# An Intention Signaling Strategy for Indirect Reciprocity: Theoretical and Empirical Studies

Tanaka, Hiroki

(Degree)
博士（学術）

(Date of Degree)
2017-03-25

(Date of Publication)
2018-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6800号

(URL)
https://hdl.handle.net/20.500.14094/D1006800

# 博士論文要旨

論文題目

An Intention Signaling Strategy for Indirect Reciprocity:
Theoretical and Empirical Studies
（間接互恵状況における意図シグナル戦略：理論・実証研究）

氏名：田中大貴

神戸大学大学院人文学研究科博士課程後期課程社会動態専攻

指導教員氏名　（主）大坪庸介　准教授
（副）石井敬子　准教授
（副）松田毅　教授

## 1. 序論

協力―自身にコストを課し、他個体に利得を与える行動―は、自然界の様々な種において見出すことができる。しかし、その中でもとりわけヒトの協力は、血縁者や長期的な関係を持つ知り合いに対してだけでなく、見知らぬ他者、見返りの期待できない相手に対してもなされるという点において、他の種のそれと明確に異なっていると言える。この、直接的な返報のないヒト特有の協力行動がなぜ進化してきたのかという問いに対し、理論的・経験的に妥当な適応的説明を与えることが本研究の最も大きな主題である。

### 1.1 先行研究において提案されたモデルとその問題

#### (a) 全ての非協力者が「悪い」わけではない

Alexander (1987)はこうした協力行動を進化的観点から説明するために、「間接互恵性」という原理を提唱した。これは、A が B に協力すると、A の評判が上がることにより、B ではなくその評判を参照した別の他者、C から返報を受けることができるというシステムである。こうした評判に基づく協力システムにより、ヒトは直接的なつきあいを越えた大規模な協力関係を築くことができたと考えられる。Nowak and Sigmund (1998)はこのシステムをモデル化し、協力したプレイヤーに「良い」評判が、協力しなかったプレイヤーに「悪い」評判が割り当てられ、さらに「良い」プレイヤーのみが他者から協力を受けるという条件下で集団の協力的均衡が維持されることを示し、こうした評判割り当てと行動規範の総称を image scoring (IS)戦略と名付けた（ただし、この場合の「戦略」は「合理的・意識的な意思決定」を含意しない）。

しかし、IS 戦略には理論的な欠陥が存在する。今、IS 戦略を用いる A が「悪い」評判を持つ B に対し非協力的に振舞ったとする。すると A は、その非協力により「悪い」評判を割り当てられ、IS 戦略を用いる C から協力を断られてしまう。したがって、「良い」評判を維持するためには、プレイヤーは相手の評判にかかわらず常に協力し続ける必要があることになり、IS 戦略は進化できない（無差別的な協力戦略は非協力者のただ乗りを許してしまう）。さらに、こうした非協力の連鎖は、集団内にフリーライダーが侵入する場合だけでなく、協力をしようとした者がその意図に反して非協力的に振舞ってしまう（協力に失敗してしまう）という「実行エラー」によっても起こる。すなわち IS 戦略の問題は、フリーライダーに対する非協力と、フリーライダーによる非協力を区別できないという点にある。

#### (b) 二次情報

この非協力の連鎖という問題は、Standing (ST)戦略によって解決可能である (Leimar & Hammerstein, 2001)。この戦略は、非協力を「正当化可能な非協力」と「正当化不可能な非協力」の二種類に区別する。例えば、IS 戦略を用いる A が「悪い」評判を持つ B への協力を断ったとする。このとき、ST 戦略においては、A の非協力は正当化され、A には「良い」評判が割り当てられる。一方、フリーライダーである F は相手の評判に依存せず一貫して協力しないので、「良い」評判を持つ D に対しても非協力的に振る舞う。ST 戦略ではこの非協力は正当化さ

れず、F には「悪い」評判が与えられる。このように、ST 戦略はプレイヤーに評判を割り当てる際、対象者 (A, F)の行動だけでなく、その行動がとられた相手(B, D)の評判も考慮に入れる。この「対象者がとった行動の相手の評判」は「二次情報」と呼ばれ、これまでの理論研究において、間接互恵性の進化のために重要な要素とされている (Ohtsuki & Iwasa, 2004, 2006)。

それでは、現実に人々は ST 戦略と合致した行動をとっているのだろうか。実は、この点に関して実証研究の結果は一貫していない (Milinski et al., 2001; Swakman et al., 2016)。なぜ人々は二次情報を用いた非協力者の区別を必ずしも行わないのか。その理由として、二次情報の利用に認知的な負荷がかかっているという可能性が考えられる。実際に、人は認知的倹約者であり、経済ゲーム下の意思決定においても参照できる情報を全て利用しないことが知られている (Fiske & Taylor, 2013; Ohtsubo & Rapoport, 2006)。こうした先行研究が示しているのは、間接互恵性を維持するための評判割り当てシステムをいくら精緻にしたところで、そのために参照が必要な情報が増えてしまっては、それは経験的に妥当なモデルとなりえないということである。

### 1.2. 良い意図のシグナルによる評判維持

間接互恵性に関する従来のモデルは、プレイヤーが自身の評判が他者から割り当てられる過程に一切関与しないということを暗黙の前提としている。しかし実際には、人々は自発的に自身の評判を維持することに努める。例えば、行動実験において利己的意図のない非協力的行動をとった参加者は、その後他者への協力頻度を増やしたり、謝罪や自己罰といった反省行動をとったりする (Milinski et al., 2001; Tanaka et al., 2015; Watanabe & Ohtsubo, 2013)。したがって、間接互恵状況において「正当化可能な非協力」をとった者が自身の良い意図をシグナルできれば、そのシグナルを受け取った者は二次情報を利用せずともその非協力を「正当化不可能な非協力」と区別できるようになるだろう。

そこで本研究では、間接互恵性の進化を説明するモデルとして新たに意図シグナル (intention signaling, intSIG)戦略を提案する。この戦略は、非協力をとった際にシグナルを発し、さらにシグナルを発した非協力者に「良い」評判を割り当てる。以下、「2. 理論研究」において intSIG 戦略の詳細な説明を行い、数理解析とシミュレーションによってその理論的妥当性を確認する。その後、「3. 実証研究」において、人々が実際に intSIG 戦略に合致した行動傾向を持つことを示す。



## 2. 理論研究

### 2.1. 間接互恵状況における意図シグナル戦略

intSIG 戦略を説明する前に、その前提となる間接互恵状況のモデルを説明する。このモデル状況で IS や ST といった具体的な戦略の記述が可能になる。間接互恵状況を数理モデルとして表現する場合、無限数のプレイヤーが集団内に存在し、複数ラウンドで構成されたドネーションゲームという資源のやり取りを行うと仮定する。各ラウンドの頭で、プレイヤーがランダムにペアリングされ、その内の片方がドナー、もう片方がレシピエントという役割を割り与えられる。ドナーとなったプレイヤーはペアであるレシピエントに協力するか否かを選択する。協力する場合、ドナーはコスト(c)を支払い、レシピエントは利得(b)を受け取る($b > c > 0$)。このとき、ドナーはペアであるレシピエントの評判を参照できる。ただし、ドナーが協力を選択したとしても、確率 e で実行エラーが発生し、行動が非協力に強制的に切り替わる。全てのドナーが選択を終えたら、確率 $\omega$ で次のラウンドに移行する（逆に言うと、確率 $1-\omega$ でゲームが終了する)。どのようなプレイヤーに「良い／悪い」評判を割り与えるか、またどのような評判に基づき協力／非協力の選択を行うかはプレイヤーの戦略に依存する。

こうした状況のもと、intSIG 戦略は以下のような行動規範を持つ。まず intSIG 戦略は、実行エラーが起こらない限り「良い」評判を持つ相手に協力をし、「悪い」相手を持つ相手に非協力的に振る舞う。さらに、この戦略は非協力的に振る舞った後、コスト s を支払いシグナルを用いる。そしてシグナルを用いた非協力者には「良い」評判が割り当てられる。すなわちこのコストの支払いが、非協力者による良い意図の表示として機能する。シグナルの使用にコストがかかるのは、このコストがシグナルの信頼性を保証するからである(cf. Zahavi & Zahavi, 1997)。もしも s が協力のコスト c よりも小さい場合、フリーライダーが偽のシ

グナルを発して「良い」プレイヤーを装い他者から協力を受けることができてしまう。したがって、シグナルに必要なコストは少なくとも協力のコストと同値でなければならない。また、以上の規範より、intSIG 戦略が「悪い」相手と見なすのは、シグナルを用いない他の戦略を使用するプレイヤーが非協力的に振る舞ったときのみに限る。

### 2.2. 進化ゲームによる数理解析

協力の進化を説明する特定の戦略の理論的妥当性を確認する際に行われるのが、進化ゲーム理論にもとづく進化的安定性の分析である (Maynard Smith & Price, 1973; Maynard Smith, 1982)。この分析では、特定の戦略を用いる無限数のプレイヤーで構成された集団に、別の戦略を用いる少数のプレイヤーが侵入した状況を仮定する。その上で、ドネーションゲームを通して被侵入個体と侵入個体が獲得する利得を比較する。前者の利得の方が大きい時、その侵入された側の戦略は、侵入してきた代替戦略に対し「進化的に安定」であると結論付けられる。本研究では無条件非協力戦略 (完全なフリーライダー) と無条件協力戦略の２戦略を代替戦略とし、これらに対する intSIG 戦略の進化的安定性を検証した。解析の結果、ゲームのラウンド数が多く、また協力の実行エラーの確率が小さいとき、intSIG 戦略は上記の２戦略に対し進化的に安定であることが示された。

### 2.3. コンピュータシミュレーション

intSIG 戦略は協力時だけでなく、非協力的に振る舞った後シグナルを発する際にもコストを支払うために、従来提案されてきた戦略、特に二次情報を用いる ST 戦略と比べ非効率であるかもしれない。そこで、intSIG 戦略を用いるプレイヤーで構成された集団と ST 戦略を用いるプレイヤーで構成された集団がそれぞれにドネーションゲームを行った際の獲得利得をシミュレーションで比較した。その結果、協力の実行エラーの確率が高くとも、協力の利得-コストの比率が大きければ、intSIG 戦略は ST 戦略に比べ効率的に協力的均衡を達成できることが示された。



## 3. 実証研究

実際に人々が intSIG 戦略に合致した行動をとっているならば、以下の２つの仮説が支持されるはずである: (a)「良い」相手によりも、「悪い」相手に対し非協力的に振る舞った場合の方が、頻繁にコストのかかったシグナルを用いる (前者は無条件非協力戦略をとっていると考えられるため)。(b)シグナルを用いなかった非協力者に対してよりも、シグナルを用いた被協力者に対し、頻繁に協力する。さらにこうした行動傾向は、ゲーム内での社会的学習や強化に先立って現れるだろう。そこで実験 1 では、参加者がシグナルを用いても他のプレイヤーから必ずしも協力的に振る舞われない状況を設定し、100 ラウンドのドネーションゲームを実施した。実験の結果、参加者の行動傾向は仮説(a), (b)をともに支持した。この結果を受け、実験 2 では、参加者がシグナルを用いた場合に他のプレイヤーから高確率で協力的に振る舞われる状況を設定し、実験 1 で見られた行動傾向がどのように変化するかを検証した。その結果、参加者の行動傾向は仮説(a), (b)を支持するだけでなく、実験 1 で見られた行動傾向をさらに強化したものになった。実験 1・2 より、intSIG 戦略の経験的妥当性が示された。

## 4. 考察・結論

本研究は、間接互恵性を成り立たせるモデルとして新たに意図シグナル戦略を提案し、その理論的・経験的妥当性を確認した。従来、ヒトの協力をモデル化する際には、相手の意図を推測する能力のみが重要視されてきた。しかし、本研究はその限界を示すとともに意図を推測される側の積極的な意図の発信に焦点を当て、実際にそうした行動を組み込んだモデルが人々の協力行動をよりよく記述できることを数理解析やシミュレーション、行動実験によって示した。これらの知見は、ヒト特有の大規模な協力の進化に、協力それ自体だけでなく、協力的な意図の伝達が重要であることを示唆している。

［課程博士用］

（別紙１）

# 論文審査の結果の要旨

| | |
|---|---|
| 氏　　名 | 田中　大貴 |
| 論文題目 | An intention Signaling Strategy for Indirect Reciprocity: Theoretical and Empirical Studies<br>（接互恵状況における意図シグナル戦略：理論・実証研究） |

## 要　　旨

この論文では、ヒトの大規模な協力行動の進化メカニズムとして注目されている間接互恵性を維持するために有効な戦略について検討されている。間接互恵性とは、AがBを助けると、Bから直接返報があるのではなく、それを見ていた別の誰か（例えばC）から将来援助が与えられるという状況である。善い行いには返報があるが、それが直接助けた相手からではなく、別の誰から与えられるために間接互恵性と呼ばれている。ヒトだけが間接互恵性を維持できるのは、ヒトが評判情報を利用できるからであると考えられている。したがって、間接互恵性とは評判に基づく大規模な協力ということができる。

本論文第１章では、従来の間接互恵性研究の問題点と限界が議論されている。間接互恵性により大規模な協力が進化可能であることを最初に示したNowak & Sigmundの論文では、社会の成員がImage Scoring戦略と呼ばれる行動規則に従って行動するならば、その社会において間接互恵状況が進化的に安定となると議論されていた。Image Scoring戦略とは、これまでの協力の履歴に応じて評判が決まるというものである。例えば、全ての成員が毎回援助が必要な者か、他者を援助できる者のいずれかになり、援助可能な者は目の前のパートナーを一定のコストをかけて助けるかどうかを決定する状況を考える。このとき援助のコスト(c)は相手が受け取る利益(b)よりも小さいと考える。Image Scoring戦略は、直近の援助可能状況で相手を援助していた者をgoodな人、援助していなかった者をbadな人とみなし、goodな人が困っているときにだけ援助する。ここに突然変異で非協力的な成員が生じると、他者を援助しない非協力的成員は誰からも援助を受けることができず、Image Scoring戦略を使う集団に侵入することができない。

ところが、後の研究で、突然変異で生じた非協力的な成員がもたらすマクロな問題が指摘された。具体的には、非協力的成員を援助しなかった者は一時的にbadになり、次に誰かを助けるまで他者から援助を受けられなくなる。同様に、この成員を援助しなかった者も一時的にbadになる。つまり、一人の非協力的成員を援助しないことによって、本来協力的なImage Scoring戦略を使う成員の評判を次々にbadになるため、集団全体の協力率が下がってしまうのである。同様の問題は、突然変異体の出現ではなく、本来協力的な成員が意図せずに協力しそこなうエラーによっても引き起こされる。

Image Scoring戦略に関わる上記の問題は、Standing戦略（自分の現在のパートナーがなぜ前回援助しなかったのかを考慮する戦略）により回避可能である。例えば、自分の現在のパートナーが前回援助しなかった相手がbadな相手であれば、これは正当化可能な非協力とみなして、この成員の評判をbadにしないという規則である。この戦略は、理論的には上記のImage Scoring戦略の問題を回避し、非協力的戦略の侵入を退けつつ非協力の連鎖を生むこともないことが示されている。しかし、実験室で人々がこの戦略を用いるかどうかを検討した研究の結果は必ずしもこの戦略が人々から使用されないことを示している。つまり、理論的には間接互恵性の進化の問題は解けているにも関わらず、実験研究では人々がどのような戦略を実際に使って間接互恵性を成立させているのかがわかっていないということである。

| | |
|---|---|
| 主査記載<br>氏名・印 | 喜多　伸一 |

第２章では、間接互恵状況における新たな戦略を提唱し、その進化的安定性を進化ゲーム理論を用いて検討している。提唱されているモデルは意図シグナル戦略といい、エラーや正当化可能な非協力をした後に評判の形成を他人まかせにするのではなく自分自身で積極的に悪い意図がなかったことを示すというものである。進化ゲーム理論の分析により、意図シグナル戦略が無条件非協力戦略、無条件協力戦略の侵入に対してほぼ安定であることが示されている。また、コンピュータ・シミュレーションにより意図シグナル戦略だけの集団のネットの利得とStanding戦略だけの集団のネットの利得を比較している。その結果、協力により相手が受け取る利益（b）と協力に必要とされるコスト（c）の比が2:1を上回っている場合には意図シグナル戦略集団の方がネットの利得が大きくなることが示された。

第３章では２つの実験室実験が報告されている。いずれの実験でも意図シグナル戦略が使用可能な状況とStanding戦略が使用可能な状況を設定し、実験参加者が果たしてそれぞれの戦略を自発的に使用するようになるかを検討している。意図シグナル戦略に関しては、エラーによる非協力、非協力者に対する非協力の後に悪意がないことの印としてシグナルが使用されていた。また、それを見た他の参加者はシグナル使用者に協力する傾向があった（良い評判を付与していた）。これらの結果は、意図シグナル戦略からの予測に合致している。Standing戦略が使用可能な条件の結果は曖昧で、一方の実験ではStanding的振る舞いが見られ、もう一方の実験ではStanding的振る舞いは観察されなかった。

第４章の考察においては、Standing戦略が想定するパートナーの以前の行動の相手の評判という複雑な情報を利用することの意味と、それより簡便なシグナルを利用することの意味が比較検討されている。また、シグナルという協力・非協力とは違う次元の行動オプションを導入することの意義が考察されている。

本博士論文では、意図シグナル戦略という間接互恵状況での新たな戦略が進化的安定性を備えていることを理論モデルにより示すとともに、実際に人々がそのような戦略を自発的に用いることが実験により示されている。間接互恵性を成立させる戦略については多くの議論が蓄積されてきているが、本博士論文で提唱された戦略はこれまで考慮されていない新たな行動オプションを組み込んだもので、かつそれが実験研究でも強く支持されているという点で高い学術的意義を備えている。このことは、本博士論文研究のシミュレーション以外の部分が*Proceedings of the Royal Society B*誌に掲載されていること、申請者の田中大貴が本研究の発表に対して日本グループダイナミックス学会と日本人間行動進化学会から発表賞を受けていることからも明らかである（グループダイナミクス学会からは実験パートの発表に2015年度優秀学会発表賞を、日本人間行動進化学会からはシミュレーション・パートの発表に若手奨励賞を受けている）。以上のことから，学位申請者・田中大貴は、博士（学術）の学位を得る資格を有すると認める。

## 審査委員

| 区分 | 職名 | 氏名 | 区分 | 職名 | 氏名 |
|---|---|---|---|---|---|
| 主査 | 教授 | 喜多　伸一 | 副査 | 教授 | 松田　毅 |
| 副査 | 准教授 | 石井　敬子 | 副査 | 准教授 | 野口　泰基 |
| 副査 | 准教授 | 大坪　庸介 | | | |