



Title	協力と罰の共進化 : 罰に対する感受性を中心とした理論的および実証的研究 [論文内容及び審査の要旨]
Author(s)	土田, 修平
Degree Grantor	北海道大学
Degree Name	博士(文学)
Dissertation Number	甲第16499号
Issue Date	2025-03-25
Doc URL	https://hdl.handle.net/2115/94963
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Shuhei_Tsuchida_abstract.pdf, 論文内容の要旨



学位論文内容の要旨

博士の専攻分野の名称：博士（文学）

氏名： 土 田 修 平

学位論文題名

協力と罰の共進化：罰に対する感受性を中心とした理論的および実証的研究

・本論文の観点と方法

他の動物種と比較した時、人間の社会は協力で満ち溢れている。人間社会に見られる大規模な協力は、社会的ジレンマによって定式化できる。社会的ジレンマ状況では、集団全員が協力することが集団にとっては望ましいが、個人にとってはコストを負担せずに利益だけを享受する非協力が合理的である。そのため進化という観点からみれば、自らコストを負担する協力は淘汰されるはずである。だが、なぜ人間社会には、大規模な協力が広範に存在しているのだろうか？

この謎は、実験と理論の両方から探究されてきたが、一貫して罰の重要性が見出されてきた。非協力に対して罰が与えられるならば、協力の誘因が生まれる。実際、罰の機会が存在すれば高いレベルで協力が維持されることは実験と理論の双方で頑健に確認されてきた。だが非協力者を罰するためにはコストが伴う。個人にとっては、罰を与えないことが合理的となる一方で、誰も罰を与えなければ集団内で非協力が優越する。つまり、罰自体が社会的ジレンマの構造を持っているのである（二次のフリーライダー問題）。そのため、罰によって協力が維持されることは事実だとしても、なぜコストを伴う罰が行使されているのか、謎だとされてきた。

本論文は、この謎を解決するために、罰を受けた側の反応に着目し、罰に対する感受性と罰の共進化によって、大規模な協力が維持されることを、進化ゲーム理論モデルの解析、エージェントベースモデル・シミュレーション、集団実験、計算論モデルの解析を通して検討したものである。

・本論文の内容

本論文は6章からなる。第1章では、協力の進化の研究史を概観したのち、社会的ジレンマの一種である公共財問題ゲームを用いた行動実験について、代表的な知見を紹介した。本論文における研究方略として適応論的アプローチ、すなわち、現在において観察される行動や心理を進化のアナロジーを通して探求する方略を採用することが論じられた。続いて、多くの研究を通して罰が協力を維持するために重要な役割を果たしているにも関わらず、罰には二次のフリーライダー問題が伴うため、なぜ罰が進化するのか十分に理解されていないことが論じられた。最後に、罰を受ける側の反応に着目すれば、協力と罰の共進化を説明できるという本論文の仮説が提示され。

第2章では、研究の出発点として、罰と協力の進化の分野で大きな影響力を持つ連携罰モデル(Boyd, Gintis & Bowles, 2010, *Science*) の再解析が行われた。これは、協力者が連携して罰のコストを分担することで、協力と罰の進化が説明できると主張したモデルである。人類学的知見に基づいて構築され、複数の要因が組み込まれた複雑な数理モデルだった。申請者は、実験計画法のロジックに基づいて、連携罰モデルから要因を丹念に一つずつ外しながら分析をした。その結果、原著者らの主張に反し、連携罰が進化したのは、一回でも罰を受けると即座に協力者へと転ずる非協力者がモデルの中で想定されていたからであることを見出した。さらに罰の進化を説明したと主張する複数の理論モデルをレビューし、いずれの研究においても、罰を受けたら協力者へ転ずる日和見主義者が想定されていることを指摘した。

第3章では、より一般的なモデルを独自に構築して、罰に対する感受性と罰が共進化しうることを立証した。進化ゲーム・シミュレーションの結果、罰に対する感受性の低い鈍感な非協力者が支配する集団の中に、罰に対して敏感な感受性の高い非協力者が侵入できることが見出された。そし

て、罰に対する感受性の上昇に伴い、非協力者に罰を与える戦略と協力戦略が共進化することが明らかとなった。さらに厳密な数理解析を行った結果、罰に対する感受性は一定程度までは確率的な浮動によって増加するが、閾値を超えると淘汰が生じ、高い感受性が適応的になることが示された。

第4章は、罰に対する感受性の一般性を実証的に検討した。第3章の理論的研究から、人々は小さな罰に対しても敏感に反応して、協力を転向することが予測されるが、これは象徴罰と呼ばれる現象、すなわち金銭的な罰を受けずとも他者から不満を伝えられるだけで、協力が維持される現象と酷似している。だが、先行研究では公共財ゲームの繰り返し回数が短く、対立仮説として参加者が象徴罰とコストのかかる金銭罰を混同している可能性が拭えない。そこで、通常よりも長い50ラウンドの公共財ゲーム実験を行い、金銭罰と象徴罰の効果を比較検討した。その結果、象徴罰によって協力は長期的にも維持されることが明らかとなった。また、人々は金銭罰と同様に象徴罰を受けることで、協力的に振る舞うようになること、特に金銭罰と象徴罰の双方を行使可能な条件では、象徴罰が金銭罰の予期的なシグナルとして頻繁に利用されることで、金銭罰のコストを抑えながら高いレベルの協力維持に貢献していることが確認された。

第5章は、罰に対する感受性を強化学習という、より一般的な計算論モデルによって再解釈することを試みた。罰を受けると行動を変えろという行動原理は、強化学習の産物と理解できる。また近年、他者の利益をも考慮する社会的選好を、強化学習によって理解できるという主張も注目を集めている。そこで、公共財ゲーム実験を実施し、参加者の行動に対して強化学習モデルをフィッティングし、協力性の個人差を強化学習モデルにおけるパラメータの違いとして記述することを試みた。その結果、向社会的な参加者ほど学習率が高い傾向が見出された。これは従来の社会的選好による説明とも整合することが論じられた。

第6章では、各研究で得られた知見に基づいて、本論文の主張の妥当性を検討した。最後に、今後の展望として、協力と罰の共進化を解明するためには、象徴罰の理解が重要であること、より積極的に強化学習の枠組みを、協力行動と罰行動の研究に拡大する必要があることが論じられた。