

評判にもとづく協力行動の数理



村瀬 洋介

理化学研究所数理創造研究センター
yohsuke.murase@riken.jp



藤本 悠雅

サイバーエージェント
fujimoto.yuma1991@gmail.com



大槻 久

総合研究大学院大学統合進化科学研究センター
ohtsuki_hisashi@soken.ac.jp

集団現象の解明には、物理学の知見や手法が有効であることが多い。この視点を社会現象に応用する研究は、社会物理学として発展してきた。その重要な対象の一つが、集団における協力行動である。協力行動とは、自らコストを支払って他者を助ける行動であり、一見すると自然選択の原理と相反する。それにもかかわらず、協力は自然界にも人間社会にも広く見られ、その成立と維持は進化生物学、社会科学、物理学などにまたがる長年の問いとなっている。

この謎を解く鍵の一つが、評判を介して協力を循環させる仕組み、すなわち**間接互恵性**である。これは、個体が協力行動を取ると、その行動が良い評判として集団に伝わり、後に第三者から返報を受けるというメカニズムである。逆に個体が非協力行動を取るとそれは悪い評判につながる。「情けは人のためならず」ということわざがあるが、これは間接互恵性のことそのものである。どのような行動や評価を行う集団であれば協力が達成され、さらにフリーライダーに対しても頑健にそれを維持できるのか。これが間接互恵性の理論が取り組む中心的な問いである。この問題に対し、**進化ゲーム理論**を用いるさまざまな理論研究が進展してきた。

初期の研究では、間接互恵性で協力を維持するためには「処罰の正当化」が重要になることが明らかになった。これは、評判の悪い相手に対しては自分は裏切ってよく、その行動は良いと評価されるという性質である。協力を維持可能として数理的に特定された「**リーディング・エイト**」と呼ばれる評価ルールは、いずれもこの性質を備えて

おり、我々が持つ勤善懲惡的な直感や道徳的判断の一側面を捉えているように見える。

一方で、リーディング・エイトの成功は、評判を集団の全員で共有しているという強い仮定（**公的評価**）の上に成立しているが、現実社会においては評判が完全に共有されることはほとんどなく、意見の不一致は不可避である。このより現実的な状況は**私的評価**と呼ばれる。私的評価の下では、同じ評価ルールを持った集団でも、小さな意見の違いがより連鎖的に拡大し、協力が崩壊しやすくなる。

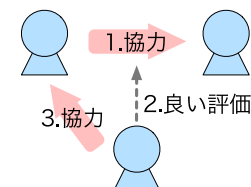
近年の研究により、私的評価の下で協力を維持するには、処罰の正当化に加えて、意見を同調させる仕組みが不可欠であることが明らかになった。たとえば、評判の悪い人への協力行動を許容する寛容さがあると、意見の違いがエスカレートせず、協力が維持されやすくなる。さらに最新の研究では、公的評価と私的評価のモデルを統一的に扱う枠組みが提示され、意見の同調度を考慮することで協力の進化条件を調べるのが可能になっている。我々のゴシップや同調心理が意見の同期に役立ち、不確実な状況下でも間接互恵性を成り立たせることに寄与していると考えられる。

さらに視野を広げれば、現代の我々が直面しているのは、接続が疎で動的な大規模社会ネットワークである。SNS や AI が介在し、情報にノイズが含まれる現代の社会において、いかにして頑健な信頼を醸成するか。この問いに対して、意見形成の相転移や集団ダイナミクスといった統計物理学の知見は、間接互恵性の理解を次の段階へと押し上げる可能性を秘めている。

用語解説

間接互恵性：

特定の相手との直接的な相互協力（直接互恵性）ではなく、自身の協力行動に対する評判を介して、第三者から協力が返ってくる仕組み、見返りが連鎖することで、大規模な集団内でも協力を可能にする。



間接互恵性のイメージ。ある人の協力（図中1）が第三者によって良い行動と評価（2）され、その評価に基づいて協力（3）が返ってくる。

進化ゲーム理論：

個々の戦略がもたらす利得を、生存や繁殖の成功率（適応度）と見なし、集団内でどの戦略が生き残り広まっていくかを分析する枠組み。

リーディング・エイト：

協力を高いレベルでかつ安定に保つことのできる8つの評価ルール。

公的評価・私的評価：

個人に対して集団全員で同一の評判を共有している状況を公的評価と呼ぶのに対し、同じ個人に対する評価が（異なる情報や誤解などにより）個人間で異なる状況を私的評価と呼ぶ。