特集「計算社会科学」・論文

社会的ジレンマに適応的な規範の計算社会科学：理論・実験・シミュレーションの統合

Computational social science on adaptive norms in social dilemmas: Integrating theory, experiments, and simulations

キーワード：
社会的ジレンマ，間接互恵性，進化ゲーム理論，被験者実験，エージェントベース・シミュレーション
keyword：
Social dilemmas, Indirect reciprocity, Evolutionary game theory, Subjects experiments, Agent-based simulations

創価大学　岡　田　　勇
Soka University　Isamu OKADA

要　約
本論文では，人間はなぜ社会的ジレンマ状況において自ら進んで協力を行ってきたのかという「協力の進化」問題に対し有力なメカニズムである間接互恵性を扱う。間接互恵性研究では，どのような情報で他者を評価すべきかについて，理論と実証において大きな対立がある。理論研究では，行動情報のみを用いた評価は進化的安定性を有しないことから複雑な情報処理の必要性を主張している。一方，実証研究では，人間を対象にした実験の蓄積から，人間はそこまで複雑な情報処理を行っていないと主張している。我々は，理論研究で用いられてきた公的評価仮定の非現実性に着目し，これを緩和した私的評価系の分析を行った。この系の解析には無限本の連立方程式を解く必要があり理論解析を困難にする。そのため，我々は別の仮定を導入し厳密解を導出した。この仮定が解に与える影響を確認するため，補完的にエージェントベース・シミュレーションを行い，解の信頼性を確認した。その結果，協力社会を維持できる間接互恵規範は，私的評価系においてはいくつかの特徴がこれまでの知見とは異なることを明らかにした。特に，私的評価系で顕在化する問題を解消するために導入した留保規範の優位性が明ら

原稿受付：2019年９月21日
掲載決定：2019年11月13日

かとなった。この理論結果を実証的に確認するため人間を被験者とする実験を行い，留保規範が許容されることを統計的に検定した。間接互恵規範を探求するため，理論・シミュレーション・実験という異なるアプローチを統合することは，計算社会科学に新たな貢献を提供する可能性がある。

Abstract

In this paper, we consider indirect reciprocity which is an important mechanism on the evolution of cooperation asking why the humans can voluntarily cooperate with others in social dilemma situations. There is a severe conflict between theoretical studies and empirical ones in the aspect of what information is required for assessing others. Theoretical studies claim that complex information should be required because an assessment rule using behavioral information only cannot have an evolutionary stability while empirical studies object because many experiments show that they may not proceed such the complex information. Here we analyzed a private assessment system relaxing a public assessment assumption which is unrealistic but often used in theoretical analyses. Its rigorous analysis is extremely difficult because the private assessment system must solve a system with infinite equations. To do so, we introduced another assumption to solve it strictly. Moreover, we complementarily performed an agent-based simulation to confirm a reliability of the solution due to test the effects on the solution of the assumption introduced. As a result, we show that an adaptive norm on indirect reciprocity to keep cooperative regimes in a private assessment system has different features from well-known norms considering in a public assessment system. We also show that the staying norm we defined for resolving the issue actualized in the private assessment system has a superiority. To confirm our result empirically, we conducted subjects experiments and tested allowance of the staying norm statistically. Integrating theory, experiments, and simulations to explore an adaptive norm on indirect reciprocity may be possible to provide a new contribution to computational social science.

1　はじめに

社会科学研究において，理論と実証の統合は他の分野にもまして重要なトピックである。理論的に精緻な分析は，厳密な議論を支援し，しばしば驚くべき帰結を説明するものの，対象が人間を含む社会科学領域であることから，物理法則によって捕捉できる自然科学とは異なり，理論が意識的に，あるいは，無意識的に仮定した非現実性について常に考慮しなくてはならない。一方，実証分析は帰納的な推論によって一般化されるものの，しばしば予測しえなかった要因によって覆される。計算社会科学は，新興の学問勢力としてそれを定義し評価するには時期尚早とはいえるものの，理論と実証の統合というトピックにおいて，新たな可能性を提供しうると期待される。

本稿では，間接互恵性による協力の進化研究においてこれまで対立してきた理論研究と実証研究の乖離を埋めるべく，統合的な視点で構築した仮説を理論・シミュレーション・実証の各手法を用いて分析し，新たな規範の提案とその効果について検討する。間接互恵性を成立させる規範を特定することは，協力の進化研究において重要なテーマであり，理論的にも実証的にも膨大な研究の蓄積がある。しかし，それらは必ずしも整合的な知見とはなっておらず，特に理論と実証における知見には大きな乖離がみられる。その統合には両者の歩み寄りが必要であるものの，これまで十分な努力がなされてこなかった。

我々は，理論研究において解析の可能性のために置かれてきた仮定の一つである公的評価仮定に注目し，それを緩めた私的評価系を分析した。その結果，私的評価系において協力をもたらす規範は，公的評価系におけるそれとは大きく異なる特徴を有することを明らかにした。この分析のため，我々は理論解析とシミュレーションを補完的に用いた。なぜなら私的評価系は厳密な理論解析に非常な困難が伴うために，理論的には近似的な分析とならざるをえない。そこで，知見の妥当性についてエージェントベース・シミュレーションを用いて確認する必要があるからである。

私的評価系の分析から，これまでの理論研究では検討されてこなかった新たな規範に注目する必要性が明らかとなった。そこで，我々はそれを留保規範（Staying）と名付け，理論的検討のみならず，実証的な妥当性についても検討した。間接互恵性研究のこれまでのほとんどの実証分析では，規範によって物事の善悪を判断するときにどの情報を用いるかについて，その取得順序に暗黙的な決まりがあった。しかし留保規範はそれとは異なる順序で情報を取得する可能性を示唆しており，それを確認すべく，われわれは情報の取得順序を自由にできる環境で被験者実験を行った。その結果，留保規範が予言していた情報所得順序が観察されたのみならず，留保規範の特徴である「判断を留保する」というケースが特定状況で起こりえることを統計的に明らかにした。

本稿では，２節において理論と実証の両面から間接互恵性研究のこれまでの蓄積について概観する。３節において私的評価系の理論分析とシミュレーションによる補完結果についてまとめる。４節において留保規範の妥当性に関する被験者実験の結果をまとめる。最後に５節において留保規範が理論と実証の両面をどのように統合した観点を提供しうるのかについて議論するとともに，理論・シミュレーション・実証の各手法を統合する意義について計算社会科学の文脈から位置付ける。

2　間接互恵性研究のこれまでの蓄積

人類社会が他の種に比べて高度な文明を築き得た原理の一つとして協力原理が指摘されている（Seabright, 2010; Boehm, 2012; Harari, 2015）。これは，人間が互いに協力し合うことで，他の種に打ち勝つのみならず，複雑な社会システムの構築を可能にし，文明社会を創造しえたとす

るパースペクティブである。ここで協力行動とは，自らは時間や金銭などのコスト，あるいは労働力を提供し，他者あるいは社会システムに実利をもたらす行動（Hamilton, 1963; Trivers, 1971; 山岸，1990; Nowak, 2006; Sigmund, 2010）を指す。合理的思考を信奉する場合，協力行動は単純な説明論理を持たない。なぜなら協力するには自らの犠牲を必要とするため，協力しないことへの誘因が絶対的に存在するからである。

そのような協力原理を，なぜ人類が獲得しえたかを探る一連の学術的努力は「協力の進化」研究として，進化論のみならず心理学（Wedekind and Milinski, 2000; Milinski et al. 2001; Milinski et al. 2002; Takahashi and Mashima, 2006; Yoeli et al. 2013）・経済学（Sugden 1986, Kandori, 1992）・生物学（Alexander, 1987; Nowak and Sigmund 1998a; Ohtsuki and Iwasa, 2004; Sasaki et al., 2017）・社会学（Nakai and Muto, 2005; Nakai and Muto, 2008）・政治学（Axelrod, 1984）あるいは，数学（Pacheco et al., 2006; Santos et al., 2016）・物理学（Uchida 2010; Yamamoto et al., 2019）・情報工学（Toriumi et al., 2016）など多くの学問分野においてなされてきた。なかでも，協力コストのため非協力への誘因が存在するにもかかわらず，社会構成員の全員が協力する場合の各構成員が受け取る利得が，全員が非協力の場合のそれよりを上回る状況は社会的ジレンマ（Sigmund, 2010; van Lange et al., 2014）と呼ばれ，近年の集中的な研究により，いくつかの成果をもたらした。

社会的ジレンマ状況での協力行動を説明する有力なメカニズムとして互恵性原理（Trivers, 1971; Nowak and Sigmund, 1998a; Wedekind and Milinski, 2000; Fishman, 2003; Brandt and Sigmund 2005; Brandt and Sigmund 2006; Masuda and Ohtsuki, 2007; Uchida, 2010; Uchida and Sigmund, 2010; Panchanathan, 2011; Uchida and Sasaki, 2013; Ghang and Nowak, 2015）が挙げられる。協力するのは，将来その見返りが期待できるからであるという原理である。互いに互恵性原理を持った二者は，相手は互恵性原理を採用していると互いに信じることで協力による高い利得を獲得することができる。この互いの行動が特定のルールに従っていると互いに信じている状況を，ここでは規範（Kandori, 1992; Seinen and Schram, 2006; Pacheco et al., 2006; Chalub et al., 2006; Santos et al., 2018）と呼ぶ。互恵規範は広く社会において観察されるが，この成立には，協力し合う二者間が長期にわたり繰り返して，社会的ジレンマ状況で協力し続けるという前提が必要である。つまり長期的関係の維持は，互恵規範を有効に機能させるための必要条件となる（Sigmund, 2010）。

直接互恵性は，長期的関係が成立していない二者間における社会的ジレンマ状況における協力行動を説明できない。にもかかわらず，現代社会のように，見知らぬ他者とその場限りの関係しか保証されていない状況にもかかわらず，人々はしばしば頑健な協力する（Pancha-nathan and Boyd, 2003; Panchanathan and Boyd, 2004; Nowak 2006; Ohtsuki and Iwasa, 2006; Sasaki et al. 2017, Okada et al., 2017; Okada et al., 2018a）。これを説明する有力な理論の一つに間接互恵規範がある（Nowak and Sigmund, 1998a; Nowak and Sigmund, 1998b; Leimar and Hammerstein, 2001; Milinski et al., 2002; Nowak and Sigmund, 2005; Rockenbach and Milinski, 2006; Sommerfeld et al., 2007; Ohtsuki and Iwasa, 2007; Ohtsuki et al., 2009; Suzuki and Kimura, 2013; Martinez-Vaquero and Cuesta, 2013; McNamara and Doodson, 2015; Ghang and Nowak, 2015; Grimalda et al., 2016; Sasaki et al., 2016）。これは，私が協力するのは，協力しようとする対象

が, 以前, 第三者に協力していたということを知っていたからだという発想による。この場合，互恵性は間接的に成立しているとみなせる。つまり，その相手とは長期的関係を有していないが，その相手は過去に良いことをしたために協力する。間接互恵規範が成立していれば, 自分の協力行動は, 第三者が自分を良いとラベリングすることにつながり，そのラベリングが将来第三者からの協力行動を保証するという論理である。

多くの協力の進化研究の蓄積にもかかわらず，見知らぬ人々の間でなされる協力メカニズムとして，間接互恵規範は，懲罰や報酬といった誘因制度（van Lange et al., 2014）と並ぶ説明可能な数少ない説明原理の一つとなっている。この規範が成立するには，見知らぬ他者の協力行動に関する履歴の流通システムが存在しなければならない。すなわち，履歴情報が流通することで他者の評判が決定可能となる。言い換えると，間接互恵規範が機能するには評判情報による他者のラベリングが必要となる。

このとき流通される評判情報は，これから協力しようとする相手がどのような人物であるかを判断するために用いられる。単純に言えば，相手が「良い」評判を持っていれば協力し，「悪い」評判を持っていれば協力しない，というような判断に用いられることになる。ではその評判はどのように決定されるのであろうか。人々はある評判ルールに従って善悪の価値判断を行い，他者の評判情報を確定させている。つまり，何が「良い」と判断されるかは，評価ルールとして何が採用されているのかに依存することが分かる。よって，間接互恵性研究において，評価ルールの特定化は重要な研究課題となっている（Brandt and Sigmund, 2004; Ohtsuki and Iwasa, 2006; Uchida and Sigmund, 2010; Sigmund, 2012; Watanabe et al., 2014; Ohtsuki et al., 2015; Okada et al., 2017; Okada et al., 2018a; Yamamoto et al. 2017）。

Nowak and Sigmund（1998a）は，イメージ・スコアリングという評価ルールに関する理論的検討を行った。このルールでは，行為者が協力したかしなかったかで，その者のスコアを更新させ，協力数が非協力数を上回った場合に, その者を「良い」とみなす。もちろん，この規範の採用者は良いとみなされた者に対してのみ協力する。Nowak and Sigmund（2005）はそのルールはさらに単純化し，単純に直前の行動が協力・非協力のどちらかであったかで常に評判情報は更新されるとしたケースについて理論的な検討を行った。その結果，このルールは進化的安定性という動学的性質を有していないことを明らかにした。なぜなら，もしこのルールを採用しているものが，悪い人間に出会ったときは協力しないことになるが，この非協力行動は，このルール自身によって悪いと判定され，以後の協力を得られなくするからである。また，協力を意図しつつも何らかの問題からそれが実行できないというようなエラーに対しても，同じ理由から脆弱である。このようにイメージ・スコアリングとは，そのルールの単純性と引き換えに, 協力の進化によって致命的な「アキレス腱（Sigmund, 2010）」を有していることが明らかとなった。Sigmund（2010）はこれを「スコアリングのジレンマ」と名付けている。

このジレンマは，評価ルールを構築する際に，非協力行動をした動機をも考慮することの重要性を示しているともいえる。すなわち，非協力行動をとったのは，相手が悪いからなのか，それとも自分が非協力的だからなのかを識別できることは，スコアリングのジレンマを解消する。この点を考慮する，すなわち「正当化された裏切り」かどうかを考慮するには，評価ルールとして単に行動を見るだけではなく，誰にそれをしたのかまで見る必要がある。そこで理論研究では，前者を一次情報，後者を二次情報と名付け，二次情報あるいはそれ以上の情報を考慮した複雑な評価ルールについて検討を行ってきた。

Ohtsuki and Iwasa(2006) らの研究は，高次情報を対象に，イメージ・スコアリングが有していなかった安定性を有しており，しかも協力社会を維持できるものを網羅的に探索し，リーディング・エイトと名付けられた８つの評価ルールの特定化に成功した。これらには正当化された裏切り（悪人に対する非協力）は良いと判断されるなど，我々が規範として持っている価値観と整合的な性質を複数有している。理論家はその後も分析を続け，最近では Santos et al.（2018）らがさらに複雑な情報を考慮して体系化を試みている。

理論的な精緻化にも関わらず，「望ましい」規範の特定化は困難である。なぜなら実証分析の知見はしばしば理論分析の知見と矛盾しており，理論と実証のどちらが正しいのかという論争が終結していないからである。Milinski et al.（2001）の研究によると，被験者実験では，理論家が提唱するような高度な情報を用いている証拠は見られず，一次情報のみで意思決定をしているとの結果が支持されている。もちろん，これと対立し，コストがかかるにもかかわらず二次情報を取得し，それを意思決定に反映させているという実験結果（Swakman et al., 2016）も存在しており，実証的にも統一見解の確定は程遠い。いずれにせよ，理論的研究では一次情報だけでは協力体制を安定的に構築できる規範は存在せず，二次情報以上の複雑な評価ルールが必要であるという知見を導出しているにもかかわらず，実証研究では，一次情報だけで協力体制を十分構築可能であり，人々は理論家が求めるような複雑な評価ルールを用いていないと反論しており，その対立は深刻なままである。

この点を止揚すべく，我々は両アプローチに対して技術的な再検討を行った。その結果，理論分析でしばしば用いられている解析容易性を確保するための公的評価仮定を緩和し，実証分析で暗黙的に仮定されている情報取得の順序性を緩和することで，新たな規範の導出に成功した。以後，その点について詳細に検討する。

ほとんどの理論研究では，進化ゲーム理論を用いて厳密な解析解を導出するために公的評価を仮定している。公的評価とは，評判情報は対象者ごとに一意に特定化され，個々の評価者が自由に対象を評価することはできない状況を想定する。確かに評判とは，対象者に対する評価の代表値として一意に限定される傾向をもつものの，かといって個々の評価者が全く個別の評価を行えないというのは強すぎる制約と言わざるを得ない。評価ルールは規範として社会で共有されていたとしても，そのルールを用いた個別の評価は自由に行いうるのが現実的であろう。しかし，私的評価系を理論的に解析しようとすることは大きな困難を伴う。集団の構成員の数をNとした場合，公的評価系では評判情報は社会全体でN個となるが，私的評価系では誰の誰に対する評価かが個別に異なるため，Nの２乗だけ必要になる。また理論分析では，しばしば無限集団を仮定するが，そのとき，ある評価者が特定の対象者をどう評価するかを定義する方程式の数が無限に発散するため，厳密解の導出が非常に困難となる。そのような事情から，ほとんどの理論研究は公的評価系の分析に終始し，私的評価系の分析はわずかな研究が近似解を導出して検討しているに過ぎない（Uchida, 2010; Uchida and Sasaki, 2013; Olejarz et al., 2015）。

私的評価系を分析する場合，公的評価で重視されてきた「正当化された裏切り」が有効に機能しなくなると予想される。これは悪い評判を持っている者に非協力行動をとったとしても，その行為は悪いことと判断されないことを意味するが，このためには行為者と評価者で，被行為者に対する評価が一致していることが前提となる。一方，私的評価系では，行為者が正当化された裏切りと認識していても，評価者は行為者の認識とは異なり，良い評判を有している者への裏切りと映る場合があり，その場合は，行為者による正当化された裏

切りは評価者において正当化されない。この事態は，私的評価系において協力体制を維持する安定規範は公的評価系におけるそれとは異なることを暗示する。私的評価系においても正当化された裏切りがそれなりの機能をするには，評価者は裏切り行為に直面した際に，それを寛容に評価するルールを採用することが求められる。この点を解決すべく，我々は本稿において，評価者が悪いと判断しているものに対する行為は，その行為の評価を留保するという評価ルールを検討する。

次に，実証的に評価ルールの特定化を行おうとするアプローチについて概観する。実証分析の場合は，操作性を確保するため被験者実験を用いるのが一般的だが，そのとき被験者がどのような場面でどのような行動をしているのかのデータを収集することで，評価ルールの特定を行う。つまり，被験者に与える情報は実験者側が制御する。これまでは被評価者が何をしたのかという行動情報（一次情報）を与えた場合や与えなかった場合，またその取得にコストがかかるとして情報を積極的に取得するかどうかを観察して評価ルールとして一次情報を用いているかいないかを統計的に明らかにしてきた。その中で，理論研究が提供した一次情報不安定説を実証的に検討するため，一次情報取得者が，二次情報（被評価者がどのような評価を有する者にその行動をしたかに関する情報）を取得するかどうかを検証する実験が行われてきた。このような研究の蓄積によって，一次情報だけで充分であるとする説を支持する実験とともに，コストをかけてでも二次情報まで取得する説を支持する実験もあるなど，論争は終結していない。しかし，対象者を評価する際に，まず行動を見て，次にその行動を誰にしたかを確認するという順序性が常に保証されているとは考えにくい。例えば，泣いている人を見て（二次情報を先に取得し），その人が何をされたか確認して（一次情報を後に取得し），その行為をした者を評価するといった場面は十分に想定されうる。つまり，被験者実験が暗黙的に仮定していた情報取得の順序性は，それ自体検討すべき点なのである。この見過ごされた制約は，一次情報・二次情報というネーミング自体にその原因の一端があるかもしれない。

私的評価系の研究蓄積について概観した際，我々は留保規範について検討を行うと述べた。留保規範は被行為者が悪い評価を持っている場合は，行為の評価を留保するという評価ルールを採用するため，事実上，二次情報を先に検討していることになる。つまり，被験者実験において，これまでの実験研究が仮定していた情報取得の順序性が満たされないデータの存在は，留保規範の妥当性を示す根拠になりうると考える。

3　私的評価系の理論的検討

私的評価系を分析するため，進化ゲーム理論の枠組み（Hofbauer and Sigmund, 1998）に従い，無限のプレイヤー集団からなるゲームを想定する。プレイヤーは完全協力規範（=X），完全裏切り規範（=Y），間接互恵規範（=Z）のいずれか一つを自分の規範として採用しているものとする。Zは全プレイヤーに対し個別に良い（=G）か悪い（=B）かのいずれかのイメージを付与している。

無限の離散時間の中で，任意の時点では，一組の寄付者と受益者の組がランダムに選ばれる。寄付者はコスト 1 を支払って受益者に寄付するか，寄付せずコストを負担しないかのいずれかを選択する。寄付を選択した場合のみ，受益者は 1 より大きい利得rを得る。Xが寄付者になった場合は，相手によらず常に寄付を行い，Yが寄付者になった場合は，相手によらず常に寄付をしない。Zが寄付者になった場合は，受益者のイメージがGである時に寄付をし，Bである時は寄付しないものとする。

私的評価系を近似なく解析するため，寄付ゲー

ムごとの観察者をZの中の１人に限定する制約を与えることにする（Okada et al., 2018a）。つまり毎時点で観察者は常に一人だけランダムに選ばれるものとする。観察者は寄付者の行為と自分が持つ受益者のイメージに基づいて，寄付者のイメージを更新する。

系を一般化するためゲームに二種類のエラーを導入する。一つは行動エラー e1 で，寄付者はわずかな確率で寄付しようとする意図を実現できないものとする。もう一つは認知エラー e2 で，観察者は寄付者のイメージを更新する際，わずかな確率でその更新を誤るものとする。

このようなゲームを無限回繰り返すと，ZのタグづけるイメージがGとなる確率は，規範ごとに特定の値に収束する。この値を用いると各規範の期待利得を計算できることになるので，リプリケータ･ダイナミクス方程式を解くことができる。このようにして任意の間接互恵規範の動学分析が可能になる。本稿では，留保規範と比較するうえで，リーディング・エイトで特定化された規範のうち，比較的頑健に安定的な協力体制を維持する代表的なSimple-standing（SS）規範と Stern-judging（SJ）規範を取り上げ分析する。

各規範が採用する評価ルールを表１にまとめる。これらの間接互恵規範の動学分析を行った（図１）。その結果，公的評価系と比べ私的評価系では

１．協力を達成できる規範が限定的となる
２．均衡状態では完全協力規範と間接互恵規範の共存となる

表１　間接互恵規範の評価ルール

規範名	CtoG	DtoG	CtoB	DtoB
留保	G	B	K	K
SS	G	B	G	G
SJ	G	B	B	G

※CtoG，DtoG，CtoB，DtoB とはそれぞれ「Gに対して寄付する」，「Gに対して寄付しない」，「Bに対して寄付する」，「Bに対して寄付しない」状況を意味する。各規範はそれぞれの状況で寄付者のイメージをどう更新するかをG，B，Kであらわしている。それぞれ「Gとする」「Bとする」「イメージ更新を留保しこれまでのイメージのままとする」を意味する。

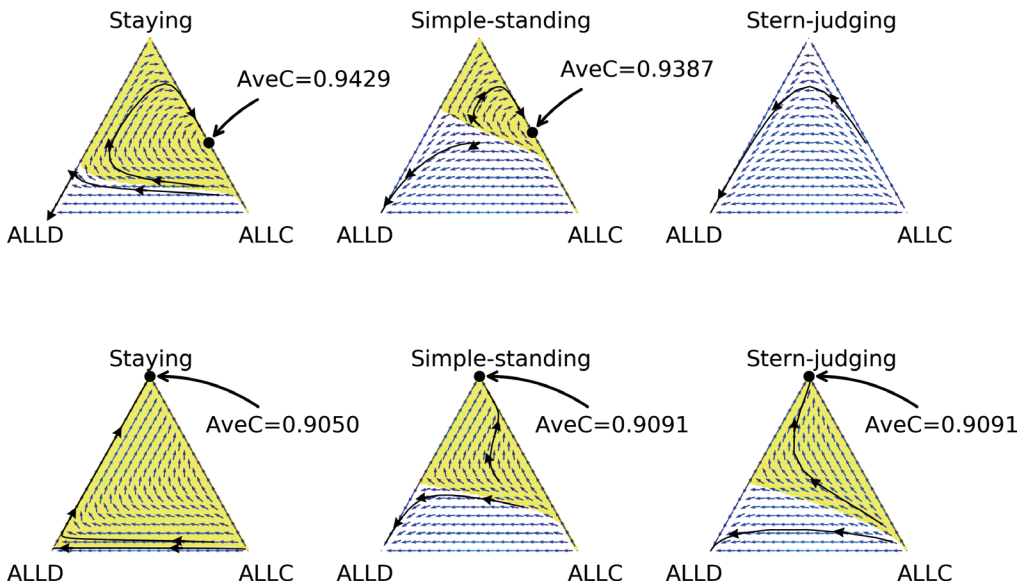

図１　間接互恵規範のリプリケータ・ダイナミクス分析結果

上段は私的評価系の，下段は公的評価系の，また左から順に，留保規範，SS規範，SJ規範の動学を示す。各三角形は集団構成員であるX, Y, Zの存在比を状態空間とする単位単体を表す。すなわち，三角形の内部や境界上の点において，そこから各辺への垂線を引いた足の長さと各規範の存在比とが一対一に対応し，その構成員でゲームをした場合，リプリケータ・ダイナミクス方程式に従って構成員存在比がどう変化するかを矢印で表している。図中の●点はその周辺の点を吸引する吸引点（局所均衡点）を表す。吸引点における平均寄付率を明示するとともに，吸引域を着色した。その領域の三角形全体に占める割合は，左上から順に74%，35%，０%，95%，52%，54%である。すべての動学においてr=３，e1=e2=５%である。

３．均衡状態では公的評価系よりも協力率を改善する

という特徴があることが明らかとなった。これはOkada et al.（2018a）の結果と整合している。

ところで，理論解析は解析可能性を確保するために，ゲームの観察者を一人に限定するといった極端な仮定を強いており，これを緩和した場合の結果に与える影響については不明である。この点を明らかにすべく，エージェントベース・シミュレーション（Okada et al.，2017）を実行して検討した。

シミュレーションでは，プレイヤー数Nは有限となり，ゲームの観察確率はqとした。図２に示されるように，図１における吸引点の集団構成比とシミュレーション結果の集団構成比がほぼ一致することから，私的評価系の挙動については一定の妥当性があると判断できる。

これらの結果から，私的評価系において頑健に協力体制を構築できると予測される留保規範の特徴を整理する。留保規範は，リーディング・エイトの代表的な戦略である SS規範と比べると，公的評価系においても私的評価系においても，協力体制への吸引域（図１の着色域）が広範囲にわたっている。特に，ほとんどすべてのプレイヤーが完全裏切り規範を採用している状態を表す頂点Y付近にも吸引域があることから，ほとんどが非協力者からなる社会構成であっても，協力体制を構築できる特徴を有していることが分かる。また，吸引点は非吸引域から遠いことから，安定状態に到達したのち，相当大きな突然変異によるミュータントの侵入にもその体制を頑健に維持できることが分かる。つまり，一度協力体制に達した場合は，集団構成比に大きな変化が生じても，協力体制を維持できる頑健性を有していることが分かる。また，その協力体制における平均協力率は高い。これらはすべて留保規範の優位性を示している。

４　被験者実験による評価ルールの特定

留保規範に関する理論解析により，その優位性が明らかとなったが，これが間接互恵規範として許容されているか否かについては，人間を対象と

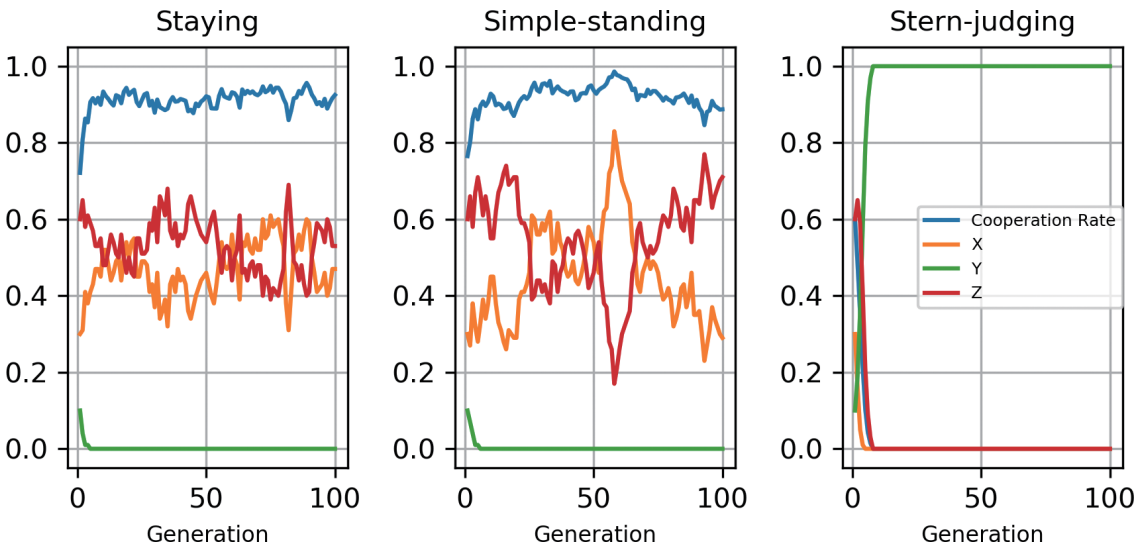

図２　間接互恵規範のエージェントベース・シミュレーション結果

左から順に，留保規範，SS規範，SJ規範のシミュレーション結果を示す。各グラフはX，Y，Zの各規範がフェルミ関数型学習過程（逆温度係数はβとする）に従って構成比を変化させるとしたときの，世代を単位とした時系列データを表す。各世代では初期状態としてZが持つ他者イメージをランダムに割り振ったのち，イメージの値を安定させるため，10万離散時間のシミュレーションを行い，最後の１万離散時間のゲームの平均利得を学習過程の適応度とした。すべてのシミュレーションにおいて，N=100，q=0.3，β=10，r=３，e1=e2=５% である。

した実験結果を待たねばならない。もし，留保規範を採用している場合，次のような情報行動と寄付行動の関係が見出されるはずである。

仮説１：二次情報を先に取得し，その値がGのとき一次情報を取得する確率は，その値がBのときに一次情報を取得する確率より高い。

仮説２：二次情報がBであるときは，取得情報が寄付行動を説明しない。

Okada et al.（2018b）では，この点を明らかにするための被験者実験を行った。実験は大学の学部生を対象にコンピュータ教室で複数回実施され，全部で152人分のデータを得た。各被験者は50ラウンド以上の寄付ゲームを行った。間接互恵状況を作るため，被験者はラウンドごとにランダムに選ばれた受益者に対して，その過去の寄付行動（一次情報）とその相手に関する過去５ラウンド分の寄付数（二次情報）を取得することができる。被験者は受益者の情報を取得したのちに寄付行動をするかしないかを選択する。

はじめに仮説１を検討するため，一次情報を先に開示し，その値が寄付した（=C）であったか，寄付しない（=D）であったかによって，二次情報の開示割合が変わるか，さらに，二次情報を先に開示し，その値がGであったかBであったかによって，一次情報の開示割合が変わるかについて，フィッシャーの正確確率検定を行った。その結果，一次情報を先に取得したときの二次情報開示率はCの時が27.5%，Dの時が28.8%となり，５%水準で帰無仮説を棄却できなかったのに対し，二次情報を先に取得したときの一次情報開示率はGの時が67.7%，Bの時が62.6%となり，帰無仮説を0.1%水準で棄却した。つまり，仮説１は支持されたことが分かる。

次に仮説２を検討するため，それぞれの情報取得行動について，寄付行動を被説明変数とするロジスティック回帰分析を行った。直近の行動，ラウンド数，現在の累積利得，直近の利益を統制変数として，一次情報と二次情報が寄付行動に有意に影響を与えるか検定したところ表２の結果を得た。この表は，二次情報を先に取得しそれがGであったときは，一次情報が寄付行動を説明するのに対し，Bであったときは一次情報も二次情報も寄付行動を説明しないことを示している。つまり，仮説２は支持されたことが分かる。

以上の検定から，留保規範は人間が採用しうる間接互恵規範として許容できることが明らかとなった。

5　議論

我々は頑健に協力体制を構築しうる間接互恵規範の有力な候補として留保規範を新たに定義し，その規範について理論・実験・シミュレーション

表２　被験者実験のロジスティック回帰分析結果

モデル	説明変数のうち有意になったもの
C型	直近の行動（正 ***）
	直近の利益（正 ***）
	現在の累積利得（正 **）
D型	**一次情報（正 ***）**
	二次情報（負 *）**
	直近の行動（正 ***）
	直近の利益（正 *）
	現在の累積利得（正 **）
	ラウンド数(負 **)
G型	**一次情報（正 ***）**
	直近の行動（正 ***）
	直近の利益（正 *）
	現在の累積利得（正 *）
	ラウンド数(負 **)
B型	直近の行動（正 ***）

モデルは上から順に，一次情報を先に取得しそれがCであった時，一次情報を先に取得しそれがDであった時，二次情報を先に取得しそれがGであった時，二次情報を先に取得しそれがBであった時に対応する。各説明変数について，係数が０であるとする帰無仮説を棄却したものを列挙し，係数がプラスであった時は「正」，マイナスであった時は「負」と記す。また棄却水準について 0.1%，１%，５%をそれぞれ ***，**，* で表す。

の各手法を用いて検討した。これまでの理論研究は公的評価系に限定した分析をしていたが，この枠組みでは正当化された裏切りが必ずしも正当化されないというジレンマが顕在化されない。このため理論解析が主張するほど，現実は協力を維持する規範が少ない可能性があるにもかかわらず，解析可能性と引き換えに，この点については多くの関心を払ってこなかった。

このジレンマが顕在化される私的評価系は，それまでの非現実的な制約を解消することができる一方，理論解析を大幅に困難にする。我々は観察者数を限定する新たな制約を設けることで，私的評価系の近似のない理論解析に成功した。一方，このために設けた新たな制約が結果に与える影響を考慮するため，補完的にエージェントベース・シミュレーションを行い結果の妥当性を確認した。

その結果，協力を維持する規範の特徴は，公的評価系での分析とは異なる特徴を有していることが明らかとなった。なかでも，私的評価系で顕在化される正当化のジレンマを解決するために評価を留保するという評価ルールについては，これまで全く検討されていなかったため，我々は新たに「留保規範」と名付け，これまで検討されてきた他の間接互恵規範と理論・シミュレーションの両面で比較した。その結果，この規範はこれまで知られていた規範よりも協力体制を頑健かつ安定的に維持できることが明らかになった。

私的評価系への理論的検討から特定した留保規範の優位性については理論的に明らかになったものの，実際に人間がそれを間接互恵規範として採用しているかどうかは別問題となる。我々はその点を確かめるために被験者実験を行った。その結果，留保規範の特徴である，寄付行動の受益者に関するイメージを先行して取得する情報行動や，そのイメージが悪い場合は寄付行動に関する情報をイメージ更新に用いないという評価ルールが統計的に有意に支持された。このことから，留保規範は間接互恵規範として許容されうることが示された。

このように留保規範の妥当性について，我々は理論と実証の両面から検討してきた。一方，間接互恵規範に対する理論研究と実証研究の主要な対立の一つである，どの情報を用いて評価ルールを構成するかという点に対しても，留保規範は止揚する立場を有していることが分かる。理論研究では一次情報だけでは進化的安定性を有しないため二次情報を用いた評価ルールが妥当であるとされてきた一方，実証研究では人間の実際の情報行動を分析した結果，そこまで複雑な評価ルールを採用していないのではないかと疑義を挟んでいた。

我々が検討した留保規範は，その情報取得順序は逆となるが，二次情報だけで評価する場合と，二次情報に加え一次情報を用いて評価する場合とが混在する。つまり，一次情報説と二次情報説の中間的な情報取得行動を示唆する。このように留保規範は理論と実証の対立を止揚する可能性を有しているといえる。

最後に理論・シミュレーション・実証の各手法を統合する意義について計算社会科学の文脈からまとめる。計算社会科学では，実証的に大量データの入手・解析が容易になってきた現状に対応するため，社会科学に計算論的視点を導入する立場を有している。Computational social science なる分野が欧米で提案されてきており，その邦訳として計算という語が充てられた。つまり，これまでよく使用されてきた計量（metric）あるいは統計（statistics）という語を用いていない。これは計量的方法論を排除しているのではなく，むしろ包摂した上位概念として，数値的あるいは数理的方法論を含んでいると想定される。また，これまでの社会科学に取られてきた方法論に対して，数理的な基礎づけを強調するという野心も含まれているであろう。

本研究では，通常の計算社会科学が用いるデータ分析は将来の課題となっている。しかし，理論と実証を統合して，これまでの社会科学がもたら

してきた知見を深化させるという本研究の目的は，計算社会科学の学問的な狙いと整合するのみならず，先導的な役割を果たす可能性があると思われる。

謝辞

本論文は，科学研究費補助金基盤研究（B）（17H02044）ならびに科学研究費国際共同研究加速基金（国際共同研究強化）（17KK0055）の助成を受けた研究に基づいたものである。

参考文献

Alexander, R.D.（1987）The biology of moral systems. New York: Aldine de Gruyter, USA.

Axelrod, R.（1984）The evolution of cooperation. New York: Basic Books, USA.

Boehm, C.（2012）Moral origins: The evolution of virtue, altruism, and shame. Basic Books, USA.（訳）斉藤隆史（2014）モラルの起源, 白揚社.

Brandt, H. & Sigmund, K.（2004）The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol. 231, pp.475-486.

————（2005）Indirect reciprocity, image scoring, and moral hazard. Proc. Natl. Acad. Sci. U.S.A. 102, pp.2666-2670.

————（2006）The good, the bad and the discriminator? Errors in direct and indirect reciprocity. J. Theor. Biol. 239, pp.183-194.

Chalub, F., Santos, F.C., Pacheco, J.M.（2006）The evolution of norms. J. Theor. Biol. 241, pp.233-240.

Fishman, M.A.（2003）Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, pp.285-292.

Ghang, W., Nowak, M.A.（2015）Indirect reciprocity with optional interactions. J. Theor. Biol. 365, pp.1-11.

Grimalda, G., Pondorfer, A., Tracer, D.P.（2016）Social image concerns promote coop-eration more than altruistic punishment. Nat. Commun. 7, 12288.

Hamilton, W.D.（1963）The evolution of altruistic behavior. Am. Nat. 97, pp.354-356.

Harari, Y.N.（2015）Sapiens: A brief history of humankind. Harper.（訳）柴田裕之（2016）サピエンス全史, 河出書房新社.

Hofbauer, J., Sigmund, K.（1998）Evolutionary Games and Population Dynamics. Cambridge University Press.

Kandori, M.（1992）Social norms and community enforcement. Rev. Econ. Stud. 59, pp.63-80.

Leimar, O., Hammerstein, P.（2001）Evolution of cooperation through indirect reciprocity. Proc. Natl. Acad. Sci. U.S.A. 268, pp.745-753.

Martinez-Vaquero, L.A., Cuesta, J.A.（2013）Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. Phys. Rev. E87, 052810.

Masuda, N., Ohtsuki, H.（2007）Tag-based indirect reciprocity by incomplete social information. Proc. R. Soc. B274, pp.689-695.

McNamara, J.M. & Doodson, P.（2015）Reputation can enhance or suppress cooperation through positive feedback. Nat. Commun. 6, 6134.

Milinski, M., Semmann, D., Bakker, T.C.M. & Krambeck, H.J.（2001）Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. B268, pp.2495-2501.

Milinski, M., Semmann, D. & Krambeck, H.J.

(2002) Reputation helps solve the 'tragedy of the commons'. Nature 415, pp.424-426.

Nakai, Y., Muto, M. (2005) Evolutionary simulation of peace with altruistic strategy for selected friends. J. Socio-Inf. Stud. 9, pp.59-71.

———— (2008) Emergence and collapse of peace with friend selection strategies. J. Artif. Soc. S11(3), No. 6.

Nowak, M.A., Sigmund, K. (1998a) The dynamics of indirect reciprocity. J. Theor. Biol. 194, pp.561-574.

———— (1998b) Evolution of indirect reciprocity by image scoring. Nature 282, pp.462-466.

———— (2005) Evolution of indirect reciprocity. Nature 437, pp.1291-1298.

Nowak, M.A. (2006) Five Rules for the Evolution of Cooperation. Science 314, pp.1560-63.

Ohtsuki, H., Iwasa, Y. (2004) How should we define goodness? Reputation dynamics in indirect reciprocity. J. Theor. Biol. 231, pp.107-120.

———— (2006) The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. 239, pp.435-444.

———— (2007) Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J. Theor. Biol. 244(3), pp.518-531.

Ohtsuki, H., Iwasa, Y., Nowak, M.A. (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457, pp.79-82.

———— (2015) Reputation effects in public and private interactions. PLoS Comput. Biol. 11. E1004527.

Okada, I., Sasaki, T. & Nakai, Y. (2017) Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. Sci. Rep. 7, 9737.

———— (2018a) A solution of private assessment in indirect reciprocity using solitary observation. J. Theor. Biol. 455, pp. 7-15.

Okada, I., Yamamoto, H., Sato, Y., Uchida, S., Sasaki, T. (2018b) Experimental evidence of selective inattention in reputation-based cooperation. Sci. Rep. 8, 14813.

Olejarz, J., Ghang, W., Nowak, M.A. (2015) Indirect reciprocity with optional interactions and private information. Games 6, pp.438-457.

Pacheco, J.M., Santos, F.C. & Chalub, F.A.C. (2006) Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput. Biol. 2, e178.

Panchanathan, K. (2011) Two wrongs don't make a right: the initial viability of different assessment rules in the evolution of indirect reciprocity. J. Theor. Biol. 277, pp.48-54.

Panchanathan, K. & Boyd, R. (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J. Theor. Biol. 224, pp.115-126.

Panchanathan, K., Boyd, R. (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432, pp.499-502.

Rockenbach, B., Milinski, M. (2006) The efficient interaction of indirect reciprocity

and costly punishment. Nature 444, pp.718-723.

Santos, F.P., Pacheco, J.M. & Santos, F.C. (2016) Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. Sci. Rep. 6, 37517.

Santos, F.P., Santos, F.C. & Pacheco, J.M. (2018) Social norm complexity and past reputations in the evolution of cooperation. Nature 555, pp.242-245.

Sasaki, T., Okada, I., Nakai, Y. (2016) Indirect reciprocity can overcome free-rider problems on costly moral assessment. Biol. Lett. 12. 201160341.

Sasaki, T., Yamamoto, H., Okada, I. & Uchida, S. (2017) The Evolution of Reputation-Based Cooperation in Regular Networks. Games 8 (1), 8.

Seabright, P. (2010) The company of strangers: A natural history of economic life. Princeton Univ. Press, USA. (訳) 山形浩生, 森本正史 (2013) 殺人ザルはいかにして経済に目覚めたか？, みすず書房.

Seinen, I., Schram, A. (2006) Social status and group norms: Indirect reciprocity in a repeated helping experiment. Eur. Econ. Rev. 50, pp.581-602.

Siegel, J.Z., Mathys, C., Rutledge, R.B., Crockett, M.J. (2018) Beliefs about bad people are volatile. Nat. Hum. Behav. 2, pp.750-756.

Sigmund, K. (2010) The Calculus of Selfishness. Princeton Univ. Press.

———— (2012) Moral assessment in indirect reciprocity. J. Theor. Biol. 299, pp.25-30.

Sommerfeld, R.D., Krambeck, H.J., Semmann, D., Milinski, M. (2007) Gossip as an alternative for direct observation in games of indirect reciprocity. Proc. Natl. Acad. Sci. U.S.A. 104(44), pp.17435-17440.

Sugden, R. (1986) The Economics of Rights, Cooperation and Welfare. Oxford: Basil Blackwell, USA.

Suzuki, S., Kimura, H. (2013) Indirect reciprocity is sensitive to costs of information transfer. Sci. Rep. 3, 1435.

Swakman, V., Molleman, L., Ule, A. & Egas, M. (2016) Reputation-based cooperation: empirical evidence for behavioral strategies. Evol. Hum. Behav. 37, pp.230-235.

Takahashi, N., Mashima, R. (2006) The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. 3. J. Theor. Biol., 243, pp.418-436.

Toriumi, F., Yamamoto, H., Okada, I. (2016) Exploring an effective incentive system on a groupware. J. Artif. Soc. S19(4), No. 6.

Trivers, R.L. (1971) The Evolution of Reciprocal Altruism. Q. Rev. Biol. 46, pp.35-57.

Uchida, S. (2010) Effect of private information on indirect reciprocity. Phys. Rev. E82, 036111.

Uchida, S., Sasaki, T. (2013) Effect of assessment error and private information on stern-judging in indirect reciprocity. Chaos Solitons Fract. 56, pp.175-180.

Uchida, S., Sigmund, K. (2010) The competition of assessment rules for indirect reciprocity. J. Theor. Biol. 263, pp.13-19.

Van Lange, P.A.M., Rockenbach, B., Yamagishi, T. (2014) Reward and punishment in social dilemmas. Oxford Univ. Press, USA.

Watanabe, T., Takezawa, M., Nakawake, Y.,

Kunimatsu, A., Yamasue, H., Nakamura, M., Miyashita, Y., Masuda, N.（2014）Two distinct neural mechanisms underlying indirect reciprocity. Proc. Natl. Acad. Sci. U.S.A. 111(11), pp.3990-3995.

Wedekind, C. & Milinski, M.（2000）Cooperation through image scoring in humans. Science 288, pp.850-852.

山岸俊男（1990）社会的ジレンマのしくみ―「自分１人ぐらいの心理」の招くもの，サイエンス社

Yamamoto, H., Okada, I., Uchida, S., Sasaki, T.（2017）A norm knockout method on indirect reciprocity to reveal indispensable norms. Sci. Rep. 7, 44146.

Yamamoto, H., Okada, I., Taguchi, T., Muto, M.（2019）Effect of voluntary participation on an alternating and a simultaneous prisoner's dilemma. Phys. Rev. E100(3), 032304.

Yoeli, E., Hoffman, M., Rand, D.G. & Nowak, M.A.（2013）Powering up with indirect reciprocity in a large-scale field experiment. Proc. Natl. Acad. Sci. USA 110, pp.10424-10429.