文のDOI（デジタルオブジェクト識別子） 10.14836/ssi.8.2_35	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、Uchida Satoshi	4．巻 11
2．論文標題 Hybrid Assessment Scheme Based on the Stern- Judging Rule for Maintaining Cooperation under Indirect Reciprocity	5．発行年 2020年
3．雑誌名 Games	6．最初と最後の頁 13
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/g11010013	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 -
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 Suzuki Takahisa、Yamamoto Hitoshi、Ogawa Yuki、Umetani Ryohei	4．巻 10
2．論文標題 Effects of media on preventive behaviour during the COVID-19 pandemic	5．発行年 2023年
3．雑誌名 Humanities and Social Sciences Communications	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1057/s41599-023-01554-9	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計18件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 山本仁志，鈴木貴久
2．発表標題 AIと人の判断はどちらが受け入れられるのか：間接互恵場面を用いた分析
3．学会等名 日本人間行動進化学会第15回大会
4．発表年 2022年

1．発表者名 佐々木達矢，内田智士，岡田勇，山本仁志
2．発表標題 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
3．学会等名 日本人間行動進化学会第15回大会
4．発表年 2022年

１．発表者名
佐々達矢,内田智士,岡田勇,山本仁志

２．発表標題
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

３．学会等名
第74回数理社会学会大会

４．発表年
2023年

１．発表者名
山本仁志,鈴木貴久

２．発表標題
間接互恵におけるAIと人間の判断に対する受容の違い

３．学会等名
2022年度社会情報学会（SSI）学会大会

４．発表年
2022年

１．発表者名
山本仁志，鈴木貴久

２．発表標題
判断に迷う状況においてAIと人のどちらの判断が受容されるか：間接互恵場面を用いた分析

３．学会等名
社会システムと情報技術研究ウィーク（WSSIT2023）

４．発表年
2023年

１．発表者名
佐々木達矢，内田智士，岡田勇，山本仁志

２．発表標題
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

３．学会等名
日本人間行動進化学会第15回大会

４．発表年
2022年

１．発表者名
山本仁志，鈴木貴久

２．発表標題
AIと人の判断はどちらが受け入れられるのか：間接互恵場面を用いた分析

３．学会等名
日本人間行動進化学会第15回大会

４．発表年
2022年

１．発表者名
梅谷凌平，山本仁志，後藤晶，岡田勇，秋山英三

２．発表標題
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

３．学会等名
2022年度社会情報学会（SSI）学会大会

４．発表年
2022年

１．発表者名
山本仁志,鈴木貴久

２．発表標題
間接互恵におけるAIと人間の判断に対する受容の違い

３．学会等名
2022年度社会情報学会（SSI）学会大会

４．発表年
2022年

１．発表者名
梅谷凌平，小川祐樹，鈴木貴久，山本仁志

２．発表標題
メディア接触が新型コロナウィルス感染症の感染拡大状況における意見の形成に与える影響の分析

３．学会等名
2021年社会情報学会学会大会

４．発表年
2021年

１．発表者名
鈴木貴久，山本仁志、小川祐樹、梅谷凌平

２．発表標題
コロナ禍における外出自粛に対するメディアの効果

３．学会等名
第28回社会情報システム学シンポジウム

４．発表年
2022年

１．発表者名
梅谷凌平，山本仁志，後藤晶，岡田勇

２．発表標題
搾取がアップストリーム互恵的協力に与える影響

３．学会等名
第28回社会情報システム学シンポジウム

４．発表年
2022年

１．発表者名
山本仁志，鈴木貴久，小川祐樹，梅谷凌平

２．発表標題
コロナ禍における向社会的行動の規定因：2時点パネル調査による分析

３．学会等名
Workshop of Social System and Information Technology (WSSIT2022)

４．発表年
2022年

１．発表者名
梅谷凌平，山本仁志

２．発表標題
資源の多寡と協力行動 －囚人のジレンマを用いた分析－

３．学会等名
2020年社会情報学会学会大会

４．発表年
2020年

１．発表者名
山本仁志，小川祐樹，鈴木貴久，梅谷凌平

２．発表標題
新型コロナウィルスによる外出自粛の規定要因：社会的ジレンマの枠組みを用いた分析

３．学会等名
2020年社会情報学会学会大会

４．発表年
2020年

１．発表者名
吉田圭太，梅谷凌平，山本仁志

２．発表標題
公共財ゲームにおいて罰の強度の非対称性が協力に与える効果

３．学会等名
第26回社会情報システム学シンポジウム

４．発表年
2020年

１．発表者名
梅谷凌平，山本仁志

２．発表標題
囚人のジレンマにおいて資源の多寡が相手選択と協力行動に与える影響

３．学会等名
第26回社会情報システム学シンポジウム

４．発表年
2020年

１．発表者名
梅谷凌平，山本仁志

２．発表標題
間接互恵状況において異なる社会階層に対して期待する規範

３．学会等名
日本社会心理学会第60回大会

４．発表年
2019年

〔図書〕　計1件

１．著者名 鳥海不二夫編著（第8章担当）	４．発行年 2021年
２．出版社 丸善出版	５．総ページ数 322
３．書名 計算社会科学入門	

〔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

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0件

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