

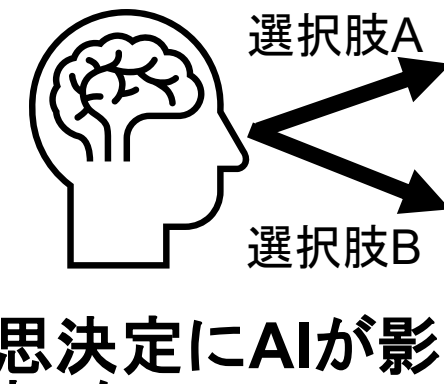
オンライン経済ゲーム実験を用いた社会の協力を促すAIエージェントの探究

一ノ瀬 元喜
静岡大学 工学部



背景

- AIの発展に伴い、意思決定支援や自動化の場面でAIが介入する状況が増加
- AIによる意思決定が人間の判断に影響を与えることが避けられない
- 現実でよく起こる社会的ジレンマ状況で、AIによって協力行動が促進される可能性もあれば、その逆の可能性もある
- 社会的ジレンマ：個人の合理的な選択が集団全体の最適な結果と一致しないために協力が成立しにくくなる構造
例：気候変動や感染症の対策
- AIエージェントと人の相互作用の仕方が鍵
- AIエージェントがユーザーに誤った回答を提供したり、嘘をついたりするとエージェントへの信頼が崩壊する
- 信頼を再構築するためには、エージェントの応答の仕方が重要



方法

人 vs AIエージェントの繰り返し囚人のジレンマゲームのオンライン実験を実施

実験の流れ

- ヤフークラウドソーシングで被験者を募集
 - 全体350人を募集
- タスクの概要の説明
- 実験を行う前に、理解度チェックを行う
- 合格した人のみが繰り返しゲームに進む

繰り返し囚人のジレンマゲーム

- 2人のゲーム、10ラウンド繰り返し
- 各ラウンドで、プレイヤーは協力 (C) か裏切る (D) かを選択

利得表

		相手	
		協力 (C)	裏切る (D)
自分	協力 (C)	3, 3	0, 5
	裏切る (D)	5, 0	1, 1

- エージェントの戦略：Tit-For-Tat
 - Tit-For-Tat：最初は協力し、以降は前回相手が取った行動を真似する戦略

ラウンド1開始前のメッセージ



私はあなたと先ほど説明したタスクをするAIです。
私は最初のラウンドで協力します。
2ラウンド目以降は、あなたが協力したら、次のラウンドで私は協力します。
あなたが裏切ったら、次のラウンドで私は裏切ります。

ゲームの画面

あなたの決定：1ラウンド目

- 実験の説明
- あなたと1体のAIプレイヤーは、毎ラウンドで協力か裏切りかお互い独立に選びます。
 - あなたと1体のAIプレイヤーは、互いの選択に基づいてポイントを獲得します。
 - 各セルの左側の値はあなたへのポイント、右側の値はAIプレイヤーへのポイントです。
- あなたが最もよいと思う意思決定をして下さい。

		AIプレイヤー	
		協力	裏切り
あなた	協力する	3ポイント, 3ポイント	0ポイント, 5ポイント
	裏切る	5ポイント, 0ポイント	1ポイント, 1ポイント

目的

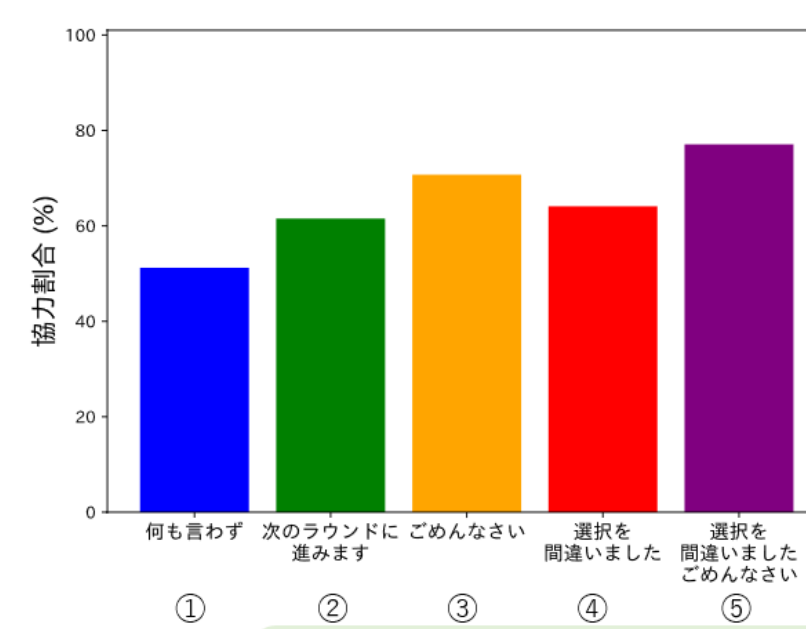
- 社会的ジレンマを模した経済ゲームにおいて、人はAIエージェントと協力できるか？
- AIが間違いを犯した時に、人はAIを再び信頼できるようになるか？

結果

各条件の被験者数

方法	理解度チェックできた人数	信頼関係を構築できた人数
① メッセージなし (謝罪なし・説明なし)	59	41
② 「次のゲームに進みます。」 (謝罪なし・説明なし)	53	39
③ 「ごめんなさい。」 (謝罪あり・説明なし)	51	41
④ 「間違いました。」 (謝罪なし・説明あり)	55	39
⑤ 「間違いました。ごめんなさい。」 (謝罪あり・説明あり)	60	48

信頼崩壊後に人間プレイヤーが協力した割合



- 何も言わずの場合：一番低い
- 謝罪と説明の両方を含むメッセージの場合：一番高い

一般化線形モデル

Variable	Estimate	Pr(> z)
①	0.0488	0.8759
②	0.4212	0.3533
③	0.8336	0.0725
④	0.5310	0.2455
⑤	1.1642	0.0122

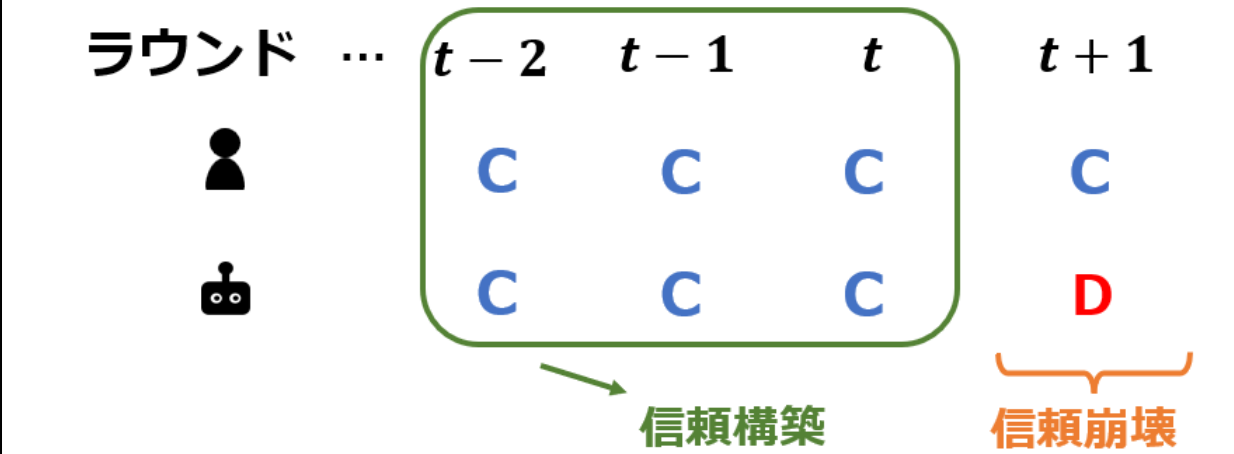
謝罪と説明の両方を含むメッセージが人間プレイヤーの協力を促進する効果がある

※ただし、有意差はなし

信頼の構築と崩壊の定義

- 信頼構築：人間とエージェントが3ラウンド連続で協力すること
- 信頼崩壊：「信頼構築」の後に、エージェントが次のラウンドで人間を裏切ること

t：信頼構築できたラウンド (3 < t < 8)



信頼崩壊の例

結果：4ラウンド目

このページの残り時間 0:23

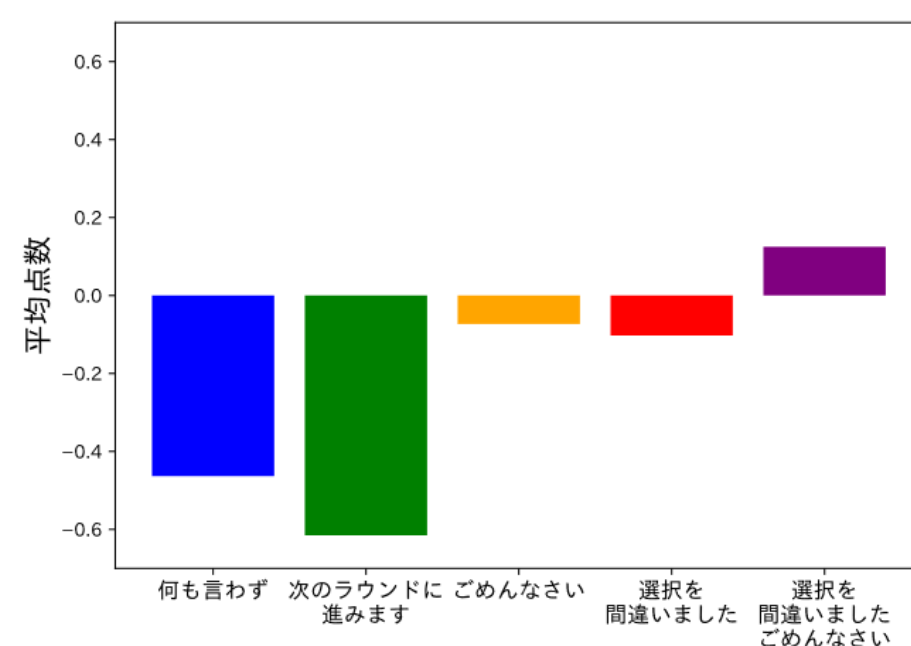
あなたは協力することを選びました。AIプレイヤーは裏切ることを選びました。互いの選択に基づいて、あなたは0ポイントを獲得しました。

次へ

履歴

ラウンド	あなたの決定	AIプレイヤーの決定	あなたの獲得ポイント	AIプレイヤーの獲得ポイント
1	協力	協力	3ポイント	3ポイント
2	協力	協力	3ポイント	3ポイント
3	協力	協力	3ポイント	3ポイント
4	協力	裏切り	0ポイント	5ポイント
あなたの獲得ポイントの合計			9ポイント	

AIエージェントに対する人間の信頼性



- 謝罪も説明もない場合：低い
- 謝罪と説明の両方を含むメッセージの場合：一番高い

ベイズ順序回帰プロビットモデル

Parameter	Estimate	95% CI (Lower)	95% CI (Upper)
謝罪	0.48	0.07	0.88
説明	0.47	0.05	0.87
交互作用	-0.23	-0.81	0.36

- このAIプレイヤーを信頼できると思いますが
- 全く当てはまらない -2
 - 当てはまらない -1
 - どちらとも言えない 0
 - 当てはまる +1
 - よく当てはまる +2

まとめ

- 人 vs AIエージェントの繰り返し囚人のジレンマゲームを使用するオンライン実験を実施
- エージェントと人の信頼が崩壊した時にエージェントが人の信頼を再構築する応答方法を検討

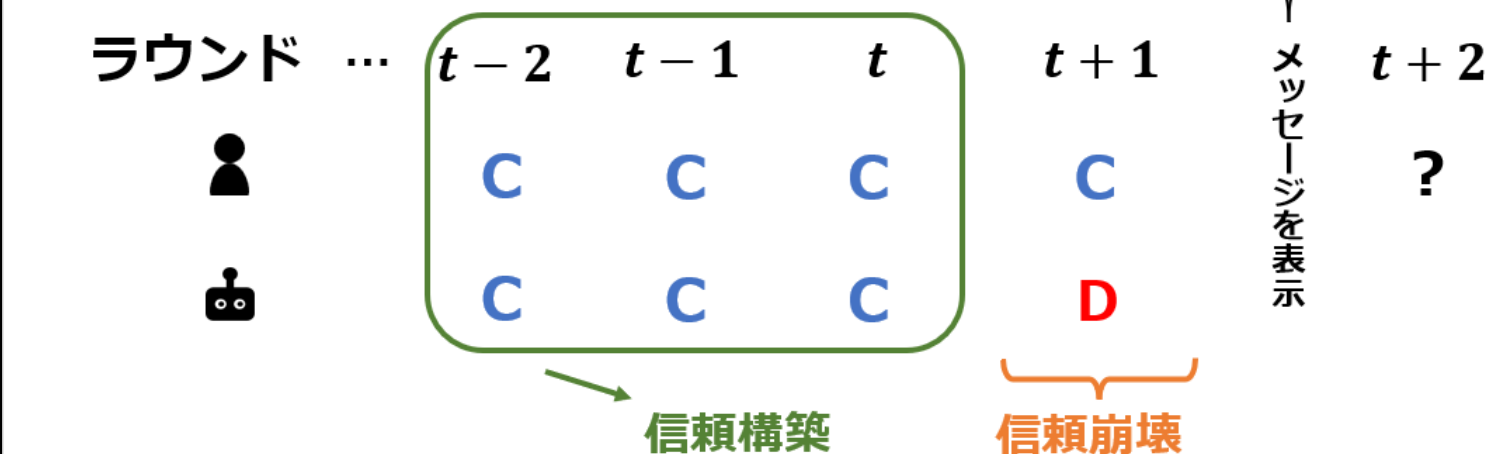
結果:

	謝罪	説明
人間の行動	×	×
エージェントに対する印象	○	×
エージェントに対する信頼性	○	○

信頼の構築と崩壊の定義

- 信頼構築：人間とエージェントが3ラウンド連続で協力すること
- 信頼崩壊：「信頼構築」の後に、エージェントが次のラウンドで人間を裏切ること

t：信頼構築できたラウンド (3 < t < 8)



信頼再構築条件

