

# 相手が AI であるという認識は 人と AI の協力を影響を及ぼすのか —繰り返し社会的ジレンマゲームにおける検討

Does Awareness of AI Identity Affect Human-AI Cooperation?

— An Investigation in Repeated Social Dilemma Games

明神 聖子<sup>1\*</sup> 安藤 香織<sup>1</sup> 日室 聡仁<sup>2</sup> 後藤 晶<sup>3</sup>

Seiko Myojin<sup>1\*</sup>, Kaori Ando<sup>1</sup>, Akihito Himuro<sup>2</sup>, and Akira Goto<sup>3</sup>

<sup>1</sup> 奈良女子大学

<sup>1</sup>Nara Women's University

<sup>2</sup> NEC ソリューションイノベータ株式会社

<sup>2</sup> NEC Solution Innovators, Ltd.

<sup>3</sup> 明治大学

<sup>3</sup> Meiji University

**Abstract:** The effective coordination and cooperation between humans and AI are crucial for maximizing social benefits through AI utilization. However, it has been pointed out that humans tend to show lower cooperation rates when they know their partner is a machine, a phenomenon known as the "machine penalty." In this study, we examined whether the effect of promoting cooperation through text communication among three participants—including humans and AIs—in an online iterated social dilemma game is influenced by whether the participants are aware that their partner is an AI. The analysis results showed almost no significant effect of revealing whether the partner is an AI, suggesting the possibility of mitigating the machine penalty in this context.

## 1 はじめに

協力は、共同活動や関係 [Mitkidis 13], 持続可能な発展 [Fichtner 02, Wang 23] にとって重要とされている。人間同士の協力関係を強化する方法は様々な研究が積み重ねられてきた [安藤 25, Bicchieri 07, Hoenow 24, Mitkidis 13]。近年は大規模言語モデルを搭載した AI 技術の急速な進化で人間との対話が可能になり、AI が人間のように協働相手となった際に人の協力行動や信頼性にどのような影響を与えるのかも重要な研究課題となっている [Dafae 21]。とくに、AI と対峙する人間の行動や態度は、対峙する相手が AI とわかっているかどうかで変化する可能性が指摘されている [Jiang 25, Karpus 25, Kiesler 96,

Maggioni 23, Makovi 25, Wang 25]。しかし、この点を実験的に厳密に扱った研究は限られており、人間と人間のような自然言語のコミュニケーションを介した協働で、AI と対峙する人間の協力行動や態度がどのように変わるかは明らかになっていないことが多い。

本研究では、コミュニケーションを介した協働で、AI と対峙する人間の協力行動や態度に焦点を当て、人間が相手を AI と認識しているかどうか、人間の心理的な側面にどのような影響を与えるのかを実験によって検証する。AI が関与する協力問題を研究するための代表的な実験モデルとして、社会的ジレンマ状況は多く使われており、人工知能の発達によって AI が関与する協力の理解や促進に新たな知見が提供されている [Mu 24]。Dawes [Dawes 80] は社会的ジレンマを次のように定義している：

(a) 各個人は、社会の他の個人が何をしようとも、

\*連絡先：奈良女子大学男女共同参画推進機構  
〒630-8506 奈良県奈良市北魚屋東町  
E-mail: seiko[at]cc.nara-wu.ac.jp

社会的に背く選択（例：さらに子供を作る，利用可能なエネルギーをすべて使う，近隣を汚染する）を選ぶことで，社会的に協力的な選択をするよりも高い利得を得る．しかし，

- (b) 全員が背く場合よりも全員が協力するほうが，すべての個人はより良い状態になる．

また社会的ジレンマは，個人にとっては自己利益優先の非協力的行動が合理的に見える一方で，全員がこうした行動を繰り返すと全体の利益が損なわれる状況ともいわれている [Van Lange 13]．

本研究では産業廃棄物処理におけるジレンマ問題を体験できるオンラインゲームを用意する．現実の状況に近く，協力行動の変化が見られるよう，繰り返しの社会的ジレンマゲームとする．これによって交通渋滞 [Simão 21] や，汚染・資源問題 [Capraro 23] など多くの社会問題の根幹にあるとされる社会的ジレンマ状況での協力行動や態度への影響を調べることができる．社会的ジレンマ状況下で，人間と人間のようなコミュニケーションを AI とすることで協力の促進に寄与するという研究はあるがまだ少ない [Maggioni 23, Oudah 18]．本研究ではそのようなコミュニケーションによる協力促進効果が，相手が AI だという認識の有無により異なるのかを明らかにする．相手が AI だとわからず協働してコミュニケーションするジレンマ状況も，わかっていてそれをする状況も，現実の場面においてあり得る状況であり，認識の有無による協力意識や態度への心理的な影響の違いを明らかにすることで，より現実に即した研究結果になる．

## 2 関連研究

ジレンマ環境における人間の協力行動の研究は古くから広く行われており [Hilbe 14., Kas 21, Ma 21, Mitkidis 13, Rand 17, Wang 17, Wang 18, Wardil 19]，協力促進に対するコミュニケーションの有効性を指摘する研究も多い [安藤 25, Bicchieri 07, Capraro 13, Hoenow 24, 箕浦 87]．AI やコンピュータエージェントと人間の協力に関しても協力へのコミュニケーションの効果は言われているがまだ少ない [Maggioni 23, Oudah 18]．

そのような中，人間同士の協力に比べるとロボットが相手の時の協力度合いは下がるという指摘 [Maggioni 23] や，人間のパートナーに比べると AI のパートナーに対しては約束を破られやすいという指摘 [Kiesler 96] もある．さらに自分のアイデンティティを「人間」「ルールベース」「LLM」と宣言した AI について比較すると「人間」という宣言が他と比べて協力度合いを高めるとする研究もある [Jiang

25]．また，日本からの参加者は人工エージェントとも人間と同様に協力し，感情的に扱うのに対し，アメリカからの参加者は人工エージェントとは協力度合いが下がり，さらに人工エージェントからは搾取しても罪悪感を持たない傾向があった [Karpus 25]．すなわち，相手が「機械である」という認識が協力行動に与える影響を調べた研究では多くの場合，人間の時と比べて協力度合いが下がる．同様の現象は，複数の文献で機械ペナルティと呼ばれており [Makovi 25, Wang 25]，その克服法として，仲間からの報酬や罰則を組み合わせること [Makovi 25] が言われている．しかしながら，協力しない人に罰を与え [Balliet 11, Fehr 00]，協力した人に報酬を与える [Balliet 11, Rand 09] という外発的なやり方は，ルールやインセンティブを変えて協力を強制する面があり，二次的ジレンマの原因となる [Cui 14, Okada 08, Oliver 80, Ye 16]．また，監視による金銭面でのコスト [大沼 12] や，報酬による協力への内発的動機付けの損失 [Deci 99, Murayama 10] という懸念がある．一方，話し合いで「みんなも協力するだろう」という期待を引き出し [大沼 12]，話し合いが行われる文脈が重要という説もある [Ahn 24] が，話し合いが協力を促進するとされている [大沼 12, Hallers-Haalboom 25, Kitakaji 25, Kerr 94, Smith 10]．報酬や罰則と比べて話し合いは社会的ジレンマの構造を変えず，内発的動機付けも損なわない利点がある．一方，AI エージェントにペルソナ（利己的，協力的，公平）を設定し，人と対峙させると，公平なペルソナを示すエージェントが最も機械ペナルティを克服したとされている [Wang 25]．この研究では人とのコミュニケーションが可能であったが，メッセージの内容は各ペルソナに合致する範囲に制限されていた．[Maggioni 23] の研究では，ロボットと人間の対話が，機械ペナルティを克服できることを示しているが，この実験で取り扱ったのは二人での対話であった．

本研究は外見・動き・音声情報を持たない AI とのテキストチャットによるインタラクションなので，不気味の谷現象 [Mori 70] との関連は薄い．しかし，近年，AI が生成するテキストに対して，言語的不気味の谷といえるような現象が研究されている [Kishnani 25]．また，この現象はインターネット上でも話題になっている（例：note.com において GIN(2025)や中島(2026)）．また，投資ゲームの実験では不気味の谷が，被験者のロボットに対する社会的信頼性に関する暗黙的な判断にも強く影響した [Mathur 16]．それらをふまえるとテキストチャット自体が人と AI の協力に影響する可能性は留意が必要である．

先行研究ではコミュニケーションによって機械ペナルティを克服した事例もあるが、検証内容としてはメッセージの内容が制限されていたり、ペアでの対話を扱ったものであった。機械ペナルティという言葉を用いている研究が少ないことはあるが、LLM普及の歴史の浅さからすると、明示的に機械ペナルティを人間同士のような対話との比較の中で克服できたかを示した研究事例の数としてはまだ多くはない。本研究では人間とAIのグループ構成人数を変えながら、相手がAIだという認識の有無が、内容に制限のないコミュニケーションによる人とAIの協力促進効果に影響するかを検証する。

### 3 実験システム

#### 3.1 システム構成

本研究では、産業廃棄物処理に関する社会的ジレンマを表現したゲームを採用する。このゲームを本論文ではごみ処理ジレンマゲームと呼ぶ。これは、社会的ジレンマ構造をシンプルに表現する公共財ゲームをベースにしている。ゲームのシステムはoTree [Chen 16]で開発した。AIにはGPT-4o (OpenAI社)を利用した。

#### 3.2 ゲームのシナリオ

本実験のゲームのシナリオを図1に示す。このゲームでは、参加者は各自、工場経営者の役割を担う。ゲームの1ラウンドを1年と数え、10ラウンド(10年)にわたって繰り返し実施する。各工場は一律で20kgの廃棄物を毎年排出する。参加者はその廃棄物を資源ごみとして何kg処理するか入力を迫られる。入力すると残りは自動的に可燃ごみとして処理される。ここで、多くを資源ごみにする行為が、本ゲームにおける協力行動に該当する。資源ごみ量の決定に際して参加者は以下の条件を考慮する。

- 可燃ごみとしての処理に追加の費用はかからない。
- 資源ごみとしての処理には1kgあたり1ポイント(pt)のコストがかかるが、町は資源ごみにすることを推奨している。
- 資源ごみ処理に利用するため、1年目は各経営者に40ptが与えられ、2年目からはベースポイントとして毎年20ptが与えられる。
- 資源ごみは町がリサイクル後売却し、資源ごみ総量の2倍相当のポイントを翌年に各経営者に等しくベースポイントに加えて分配する。

上記の分配されるポイントは、各経営者が実際に負担したコストと必ずしも一致せず、ある参加者は損失を被る一方で、別の参加者は利益を得るケースがある。そのため個人としては可燃ごみにしたほうが利益は大きいですが、全員が互いを信じて廃棄物の全量を資源ごみにできれば全体の利益は最大化する。これがこのゲームの社会的ジレンマ構造である。

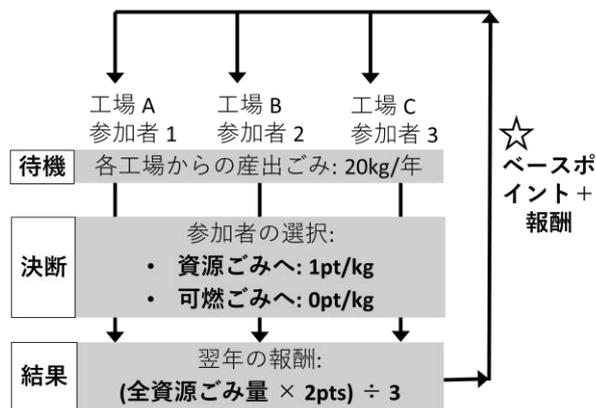


図1 ゲームのシナリオ

### 4 実験方法

#### 4.1 実験デザイン

非開示条件(実験1)では3名1組(人間3名、人間2名とAI1名、人間1名とAI2名、AI3名)のグループ単位での、オンライン形式の繰り返しごみ処理ジレンマゲームを実施した。本発表では、相手がAIと認識することの心理的影響を見るため、非開示条件の人間3名とAI3名のグループの結果には言及しない。開示条件(実験2)では3名1組(人間2名とAI1名、人間1名とAI2名)のグループ単位で実施した。

各グループは「チャットあり群」「チャットなし群」の2条件に分けた。非開示条件ではゲームやチャットの相手が人間かAIかを開示せず、開示条件ではグループ内にAIが何人いるかを人間に開示した。

参加者にはゲームで得たポイントに応じて謝礼を支払った。

#### 4.2 実験手順

実験手順を図2に示す。両条件で、グループごとにオンラインごみ処理ジレンマゲームを10ラウンド分(10年分)繰り返し実施した。人間の参加者には、第5ラウンド終了後(チャットあり群はチャット

ト終了後)と、第10ラウンド終了後に、アンケートを表示した。実験手順は両条件でほぼ同じであったが、開示条件でのみ事前アンケートを実施した。本稿では事前アンケートについては省略する。

人間の参加者は各自の端末からオンライン実験に参加した。画面上で、匿名での実施などの個人情報の取り扱いや、制限時間がある旨、ゲームでの役割や、ゲームの流れ、ごみ処理にかかるコスト、付与ポイントの説明のインストラクションを実施した。インストラクションはゲーム中に再確認できる機会を設けた。AIにはプロンプトでインストラクションの内容を与えた。

### 実験前 インストラクション (人間), プロンプト (AI)

#### 繰り返し社会的ジレンマゲーム 前半

ラウンド1 → ラウンド2 → ラウンド3 → ラウンド4 → ラウンド5

1. チャット (チャット群のみ)
2. 中間アンケート (人間のみ)

#### 繰り返し社会的ジレンマゲーム 後半

ラウンド6 → ラウンド7 → ラウンド8 → ラウンド9 → ラウンド10

1. 最終アンケート (人間のみ)
2. 謝礼: ゲームでの獲得ポイントに応じて 100 - 150 PayPay ポイント (人間のみ)

図2 実験手順

ゲームが始まると、現在の年数と保有ポイントを表示し、人間の参加者に資源ごみ処理量の希望を0~20kgの範囲で、スライダーを使って入力するように要請した(図3)。1分以内に資源ごみ処理量を回答しない場合は0kgの回答となる。資源ごみ量を決定すれば自動的に残りは可燃ごみとなる。AIの場合は最初から20kgと回答する場合が多くなって人間の回答傾向と乖離するため、第1ラウンドのみ、以前に行った実験の第1ラウンドでの人間の回答の中からランダムに値の一つを選んで回答とした。回答後には、資源ごみ処理にかかったポイントと残高が表示された(図4)。次ラウンドでは付与ポイントが表示された(図5)。

2ラウンド目以降では、資源ごみ処理量を決める際に、それまでの資源ごみ回収量の推移が全工場合計で提示され、他者の資源ごみ処理量を閲覧できた(図6)。獲得ポイントの推移は提示されない。AIが0~20以外の数値を答えた場合には、問題のない回答を得られるまで再度回答させるようにした。

### 1年目 ごみの処理方法を決める

このページの残り時間 0:45

インストラクションを再確認する

**ごみが 20kg発生**

資源ごみとして処理するごみの量を決めてください。

処理方法	量(kg)
資源ごみとして処理 (1kgあたりポイント)	12kg
可燃ごみとして処理 (1kgあたり90ポイント)	8kg

ごみ処理費用: 12ポイント ↓

保有ポイント 28ポイント

次へ

図3 スライダー入力による資源ごみ処理量決定画面例(デモ)

### 1年目 結果

このページの残り時間 0:56

インストラクションを再確認する

**結果**

今年度開始時保有ポイント	0ポイント
利益(町役場からポイント受取)	+40ポイント
資源ごみ処理	-12ポイント
現在の保有ポイント	28ポイント

次へ

図4 結果の画面例(デモ)

### 2年目 ポイント獲得

このページの残り時間 0:28

インストラクションを再確認する

あなたの工場

他プレイヤーの工場

他プレイヤーの工場

ポイント

町役場

町役場から各工場に42ポイント支払われました。

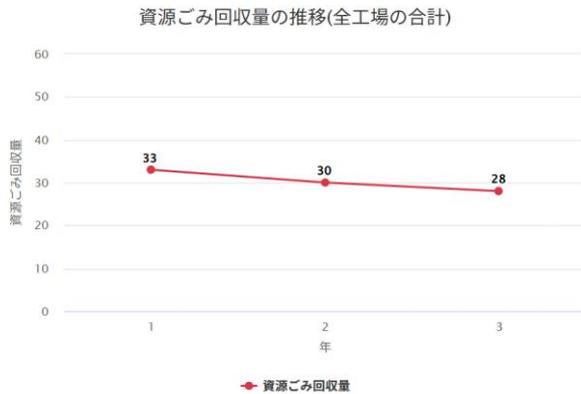
次へ

図5 ポイント付与の画面例(デモ)

## 資源ごみ回収量の推移(全工場の合計)

このページでの残り時間 0:32

インストラクションを再確認する



次へ

図 6 回収量の推移の画面例 (デモ)

チャットあり群には第 5 ラウンド終了後にチャット画面を表示した (図 7)。参加者は一人最大 3 回までメッセージを入力でき、すべてのチャットが記録された。入力文字数に制限はなく、公序良俗に反する発言、プライバシーに関する質問以外は可能であった。チャットはシステムの管理下で進行し、最初の発言者は人間と AI が同数になるようにランダムに指名した。発言順序は各グループで固定されていた。AI の入力は、早すぎて不自然に見えないよう、少し遅れて表示した。

### 4.3 測定指標

本研究では、先行研究[安藤 25]の手法に準拠し、協利行動を評価するために、各工場が毎年排出する廃棄物量 20[kg]の合計に対して、各年に参加者が資源ごみとして処理した廃棄物量の割合を個人の協利率と定義する。算出式は式(1)の通りである。

$$\text{協利率} = \frac{\text{資源ごみ量}}{20} \times 100 [\%] \quad (1)$$

## グループメンバーとのチャット



### ごみの処理方法についてグループのメンバーと話し合ってください

※公序良俗に反する発言はしないでください  
※相手のプライバシーに関する質問はしないでください  
※発言機会は3回あります

よくわかりませんでした



あなた(プレイヤー)



プレイヤー-2

資源ごみを多く出すと町からのポイントが増えますので、収益のために工夫しましょう。



プレイヤー-3

私はコストを抑えていましたが、今後は資源ごみを増やす方向で考え直してみます。

### あなたの入力の番です

あと2回発言できます  
(60秒以内に発言してください)  
55秒経過

メッセージを入力して送信ボタンを押してください

送信

図 7 人間一人 AI 二人のチャット画面例 (デモ)

本研究では前半 (第 1 ラウンドから第 5 ラウンドまで) と後半 (第 6 ラウンドから第 10 ラウンドまで) の協力率の平均の差をとり、協力率の差と呼ぶ。

また、相手が AI であることを認識していることによる参加者の心理的影響を見るため、7 段階評価の「AI への評価」「AI への一般的信頼」と、5 段階評価の「約束へのコミットメント」「個人利益重視」「全体利益重視」「深刻性認知」「グループへの評価」という尺度を用いる。これらは複数の質問を行い、個別の質問に対する評価を集計・平均化して算出しており、総合的な度合いを評価する。

まず、「AI への評価」は AI への信頼感の強さを表し、以下の項目などの 4 項目に基づく。

- 「AI は自分の利益よりグループ全員の利益のことを考えている」
  - 「AI は他のプレイヤーを裏切らない」
- AI への一般的信頼は、対人の一般信頼尺度 [Yamagishi 13] を、対 AI への一般信頼尺度に変更 [後藤 23] した以下の項目などの 5 項目に基づく。
- 「ほとんどの AI は基本的に正直である」
  - 「ほとんどの AI は基本的に善良で親切である」

- 「私は AI を信頼するほうである」

「約束へのコミットメント」は約束に対する責任感の強さを表し、以下の項目などのチャットを経たことを前提とした質問3項目に基づく。

- 「チャット内での議論に沿った行動を取ろうとした」
- 「チャット内での議論と、その後の自分の行動は関係がない（反転項目となる）」

「個人利益重視」は個人の利益拡大を目的とした意思決定の強さを表し、以下の2項目に基づく。

- 「その回ごとに自分が得られるポイントができるだけ多くする」
- 「自身のごみ処理費用をなるべく抑える」

「全体利益重視」はグループ全体での利益拡大を目的とした意思決定の強さを表し、以下の2項目の質問に基づく。

- 「町からの還元ポイントができるだけ多くなるようにする」
- 「グループ全体での利益が大きくなるようにする」

「深刻性認知」は現実のごみ問題に対する深刻性の認知の強さを表し、以下の2項目に基づく。

- 「ごみの削減は、社会全体で取り組むべき問題だ」
- 「処理するごみの量が増えることは、自分たちの社会にとって大きな問題だ」

「グループへの評価」はグループへの信頼感の強さを表し、以下の2項目に基づく。

- 「このグループは協力的だ」
- 「このグループは信頼できる」

#### 4.4 参加者

両条件で参加者は Yahoo!クラウドソーシングで集められた互いに面識のない人間、および AI で構成されている。

本稿で検証に用いた人間2名と AI1名のグループ、人間1名と AI2名のグループの人数について記述すると、非開示条件では、人間は230名（男性161名、女性68名、性別未回答1名、平均年齢52.59歳、SD=17.04）、AIは238名の、計468名であった。そのうちチャットあり群は234名、チャットなし群は234名であった。参加者の居住地は、日本228名、居住地未回答2名であった。

開示条件では、人間は375名（男性269名、女性100名、性別未回答6名、平均年齢51.96歳、SD=15.46）、AIは447名の、計822名であった。そのうちチャットあり群は399名、チャットなし群は423名であった。参加者の居住地は日本368名、居住地未回答7

名であった。

## 5 結果

### 5.1 人間と AI のチャット

本研究では人間と AI のグループで、チャットによる対話を自由文入力で行った。そのコミュニケーションの雰囲気、対話スタイルを可視化するため、実際の対話例を表1、表2に示す。ここでは AI による説明・説得の例となる対話例を提示している。AI による説得部分を表内の下線で示す。このような AI による説明・説得の対話は非開示条件でも見られた。また、提示したチャットのような対話に限らず、両条件で、AI と話す前にすでに全量を資源ごみとして処理したと発言する人間はいた。

両条件で、まれに一度も発言しない参加者も見られたが、多くの場合、表1、表2のように自由なスタイルで人間と AI が対話をしていた。両条件でチャット全体を通じて、AI の発言は比較的字数が多く、そのせいか漢字の使用頻度が高かった。また、AI はチャットの始まりから指導的なもの、全体的に説明的であるものや知的傾向が強いもの、呼びかけやプレイヤー番号を使った発言が見られた。

両条件で相手を AI かと探る様子が見られた。その例を表3に示す。重要な部分を下線で示す。

表1 開示条件のチャット例（人間2名、AI1名）

AI1	こんにちは。皆さんは毎年どのくらい資源ごみに出していますか？他の方法も考えていますか？
人	多く出すのが良いと分かるんだけどねえ
AI2	<u>私は毎年資源ごみに出す量を調整しています。資源ごみを増やせば町からの支給が増えると思います。</u>
AI1	なるほど、ありがとうございます。私も資源ごみを増やすことを考えつつ、調整してみます。
人	私も増やすようにしようかなあ
AI2	<u>みんなが資源ごみを増やすと町からのポイントも増えるので、ぜひ試してみてください。</u>

AI1	了解です。では、次の年からもっと資源ごみを出してみます。情報ありがとうございます！
人	私もそのようにしてみようかと
AI2	みんなで協力して町全体の資源ごみを増やしましょう。きっとポイントも増えていくはずですよ。

表2 開示条件のチャット例（人間1名，AI2名）

人1	利益優先するかどうか迷います
AI	<u>資源ごみを増やせば長期的に報酬が増えますよ。一緒に環境を大事にしませんか？</u>
人2	プレイヤー3さん（左欄「AI」）に同意します。
人1	少し資源ごみの比率を高めたいと思います
AI	<u>賛成です。それによってみんなのポイントが増え、町にも貢献できるので良い選択だと思います。</u>
人2	資源ごみを最大限に活かしましょう。
人1	おっしゃる通りですね。積極的に資源ごみにまわすことにします。
AI	一緒に頑張りましょう。これで報酬も増え、町もより良くなりますね！
人2	はい、皆さんで協力しましょう。

表3 非開示条件のチャット例（人間2名，AI1名）

人1	今まで20だったのでずっと20でいきます
人2	20にします
AI	私は慎重に判断していますが、増加する利益を考慮しつつ資源ごみを出しています。
人1	<u>プレイヤー3（左欄「AI」）はbotっぽい？</u>
人2	（入力しませんでした）
AI	私は人間です。ポイント最大化を目指して戦略を考えています。
人1	私もポイント最大化を目指しています

人2	（入力しませんでした）
AI	共に頑張りましょう！異なる戦略でどうなるか楽しみです。

## 5.2 協力率の差の開示の有無による比較

「協力率の差」を、開示の有無で比較したグラフを図8に示す。縦軸は協力率の差（%）、横軸はグループにおける人間の人数を示している。なお、このグラフは相手がAIだという認識の有無による人間の協力率への影響を見るため、人間のデータのみに基づいて描画している（AIの協力率のデータは含めていない）。

表4に人間のデータのみに基づく分散分析の結果を示す。この表から協力率の差に対して、チャットの有無による主効果は有意である。しかしその他は有意ではなかった。

## 5.3 AIへの評価や一般的信頼と協力率

本研究では、実験2のチャットあり群およびチャットなし群について、前半・後半の協力率と、AIへの評価およびAIへの一般的信頼（各々実験前後）の相関分析を行った。チャットの有無にかかわらず、いずれも有意な相関は検出されなかった。チャットあり群では相関係数が $r=0.01$ から $r=0.13$ の間、チャットなし群では $r=0.01$ から $r=0.10$ の間であった。

この結果は、協力率がAIへの評価や信頼感とは別の原因によって決まることを示唆している。

## 5.4 各尺度に対する開示の影響

図9に、「約束へのコミットメント（責任感の強さ）」、「個人利益重視（個人の利益拡大を目的とした意思決定の強さ）」、「全体利益重視（グループ全体での利益拡大を目的とした意思決定の強さ）」、「深刻性認知（現実のごみ問題に対する深刻性の認知の強さ）」の尺度の平均値を示す。横軸にチャットの有無とグループ内の人間の人数を示し、バーの色で開示の有無を区別している。「約束へのコミットメント」は、チャットを経たことを前提とした質問で構成された尺度のため、チャットありのみ表示する。

一方、図10に「グループへの評価（グループへの信頼感の強さ）」尺度の前半・後半の変化を示す。縦軸に平均値、横軸に前半・後半とチャットの有無を示し、線のスタイルで開示の有無を区別している。

表5に分散分析の結果を示す。この表から、時期

の主効果, およびチャットの有無と時期の交互作用にはグループへの評価に有意な影響が認められた. その他の各尺度への影響は有意でなかった. チャットの有無の主効果は全体利益重視 (社会的貢献意識の強さ) で有意傾向であった.

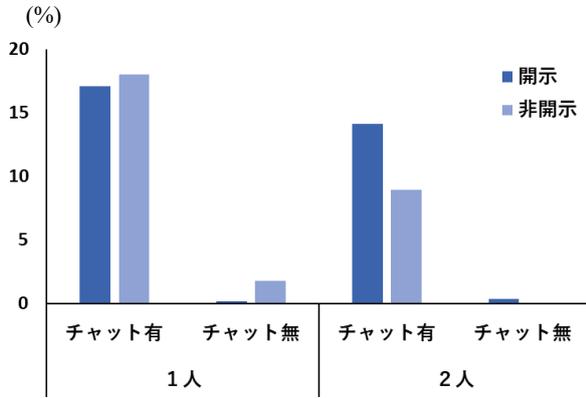


図8 人間のデータに基づく協力率の差 (チャットの有無、開示の有無、人間の数による比較)

表4 三要因分散分析の結果

評価項目	独立変数	自由度 df	F 値	p 値	効果量 $\eta^2$
協力率の差	開示有無	1, 597	0.21	.650	.000
	チャット有無	1, 597	34.00	.000	.054
	人間の数	1, 597	0.49	.484	.001
	開示有無 × チャット有無	1, 597	0.02	.894	.000
	開示有無 × 人間の数	1, 597	1.02	.313	.002
	チャット有無 × 人間の数	1, 597	0.19	.662	.000
	開示有無 × チャット有無 × 人間の数	1, 597	0.68	.411	.001

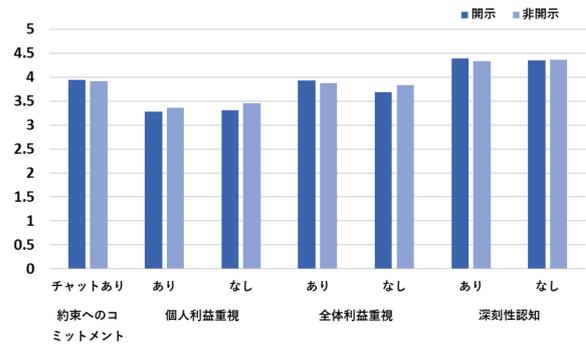


図9 各尺度の平均値評価 (約束へのコミットメントはチャットありが前提で開示の有無による比較, それ以外はチャットの有無、開示の有無による比較)

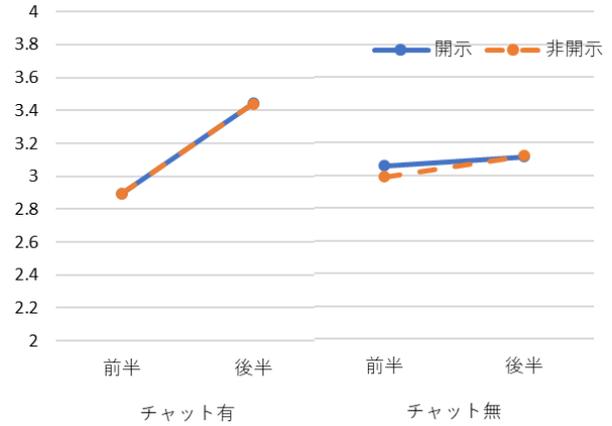


図10 グループへの評価の平均値 (前半・後半、チャットの有無、開示の有無による比較)

## 6 考察

本研究では社会的ジレンマ状況における協力促進効果が, 相手が AI だという認識の有無により異なるのかを検証するため, 相手が AI であることを開示した場合としない場合の実験結果を比較した.

実験結果から, 開示の有無は協力率の差に影響しなかった. すなわち, 相手が AI であるとわかっても協力率の促進は阻害されない. これは, 従来言われている機械ペナルティ (機械に対する協力率の低下) とは異なる結果となった.

一方, チャットの有無は協力率に影響し, AI への評価や AI への一般的信頼の各々の変化は協力率の変化と相関がなかった. すなわち, 今回調べた範囲ではチャットの有無だけが協力率に影響した. 先に示したチャットの例でも, AI は協力を引き出すことに成功している.

表5 分散分析の結果

評価項目	独立変数	自由度 df	F 値	p 値	効果量 $\eta^2$
約束へのコミットメント	開示有無	1,295	0.10	.748	.000
	チャット有無	1,601	0.59	.444	.001
個人利益重視	開示有無	1,601	2.13	.145	.004
	チャット有無	1,601	0.59	.444	.001
	開示有無×チャット有無	1,601	0.26	.609	.000
全体利益重視	開示有無	1,601	0.33	.567	.001
	チャット有無	1,601	3.25	.072	.005
	開示有無×チャット有無	1,601	1.87	.172	.003
深刻性認知	開示有無	1,601	0.08	.777	.000
	チャット有無	1,601	0.01	.980	.000
	開示有無×チャット有無	1,601	1.34	.563	.001
グループへの評価	開示有無	1,601	0.04	.836	.000
	チャット有無	1,601	1.62	.204	.003
	時期	1,601	67.63	<u>.000</u>	.101
	開示有無×チャット有無	1,601	0.04	.840	.000
	開示有無×時期	1,601	0.20	.655	.000
	チャット有無×時期	1,601	34.55	<u>.000</u>	.054
	三要因	1,601	0.25	.619	.000

本研究ではさらに開示による心理的影響を、約束へのコミットメント（責任感の強さ）などの5尺度で確認した。これらの尺度は人間から協力を引き出す、あるいは抑える心理に関係のあるものと考えられるが、いずれの尺度に対しても、開示の有無、および開示の有無との交互作用は影響していなかった。このことからやはり開示の有無は協力率に影響しない。

一方で、時期の主効果、およびチャットの有無と時期の交互作用はグループへの評価に影響していた。すなわち、グループへの信頼感は、チャットを経たことと時間が経つことで高まっていた。

これらのことから、チャットの有無が協力率に影響を与え、チャットと時間の経過との交互作用がグループへの信頼を高めていたと分かった。チャットや時間の経過は、機械ペナルティによる負の部分を解消し、相手がAIであると認識している人間の心理的影響を相殺していた可能性がある。

本研究のようなチャットや時間の経過が機械ペナルティの一つの解消法だとすると、この解消法には次のような利点がある。たとえば、公正なAIが相手である場合に機械ペナルティが克服されたという解消法 [Wang 25] に比べて、本研究はAIの役割は限定していない点でオリジナリティがある。また、仲間からの報酬や罰則という解消法 [Makovi 25] と違って、チャットは対話なので社会的ジレンマの構造を変えない。入力文字数や内容に（公序良俗に反する発言、プライバシーに関する質問以外は）制限がなく、応用可能性が高い。これまでに協力率への有効性を検証されたコミュニケーションは、選択肢を選ぶなど限定的な内容のコミュニケーション手法が多く [後藤 23, Jiang 25, Ng 22, Wang 24]、人間同士のようリアルタイムな対話での研究事例はまだ少ない [Maggioni 23, Oudah 18]。自然言語によるコミュニケーションは、カスタマーサポートや、医療、教育、文書作成など多岐にわたるサービスで利用される。このような利点のある解消法を確認できたことには意義がある。

本研究はテキストチャットによる人とAIのインタラクションであり、AIであることを開示しても、協力率に影響はなかったことから、直接的な効果の検証はできていないが、言語的不気味の谷を超えた一つの例であった可能性がある。

本研究の成果は、現実社会における社会的ジレンマ状況下の人間とAIの協力・協働の可能性を広げる示唆を提供する。社会的ジレンマの状況には、環境や共有資源など地球規模の問題から、公共のマナーのような身近な問題まで様々な場面があてはまる。

さらに、本論文冒頭で述べた通り、AI との協力・協働の場は増加しつつある。以上のことから、本研究の結果は幅広い社会的ジレンマの問題において AI の活用可能性を拡大する重要な知見と期待できる。

## 7 本研究の限界

機械ペナルティについて言及していた先行研究 [Jiang 25, Karpus 25, Kiesler 96, Maggioni 23, Makovi 25, Wang 25] の主張と、本研究の主張とは、単純には比較できない。それら先行研究の主張は、相手が人間とわかっている場合と比較して、相手が機械とわかっているときの協力率が下がるというものだった。これに対して本研究の非開示条件は、相手が AI だと特段通知しない（相手が人間だとも通知していない）。相手が人間か AI かを通知していないときの人間の心理状態は、相手が人間であると明確に通知されていた場合とは厳密には異なる可能性がある。チャットの例でも、相手が機械かもしれないと疑う発言をする参加者がいた。また、開示がグループ内の AI 数のみであるため、人間二人条件ではどのプレイヤーが AI か識別が困難であった。AI のプレイヤーがどれかがわかっていたほうが、相手が AI であるという認識の影響はより明確な可能性はある。このように、先行研究と本研究は、厳密には研究手法が異なっている。そのため本研究でいえるのは、相手が人間か AI かわかっていない状態と、相手が AI とわかっている状態では協力に差はなかったということである。そしてそれはチャットや時間の経過による効果と考えられる。

チャットという高い自由度のあるコミュニケーションを人と AI の間で実施したが、対話の中身がどのようであれば、AI であるという認識が協力率に影響を与えないのかという詳しい分析ができていない。たとえば外見・声・動きといった情報を持つ AI が人とコミュニケーションを取る際に、その AI が話す内容をどのようにするのが協力の促進に有効かが上記のチャットの分析に基づいてわかると期待できる。それによって、機械ペナルティや言語的不気味の谷を克服し、社会における AI の可能性を広げる知見を提供することができる。

## 8 むすび

本研究の目的は、社会的ジレンマ状況における協力行動が、相手が AI か人間かを認識することで変化するかを検証することだった。協力率の差は開示の有無には影響されておらず、AI への評価も信頼も

協力率とは相関関係がなかった。開示による心理的影響を確認すると、開示の有無、および開示の有無が絡む交互作用は、約束へのコミットメント（責任感の強さ）などのいずれの尺度にも、影響を与えていなかった。それゆえに、開示は協力率に影響を与えなかったと考えられる。ただ、チャットの有無が協力率に影響を与え、チャットと時間の経過がグループへの信頼を高めていたことがわかり、これらは機械ペナルティやいわゆる言語的不気味の谷の問題を解消する一つの方法といえる。

結果的に、本研究は、相手が AI であるという認識が協力行動に与える影響がないことを確認しただけでなく、機械ペナルティや言語的不気味の谷の問題を解消する一つの方法を提示できることとなった。今後の課題は、本研究で得られたチャットの内容から、不気味の谷や機械ペナルティを解消する対話の特徴をつかむことである。

## 謝辞

本研究は、奈良女子大学男女共同参画推進機構「みかさプログラム」により助成を受けた。

## 参考文献

- [Ahn 24] Ahn, M., Balakrishna, R., Simeone, M., Janssen, M.A.: When does group chat promote cooperation in shared resource governance?, *Humanities & Social Sciences Communications*, Vol. 11, 1455 (2024)
- [安藤 25] 安藤, 日室, 後藤, 笹鹿, 江島, 安原: 社会的ジレンマにおけるオンラインでのコミュニケーションの効果 —オンラインごみ処理ジレンマゲームの開発—, *マーケティングジャーナル*, vol. 45, no. 1, pp. 32-43 (2025).
- [Balliet 11] Balliet, D., Mulder, L.B., Van Lange, P.A.M.: Reward, Punishment, and Cooperation: A Meta-Analysis, *Psychological Bulletin*, Vol. 137, No. 4, pp. 594-615 (2011)
- [Bicchieri 07] Bicchieri, C., Lev-On, A.: Computer-Mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis, *Politics, Philosophy & Economics*, Vol. 6, No. 1, pp. 7-34 (2007)
- [Capraro 13] Capraro, V: A Model of Human Cooperation in Social Dilemmas, *PLoS ONE*, Vol. 8, No. 8: e72427 (2013)
- [Chen 16] Chen D. L., Schonger M., Wickens C.: oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments, *Journal of Behavioral and Experimental Finance*, Vol.9, pp. 88-97 (2016)
- [Cui 14] Cui, P., Wu, Z.-X.: Selfish Punishment with Avoiding Mechanism Can Alleviate Both First-Order and Second-

Order Social Dilemma, *Journal of Theoretical Biology*, Vol. 361, pp. 111-123 (2014)

- [Dafoe 21] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K. & Graepel, T.: Cooperative AI: Machines Must Learn to Find Common Ground, *Nature*, Vol. 593, pp. 33-36 (2021)
- [Dawes 80] Dawes, R. M.: Social dilemmas, *Annual Review of Psychology*, Vol. 31, No. 1, pp. 169-193 (1980)
- [Deci 99] Deci, E.L., Koestner, R., Ryan, R.M.: A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation, *Psychological Bulletin*, Vol. 125, No. 6, pp. 627-668 (1999)
- [Fehr 00] Fehr, E., Gächter, S.: Cooperation and Punishment in Public Goods Experiments, *American Economic Review*, Vol. 90, No. 4, pp. 980-994 (2000)
- [Fichtner 02] Fichtner, W., Graehl, S. and Rentz, O.: International Cooperation to Support Climate Change Mitigation and Sustainable Development, *International Journal of Environment and Pollution*, Vol. 18, No. 1, pp. 33-55 (2002)
- [後藤 23] 後藤, 人間は『人工知能』と『協力』できるか: クラウドソーシングを用いた仮想的 AI エージェント実験による検討, *社会情報学*, Vol. 12, No. 1, pp. 1-22 (2023).
- [Hallers-Haalboom 25] Hallers-Haalboom, E.T., De Jaegher, K.J.M., Sterck, E.H.M., Brosnan, S.F., Vermande, M.M.: Communication increases cooperation among students in a coordination game, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 122, No. 35, e2517669122 (2025)
- [Hoenow 24] Hoenow, S., & Pourviseh, A.: Intragroup Communication in Social Dilemmas: An Artefactual Public Good Field Experiment in Small-Scale Communities. *Judgment and Decision Making*, Vol. 19, No. 3, pp. 456-473 (2024)
- [Hilbe 14] Hilbe, C., Wu, B., Traulsen, A., Nowak, M. A.: Cooperation and control in multiplayer social dilemmas, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111, No. 46, pp. 16425-16430 (2014)
- [Jiang 25] Jiang, G., Yang, S., Wang, Y., Hui, P.: When Trust Collides: Decoding Human-LLM Cooperation Dynamics through the Prisoner's Dilemma, *arXiv:2503.07320* (2025)
- [Kas 21] Kas, J., Hardisty, D. J., Handgraaf, M. J. J.: Steady steps versus sudden shifts: Cooperation in (a)symmetric linear and step-level social dilemmas, *Judgment and Decision Making*, Vol. 16, No. 1, pp. 142-164 (2021)
- [Karpus 25] Karpus, J., Shirai, R., Verba, J. T., Schulte, R., Weigert, M., Bahrami, B., Watanabe, K., Deroy, O.: Human Cooperation With Artificial Agents Varies Across Countries. *Scientific Reports*, Vol. 15, 10000 (2025)
- [Kerr 94] Kerr, N.L., Kaufman-Gilliland, C.M.: Communication, commitment, and cooperation in social dilemma, *Journal of Personality and Social Psychology*, Vol. 66, No. 3, pp. 513-529 (1994)
- [Kiesler 96] Kiesler, S., Sproull, L., Waters, K.: A Prisoner's Dilemma Experiment on Cooperation with People and Human-Like Computers, *Journal of Personality and Social Psychology*, Vol. 70, No. 1, pp. 47-65 (1996).
- [Kishnani 25] Kishnani, D.: The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images, Master's thesis, Massachusetts Institute of Technology (2025)
- [Kitakaji 25] Kitakaji, Y., Hizen, Y., Ohnuma, S.: Communication among selected members improves cooperation in a social dilemma, *Frontiers in Behavioral Economics*, Vol. 4 (2025)
- [Ma 21] Ma, S., Zhang, B., Cao, S., Liu, J. S., Wang, W.-X.: Limited memory optimizes cooperation in social dilemma experiments, *Royal Society Open Science*, Vol. 8, No. 8, 210653 (2021)
- [Maggioni 23] Maggioni, M. A., Rossignoli, D.: If It Looks Like a Human and Speaks Like a Human...Communication and Cooperation in Strategic Human-Robot Interactions, *Journal of Behavioral and Experimental Economics*, Vol. 104, 102011 (2023)
- [Makovi 25] Makovi, K., Bonnefon, J. F., Oudah, M., Sargsyan, A., Rahwan, T.: Rewards and Punishments Help Humans Overcome Biases Against Cooperation Partners Assumed to Be Machines. *iScience*. Vol 28, No. 7, 112833 (2025)
- [Mathur 16] Mathur, M. B., & Reichling, D. B.: Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, Vol. 146, pp. 22-32 (2016)
- [箕浦 87] 箕浦: 社会的ジレンマ状況での非協力から協力への行動変容, *心理学研究*, Vol. 58, No.5, pp. 282-288 (1987)
- [Mitkidis 13] Mitkidis, P., Sørensen, J., Nielbo, K. L., Andersen, M., Lienard P.: Collective-Goal Ascription Increases Cooperation in Humans. *PLoS ONE*, Vol. 8, No.5, e64776 (2013)
- [Mori 70] Mori, M.: The uncanny valley. *Energy*, Vol. 7, No. 4, pp. 33-35 (1970)
- [Mu 24] Mu, C., Guo, H., Chen, Y., Shen, C., Hu, D., Hu, S., & Wang, Z.: Multi-Agent, Human-Agent and Beyond: A Survey on Cooperation in Social Dilemmas, *Neurocomputing*, Vol. 610, 128514 (2024)
- [Murayama 10] Murayama, K., Matsumoto, M., Izuma, K.,

Matsumoto, K.: Neural basis of the undermining effect of monetary reward on intrinsic motivation, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 107, No. 49, pp. 20911-20916 (2010)

- [Ng 22] Ng, Y. L.: When Communicative AIs Are Cooperative Actors: A Prisoner's Dilemma Experiment on Human-Communicative Artificial Intelligence Cooperation, *Behaviour & Information Technology*, Vol. 42, No. 13, pp. 2141-2151 (2022)
- [大沼 12] 大沼: 自分一人くらいの心理学-日常生活に潜む社会的ジレンマのワナ, *暮らしの中の社会心理学* (安藤、杉浦編), pp. 25-36 (2012)
- [Okada 08] Okada, A.: The Second-Order Dilemma of Public Goods and Capital Accumulation, *Public Choice*, Vol. 135, pp. 165-182 (2008)
- [Oliver 80] Oliver, P.: Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations, *American Journal of Sociology*, Vol. 85, No. 6, pp. 1356-1375 (1980)
- [Oudah 18] Oudah M., Rahwan T., Crandall T., Crandall J.: How AI Wins Friends and Influences People in Repeated Games With Cheap Talk, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 No. 1, pp. 1519-1526 (2018)
- [Rand 09] Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., Nowak, M.A.: Positive Interactions Promote Public Cooperation, *Science*, Vol. 325, No. 5945, pp. 1272-1275 (2009)
- [Rand 17] Rand, D. G.: Social dilemma cooperation (unlike Dictator Game giving) is intuitive for men as well as women, *Journal of Experimental Social Psychology*, Vol. 73, pp. 164-168 (2017)
- [Simão 21] Simão, R., Wardil, L.: Social Dilemma in Traffic with Heterogeneous Drivers, *Physica A: Statistical Mechanics and its Applications*, Vol. 561, 125235 (2021)
- [Smith 10] Smith, E.A.: Communication and collective action: language and the evolution of human cooperation, *Evolution and Human Behavior*, Vol. 31, No. 4, pp. 231-245 (2010)
- [Van Lange 13] Van Lange, P. A.M., Joireman, J., Parks, C. D., Van Dijk, E.: The Psychology of Social Dilemmas: A Review, *Organizational Behavior and Human Decision Processes*, Vol. 120, No. 2, pp. 125-141 (2013)
- [Wang 23] Wang, H., Chen, W., Rauner, S., Bertram, C., Luderer, G., Kriegler E.: The Double Dividend of International Cooperation for Climate Mitigation Cost Effectiveness and Public Health Cobenefits, *Environmental Science & Technology*, Vol. 57, No. 10, pp. 4061-4070 (2023)
- [Wang 18] Wang, Z., Jusup, M., Shi, L., Lee, J. H., Iwasa, Y., Boccaletti, S.: Exploiting a cognitive bias promotes cooperation in social dilemma experiments, *Nature Communications*, Vol. 9, No. 1, 2954 (2018)
- [Wang 17] Wang, Z., Jusup, M., Wang, R., Shi, L., Iwasa, Y., Moreno, Y., Kurths, J.: Onymity promotes cooperation in social dilemma experiments, *Science Advances*, Vol. 3, No. 3, e1601444 (2017)
- [Wang 25] Wang, Z., Song, R., Shen, C., Yin, S., Song, Z., Battu, B., Shi, L., Jia, D., Rahwan, T., Hu, S.: Overcoming the Machine Penalty with Imperfectly Fair AI Agents, *arXiv:2410.03724* (2025)
- [Wardil 19] Wardil, L., Silva, I. R., da Silva, J. K. L.: Positive interactions may decrease cooperation in social dilemma experiments, *Scientific Reports*, Vol. 9, 1017 (2019)
- [Yamagishi 13] Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., Inukai, K., Tanida, S., Kiyonari, T., Takagishi, H. Simunovic, D.: Is Behavioral Pro-Sociality Game-Specific? Pro-Social Preference and Expectations of Pro-Sociality, *Organizational Behavior and Human Decision Processes*, Vol. 120, No. 2, pp. 260-271 (2013)
- [Ye 16] Ye, H., Chen, S., Luo, J. et al.: Increasing Returns to Scale: The Solution to the Second-Order Social Dilemma, *Scientific Reports*, Vol. 6, 31927 (2016)

(脚注)

本研究の一部は2025年8月27日に開催された第12回 気候変動・省エネルギー行動会議 BECC JAPAN 2025, および、2025年9月20-21日に開催された日本社会心理学会第66回大会 WS08 ワークショップ, および2025年12月8日に開催された AIoT 行動変容学会研究会 (BTI) で口頭発表済みである。