2024 年 4 月 19 日 M1 西村昭賢

進捗報告

1 社会的ジレンマ状況下の人間の行動に関する論文の調査

LLM エージェントが社会的ジレンマ環境下において利他的か利己的なのかどのような判断を下すかを見たいという目的で社会的ジレンマの研究例, 社会的ジレンマ環境における人間の行動を研究していた論文を調べていた.

1.1 人間の協力行動に関する実験ゲーム研究と組織管理への応用可能性 [1]

複数の社会的ジレンマ環境を紹介し, それらの環境に相手を搾取しようとするジレンマ (GID), 相手から搾取されないとするジレンマ (RAD) の 2 つの指標を導入して評価, また協力行動の創発による互恵関係が結ばれた状態のゲーム環境の変化を GID, RAD を用いて評価した論文.

論文中では, 代表的なゲーム理論における社会的ジレンマの構造を持つゲームとして, 囚人のジレンマ, チキンゲーム, スタグハントゲーム, トリビアルゲームなど代表的なペアワイズゲームが紹介されていた. 囚人のジレンマは常に裏切りが支配戦略となり協力が促進されづらいゲームであるが, スタグハントゲームは, (C,C) と (D,D) がナッシュ均衡である双安定な構造となっており, 利他的か利己的かをフラットに判断するならスタグハントゲームのほうが良いと考えた.

また, 論文中では協力促進メカニズムとして有名な五つの互恵ルール (Five Reciprocity Rule) [2] が紹介されていた. LLMAgent の利他性の定性的な評価として使用可能な指標かもしれない.

1.2 協力行動と公共財ゲームに関する一考察：経済学実験および心理学実験を中心に [3]

公共財についてゲーム理論的に記述したゲームである公共財ゲームに関しての論文. 協力行動の促進, 社会的構成に関する分析として用いられるゲーム環境らしい. パレート最適は保有学の全額を貢献する場合であるが, 支配戦略は 0 円も貢献しないいわゆるフリーライダーになることであり社会的ジレンマ状況が成立している. 1.1 で述べられていたのはペアワイズゲームであったため, 複数人でのシミュレーションが可能な公共財ゲームが LLMAgent の特に利他性を見るにあたって最適なゲーム環境ではないかと考えた. また, 論文中では「税金」や「組織間競争」, コミニュケーションを導入することで協力活動の促進が見られたと記載されている.

1.3 協力行動と公共財ゲームに関する一考察：経済学実験および心理学実験を中心に [4]

N 人版の繰り返し社会的ジレンマゲームにおける集団内協力者数に依存する行動戦略の自生可能性に焦点を当てた論文. 実験では公共財ゲームと似た社会的ジレンマ環境を取り上げていた.

論文中で設定された社会的ジレンマ環境

7 名グループで実験をする. まずはじめに各被験者が元手として 3 円を与えられる. 次にその 3 円の元手を 「提供する」か「提供しない」かを選択する. 元手を提供すると 3 円は 6 倍され他 6 名へ 3 円ずつ配られる. 提供しなければ, 元手は被験者自身のものとなる. 各参加者は独立に, 同時に, 独立の行動決定をする.

実験の手続きとして, 被験者にインストラクションをし, 説明の中で獲得した金額が実験終了後にもらう謝礼

として設定されていた. ゲームの 1 試行が終わるごとに実験で用いるコンピュータ上には被験者自身が前試行で得た利得と, 過去 3 試行における各参加者の行動選択を表示していた. また, 10 回繰り返すことを 1 セッションとして計 6 セッションの試行をし, 参加者には事前に何試行で終わるのかは周知していなかった.

結果として, グループにおける協力率の推移, 他者の行動推移による影響, 行動パターンの時系列変化の 3 つの観点から評価をしていた.

2　自分の研究の実験設定

1.3 節で述べた実験環境, または公共財ゲームにおいては協力することがパレート最適, 協力しないことがナッシュ均衡となる環境となる. 1.3 節の実験では, 実験で獲得した金額を謝礼とすることと設定することでゲーム外にも協力を促す要素が用いられれていた. これを踏襲し, 各繰り返し試行で被験者が持つ財産を累積していき, 実験で獲得した金額を謝礼とするゲーム外からの協力を促す要素を入れつつ, 最も高い利得を得たプレイヤーは謝礼が 0 となるという非協力を促す要素を追加し, ゲーム外からでもジレンマの設定を入れることで LLMAgent の行動を評価するためのより良い実験設定になると考えられる.

参考文献

[1] 結孝堀田. 人間の協力行動に関する実験ゲーム研究と組織管理への応用可能性. 組織科学, Vol. 53, No. 2, pp. 33–42, 12 2019.

[2] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, Vol. 314, pp. 1560 – 1563, 2006.

[3] 後藤晶. 協力行動と公共財ゲームに関する一考察：経済学実験および心理学実験を中心に. 山梨英和大学紀要, Vol. 12, pp. 32–48, 2013.

[4] 瑞穂品田, 達也亀田. 社会的ジレンマ状況における行動戦略の自生に関する実験的研究. 心理学研究, Vol. 74, No. 1, pp. 71–76, 2003.