規範と利害に関わる承認理論の解釈
――イメージスコアモデルと評判動学による解釈――

川　村　哲　也

Ⅰ　序論

本論文は，八木（1998）［3］（1999）［4］Yagi（2001）［13］の，規範と利害に関わる承認理論（以下，承認理論）を，2つの進化ゲームモデルから解釈する。まず初めに承認理論の概要を説明しよう。

承認理論は，「社会状態」を「各人の財に対する支配に関して，当事者による正当化と社会による正当化が一致している状態」と定義する。この「社会状態」の動態を説明・分析する承認理論は，各人の現実世界での財に対する支配関係を記述する実定的領域と，その関係を正当化すると同時に，その関係から形成される規範的領域の2つの領域を前提とする。

規範的領域内では個別規範と一般規範の間の循環運動が存在する。ここで，一般規範は個別規範を包摂すると同時に，一般規範内の個別規範が実定的領域での正当化を通じて一般規範に反射し，一般規範の変化を促す。本論文では，一般規範を全ての個別規範と矛盾しない規範と定義する。すなわち，各人の財所有を正当化する個別規範の中で，全ての人の財所有において相互に矛盾しない個別規範の共通項を一般規範とする。

実定的領域内では各人の個別利害の間の連帯を介して類型的利害が生まれる利害の水平運動が存在する。各人は個別の利害判断にもとづき財の所有を承認しあう。財の相互承認にあたって各人は，互いに相手から財の所有を承認される立場にあるという利害状況の近さから，財の承認を求める相手の財所有に対して共感を覚える。

規範的領域と実定的領域の間には垂直の循環運動がある。規範的領域で決まる規範は，実定的領域における財所有のあり方に正当化を与える（下向き運動）と同時に，実定的領域で定まる財所有のあり方が規範による正当化を要求する（上向き運動）。

本論文は，この規範と利害の水平・垂直運動を表現する2つのモデルを紹介し，承認理論の解釈を行う。この解釈から，規範と実定的領域における承認方法の間の動学を描写することができる。ただし，承認理論が想定する個人間の相互作用のあり方にいくつかの制約を加える。

1　規範と利害に関わる承認理論

本節では，承認理論の基本モデルを説明する。承認理論は，「各人の財に対する支配に関して，当事者による正当化と社会による正当化が一致している状態」である「社会状態」の成立を，各人が互いに財の所有の承認を要求しあう2つの領域から説明した。2つの領域とは，各人の財の所有に正当性を与える規範の間の関係を描写する規範的領域と，各人の財所有の利害関係を描写する実定的領域である。この2つの領域で，相互に財の承認とその正当化を求める2人の個人の相互作用を通して承認理論を説明する。本論文では，実定的領域における2人の個人の利害関係を標準形ゲームの形で記述する。ただし，標準形ゲームへの変換によって，各人の利害関係のあり方に制限を加える。

1-1 規範的領域

規範的領域は，各人が自分の財の所有を正当化する規範の相互作用として記述できる。個人 i が財 p を所有することを D_{ip} と表す。更に，D_{ip} を自分の規範 $N_i(D_{ip})$ から正当化することを $L_i(N_i(D_{ip}))$ と表す。同様に，個人 j が財 q を自分の規範から正当化することを $L_j(N_j(D_{jq}))$ と表す。

自らの財所有に関する規範の適用方法に応じて2つのタイプのプレイヤが想定される。1つは，自らの財所有を自分の規範で正当化する自律的タイプで，他方は相手の規範で正当化を行う同調的タイプである。

ここで，各人 (i, j) のタイプに応じて，以下の4つの相互承認の形が考えられる。

- ケース1（i が自律的承認を行い，j も自律的承認を行う場合）：
$$L_i(N_i(D_{ip}))=L_j(N_j(D_{ip}))$$
$$L_i(N_i(D_{jq}))=L_j(N_j(D_{jq}))$$
- ケース2（i が同調的承認を行い，j も同調的承認を行う場合）：
$$L_i(N_i(D_{ip}))=L_j(N_i(D_{ip}))$$
$$L_i(N_j(D_{jq}))=L_j(N_j(D_{jq}))$$
- ケース3（i が自律的承認を行い，j は同調的承認を行う場合）：
$$L_i(N_i(D_{ip}))=L_j(N_i(D_{ip}))$$
$$L_i(N_i(D_{jq}))=L_j(N_j(D_{jq}))$$
- ケース4（i が同調的承認を行い，j は自律的承認を行う場合）：
$$L_i(N_i(D_{ip}))=L_j(N_j(D_{ip}))$$
$$L_i(N_j(D_{jq}))=L_j(N_j(D_{jq}))$$

ケース1では，規範の衝突が生じる。自律的承認タイプどうしが相対し，生じる規範の衝突は，実定的領域においてより多くの財所有を正当化できる規範が他の規範を淘汰する規範の競争をもたらすだろう。これは，規範的領域と実定的領域の間の垂直関係が規範どうしの水平運動に影響し，一般規範 N_g の形成に影響する構造を表す。すなわち，ケース1では，実定的領域において最も多くの財所有を正当化する個別規範 N_i が一般規範 N_g として成立し，各人の財所有の正当性の与え手 $L_g(N_g)$ として機能する。

一方で，ケース2では，相互に相手の規範を尊重しあうので規範の衝突は生じない。同調的承認タイプどうしが相対する場合，各人は相手の規範を尊重しあうので，各人の財所有を正当化する一般規範は多様になる。もしも，社会の全ての個人が同調的であれば，財の承認において個々人の規範全てが一般規範となりうる。この場合，任意の N_i が一般規範 N_g の候補となる。

しかし，ケース1でもケース2でも，各人が同じ承認方法を採用するとき，規範全体を包摂する一般規範がどのような形態をとるのかは自明ではない。

一方で，ケース3とケース4の場合，自律的承認タイプの規範が同調的承認タイプの規範に優位する。このとき，同調的タイプの財所有は自律的タイプの規範によって正当化される。もしも，社会に1人の自律的なタイプが存在し，他の人は全て同調的なタイプだとすれば，社会全体の財所有は自律的なタイプの規範に準じて正当化されることになる。

本論文では，ケース2に表される規範の同調的承認がどのような一般規範を形成するかを考察する。後の章で説明する評判動学において，各人は第三者から自分の財所有の正当化を受ける。同時に，各人は，第三者に対する正当化の与え手でもある。また，各人が第三者として下す実定的領域における財所有の正当性は，社会全体で共有される。各人の採用する規範は第三者の正当化する財の承認方法を規定し，実定的領域において各人が許容する財の承認方法を決める。各人が許容する実定的領域の承認方法が自分の承認方法と適応的ならば，各人の規範は

社会の中で存続する。どのような規範が，各人の規範が衝突しない一般規範が形成される均衡で達成されるのかは，実定的領域からの正当化の要請と，その正当化がもたらす承認方法の適合度に応じて決まる。

本論文で考察する一般規範は，実定的領域と規範的領域の垂直運動の中から選抜され，一般規範どうしは同調的に承認される。本論文は，規範的領域における規範の衝突が，実定的領域において正当化される承認方法の間の衝突として間接的に生じる場合のみを考察する。

1-2　実定的領域

実定的領域は，各人の互いの財に対する承認行動の相互作用として記述できる。実定的領域において，各人は各人の個別利害に応じて財の承認を行うと同時に，他人の財所有に対しても共感を覚える類型的利害関係にもある。いま，財の相互承認を行う２人の個人(i, j)を考える。各人(i, j)は互いに自分の財(p, q)に対する承認を求めている。ここで，$i(j)$が財$p(q)$を支配する$i(j)$の利害判断を$I_i(D_{ip})(I_j(D_{jq}))$と表す。更に，$I_i(D_{ip})(I_j(D_{jq}))$に対する$i(j)$の感情を$S_i(I_i(D_{ip})(S_j(I_j(D_{jq}))))$とする。

i, jは規範的領域と同様に自律的承認タイプと同調的承認タイプの２つのタイプを持つとする。自律的承認タイプは自分の財の所有だけでなく，相手の財を自分が所有することにも利害関心を持つエゴイストとする。iを自律的承認タイプとすると，iが財pを支配することに対する利害関心は$S_i(I_i(D_{ip}))$である。一方，同調的承認タイプは相手が財を所有することに共感を示す。jを同調的承認タイプとすると，jが財qを支配することに対する利害関心は$S_j(I_j(D_{jq}))$である。

本論文では，実定的領域における財の相互承認に以下の制約を課す。同じタイプどうしが相対した場合，それぞれ自分の財の支配が承認されるとする。自律的承認タイプと同調的承認タイプが相対した場合，同調的タイプの財にも関心を示す自律的承認タイプが全ての財を支配する。本論文では，実定的領域における同じタイプの個人の利害関心と財の対称性を仮定する。このとき，自律的タイプと同調的タイプが財の相互承認を行うときのiの利得は，行プレイヤをi，列プレイヤをjとすると，表１のようになる[1)]。

同調的タイプどうしが財の相互承認を行う場合，互いに自分の財が承認されると同時に，相手が相手の財を所有することからも共感による利益を感じる。自律的タイプと同調的タイプが相対する場合，自律的タイプは，自分の財と同調的タイプの財を共に支配する拡張的エゴイストになる。このとき，同調的タイプは自分の財が取り上げられ，相手の財の支配に対する共感からしか利益を得ることができない。自律的タイプどうしが相対する場合は，互いに自分の財を支配するだけで，他人の財支配に対する共感は存在しない。

この実定的領域における２人の個人の相互承認を記述するゲームの利得表は囚人のジレンマの構造になる。ここで，財の対称性と各タイプ

表１　実定的領域の財の相互承認の利得表

	同調的タイプ	自律的タイプ
同調的タイプ	$S_i(I_i(D_{ip}))+S_i(I_i(D_{jq}))$	$S_i(I_i(D_{jq}))$
自律的タイプ	$S_i(I_i(D_{ip}))+S_i(I_i(D_{iq}))$	$S_i(I_i(D_{ip}))$

1）ここで，利得という言葉は，各人の人格（本論文では自律的か同調的かの２つのタイプ）にもとづく感情的な利害関心を含んだ概念で，財の所有から得られる便益のみを規定するものではない。

の利得の対称性を仮定する。この仮定から，$S_i(I_i(D_{ip}))=S_i(I_i(D_{iq}))$ が成立する。この利得表において，相手のタイプに関わらず自分が同調的タイプから自律的タイプに切り替える場合の利得の差は一定で，$S_i(I_i(D_{jq}))-S_i(I_i(D_{ip})$ となる[2)]。

第Ⅱ章で，実定的領域における財の相互承認が，表1の利得表を持つ2人進化ゲームで表現されるモデルから承認理論の解釈を試みる。

1-3 社会状態

各人 (i) の財 (p) の所有が社会的に承認されるためには，次の条件が必要である。

$$L_i(N_i(D_{ip}))=L_g(N_g(D_{ip})) \quad (1)$$

(1) 式は，当事者による財の所有の正当化と社会による正当化が一致している「正義」の状態である。(1) 式は，実定的領域における財の所有を正当化する個別規範の均衡条件である。本論文では，一般規範 N_g を，全ての個別規範 N_i と矛盾しない規範と定義しているので，(1) 式が意味するのは全ての個人 i が自分の規範 N_i にもとづいた財の所有の正当化 $L_i(N_i)$ が，社会からの正当化 $L_g(N_g)$ と一致し，その判断が「正義」とみなされていることを示す。注意すべきは，$i\neq j$ なる N_i，N_j について，$L_i(N_i)=L_j(N_j)=L_g(N_g)$ を満たす可能性があることである。すなわち，社会の正当化の与え手である一般規範 N_g は複数の異なる個別規範を同時に満たす可能性がある。

$L_g(N_g)$ が正当化する財所有の範囲，すなわち一般規範の圏域が問題である。この問題は一般規範の正当化する財所有の構造の実定的領域における安定性の問題に他ならない。すなわち，実定的領域において一般規範に含まれる個別規範の正当化する財の所有構造が持続可能であるか否かが問題である。

各人 (i) の財 (p) の所有が実定的領域において持続可能であるためには，次の条件が満たされなければならない。

$$S_i(I_i(D_{ip}))=S_g(I_g(D_{ip})) \quad (2)$$

ここで，$I_g(D_{ip})$ は，i が p を支配することの社会全体から見た利害判断であり，$S_g(I_g)$ はその利害判断にもとづく社会感情を表す。(2) 式は，i が p を支配することの利害判断 I_i にもとづく感情 S_i が社会一般においても利益になるとの判断 I_g にもとづく感情 S_g と一致することを示す。

本論文では，実定的領域における各人の財の承認関係を標準形ゲームで表した。このとき，(2) 式は，実定的領域における財の相互承認の結果が社会的に望ましい配分になっていることを要求する。すなわち，(2) 式は，実定的領域での財の承認関係がナッシュ均衡ではなく，パレート最適なものになっていることを要求する。従って，実定的領域における財の承認方法は，支配戦略である自律的承認ではなく，同調的承認でなければならない。更に，本論文では，実定的領域を進化ゲームで記述する。従って，(2) 式を満たす承認方法の組み合わせは進化的に安定な戦略 (ESS) とならなければならない。ESS 条件が成立しなければ，社会の厚生を最大化し，社会全体の利害判断と一致する状態であっても，その社会の外から新たに加わる成員によってその状態が持続可能でなくなってしまうためである。(1) 式，(2) 式を同時に満たす一般規範とその正当化する財の所有構造が，「社会状態」である。このとき，(1) 式，(2) 式を同時に満たす一般規範の圏域がどのように定まるかを考察していく。

まず，全ての個人が1つの同じ規範に従って財の承認を行うモデルを紹介する。このとき，個別規範＝一般規範であり，「社会状態」を正当

2）この利得表の条件は，以下に述べる2つのモデルの結果を用いるために必要不可欠である。本論文の考察は，実定的領域における利得表がこの条件を満たす囚人のジレンマとなる場合のみに適用できる。

化する一般規範の圏域についてのみ論じる（第Ⅱ章がそれである）。次に，多数の個別規範にもとづき財の承認を行うモデルを紹介し，一般規範の成立過程と，その圏域について論じる（第Ⅲ章がそれである）。

Ⅱ　規範としてのイメージスコア：個別規範＝一般規範モデル

1　イメージスコアモデル

Nowak and Sigmund（1998a）［9］（1998b）［10］は，２人進化ゲームにおいて，相手の社会におけるイメージが観察できるモデル（以下，イメージスコアモデル）を用いて，間接互恵性のメカニズムを説明した。イメージスコアモデルでは，無限に大きな人口から２人がランダムに選ばれ，それぞれ相手を助けるか否かの選択を同時に行う。相手を助ける場合，その行動にはコスト c がかかり，相手から助けてもらうと利益 b を受け取ることができる。１世代の中でランダムにピックアップされた個人が複数回，相手を助けるか否かの意思決定を同時に行うラウンドを持つ。各人の適合度は，このラウンドにおける個人の利得であり，レプリケータダイナミクスによって次世代の各人の割合が決まる。世代間では，突然変異が起こり，人口のうちの少数の割合がランダムに選ばれた戦略に変化する。１ラウンド内での意思決定の回数は２回以上で進化的に安定な状態の性質は異ならないので，議論を簡単にするため，本論文は１ラウンドに平均２階の意思決定が行われる結果を用いて承認理論の解釈を試みる。

Nowak and Sigmund（1998a）では，進化ゲームをプレイするプレイヤは，対戦相手のイメージを整数の形で持つ。プレイヤのイメージは，ランダムにピックアップされた過去の対戦相手を助けた回数分だけ上昇し，助けなかった分だけ減少する。Nowak and Sigmund（1998b）では，相手が１期前の対戦相手に対して裏切りを選択していたなら，Bad（B），協力を選択していたなら Good（G）の２値を相手のイメージとして割り当てる。イメージスコアモデルは，十分に大きな人口の中からランダムに対戦相手がピックアップされることを仮定しており，同じ対戦相手と再びゲームを行うことはない。従って，ある期における相手のイメージはゲームに参加しない第三者に対する一期前の行動に依存して決まる。

承認理論における規範 N_i は，各人 (i) の財所有 (D_{ip}) を正当か否か判定する（$L_i(N_i(D_{ip}))=$｛正当，正当でない｝）。各人のイメージは，各人の財所有に対する規範からの正当化の有無として解釈できる。そこで，本章ではイメージの値が２値をとる Nowak and Sigmund（1998b）のモデルで承認理論の解釈を行う。このとき規範はプレイヤのイメージの決め方を規定するルールであり，これは全てのプレイヤに共通である。従って，本章で解釈を行うモデルは，個別規範＝一般規範という制約を課した承認理論の一部の動学を表現する。

次に，このモデルが承認理論のモデル候補として妥当であることを論じる。

このモデルは，相手を助けない行動は常に悪であり，助ける行動は常に善であるという１つの規範が社会の中で共有されている場合，この規範に則って，善いものを助け，悪いものを助けないという戦略が進化的に安定な戦略（ESS）となることを示した。

ここで，このモデルが承認理論を部分的に説明するのに都合のよい性質を備えていることを示す。第一に，このモデルは，個体の適合度を定める現実の相互作用関係と，その相互作用関係と相互依存する規範を備えているためである。実際，以下で述べるようにこのモデルにおけるイメージスコアを規範とみなすことは自然である。

イメージスコアの基本的アイデアは，Trivers の *Social Evolution*（1985）［12］でいう「正

義」が支える道徳システムの中で社会が間接互恵性を達成することを示すための仕組みとして考案された。この道徳システムの理論はAlexander（1987）［5］で理論化され，この理論の具体例としてイメージスコアモデルを位置づけることができる。イメージスコアは，社会に存在する行動の中で，どのような行動が正当化されるかを表す1つの指標であり，承認理論の定義する個別規範の定義を満たす。ただし，イメージスコアモデルでは，常に1種類の規範しか存在しないため，先に述べたように個別規範が一般規範と一致しており，承認理論の規範的領域における水平運動を説明することはできない。しかし，一般規範が成立している場合の実定的領域と一般規範の間の垂直循環から一般規範の圏域について論じることはできる。次に，このモデルから承認理論を解釈し，一般規範の圏域について論じる。

2 イメージスコアモデルにおける一般規範の圏域

まず始めに，前章で定義した承認理論の実定的領域の利得表をイメージスコアモデルの利得表に対応させる。イメージスコアモデルは，十分に大きな集団からなる社会から2人のプレイヤがランダムに選ばれ，相互に相手を助けるか否かの意思決定を同時に行う。各期の意思決定の際に，各プレイヤは相手のイメージを観察できる。このイメージは，1期前に相手が自分とは異なる対戦相手を助けていたなら「善い」，そうでなければ「悪い」。このイメージを利用して各プレイヤは意思決定を行う。

イメージスコアモデルにおける「相手を助ける」選択が，承認理論の「同調的承認」に対応し，「助けない」が「自律的承認」に対応する。このとき，二つの利得表は，同じクラスの囚人のジレンマとなる。イメージスコアモデルで相手の行動に関わらず，相手を助けたときにかかるコストは $-c$ で一定であり，承認理論では，相手のタイプに関わらず相手に対して自律的な承認から同調的な承認に切り替えることによる利得の減少分は $S_i(I_i(D_{jq}))-S_i(I_i(D_{ip}))$ で一定である。このモデルは，自分の選択を変えた場合の利得の変化が相手の選択に依存しないクラスの囚人のジレンマを仮定している[3)]。

イメージスコアモデルには3つの戦略がある。常に相手を助ける無条件協調戦略，常に相手を助けない無条件非協調戦略，相手のイメージが善ければ助け，悪ければ助けない区別戦略である。同様に，承認理論にも3つの戦略に対応する個人を想定する。すなわち，常に相手に対して同調的承認を行う同調的承認者と，常に相手に対して自律的承認を行う自律的承認者，相手の財所有が一般規範によって正当化されていれば同調的承認を行い，そうでなければ自律的承認を行う区別的同調者である。イメージスコアモデルにおける善いプレイヤは，承認理論における財所有が正当化されたプレイヤに対応し，悪いプレイヤは正当化されていないプレイヤに対応する。規範を参照するプレイヤは，区別的同調者のみである。

区別的同調者は，相手の財所有が一般規範に照らして正当化か否かに応じて，相手の財の承認方法を変えるが，ゲームの開始時点においては相手の財所有が正当か否か判定できない。本論文では，区別的同調者はゲーム開始時に相手の財所有方法は常に正当と見なすと仮定する[4)]。

3）この仮定は，一般規範の圏域を解析する上で必要となる。この仮定が満たされない場合，一般規範の圏域が実定的領域における各人の行動の分布に対して収束しない可能性が出てくる。

4）区別的同調者が，ゲーム開始時の相手の財所有を確率 p で正当とみなす場合を考える。イメージスコアモデルでは，p を進化の対象として扱うと，本論文で扱う囚人のジレンマのクラスのどの利得表領域でも，承認方法の分布に関わらず，$p=1$ に進化する。

イメージスコアモデルで表現した承認理論において，社会状態を正当化する一般規範の圏域について考察を行う。イメージスコアモデルでは個別規範＝一般規範なので，1章の(1)式は常に成立する。従って，社会状態を正当化する一般規範の圏域を検証するためには，一般規範に従って相手の財の承認を行う区別的同調者が，1章の(2)式を満たす場合を考察すればよい。

(2)式は，実定的領域で全ての個人が同調的承認を行い，かつその状態がESSでなければならないことを意味する。このとき，全ての個人の財所有は一般規範によって正当化されるが，全ての個人が一般規範を実際に参照するわけではない。この一般規範を参照し，自律的承認者が社会の中に入り込んできたら自律的な対応をとる区別的同調者の割合によって，(2)式が成立するか否かが決まる。

実際に，このモデルの結果を用いて，社会状態を正当化する一般規範の圏域について論じる。まず，(2)式が成立するためには，区別的同調者の割合が $\frac{S_i(I_i(D_{ip}))}{S_i(I_i(D_{jq}))}-1$ より高くなくてはならない。この水準は，相手の財所有に共感することによって得られる類型的利害関係からの利得が高いほど低くなり，自分の財所有から得られる利得が高いほど高くなる。すなわち，類型的利害関係が相対的に強い社会ほど，狭い一般規範の圏域によって「社会状態」を維持できる。

そうでない場合，(2)式が成立しないどころか，自律的承認者のみで構成される社会がESSとなる。このとき，一般規範は全ての自律的承認者の財所有を不当とみなすが，一般規範によって財の所有が正当化される同調的承認者と区別的同調者は，実定的領域における淘汰圧によって社会から排除され，また侵入することができない。従って，一般規範を参照する区別的同調者の割合が低いと一般規範は正当化の圏域を持たない。

区別的同調者と同調的承認者のみが社会の構成員となり，自律的承認者からの侵入に対して頑健となる条件から，一般規範の圏域を特定する。区別的同調者の割合が $\frac{S_i(I_i(D_{ip}))}{S_i(I_i(D_{jq}))}-1$ を上回る場合，同調的承認者と区別的同調者からなる進化的安定な割合が唯一つ定まる。その割合は，$\frac{S_i(I_i(D_{ip}))}{S_i(I_i(D_{jq}))}-1$ よりも区別的同調者の割合が必ず高くなる[5)]。自律的承認者が，この状態に侵入しても最終的には淘汰され，社会状態が維持される。

重要なのは，区別的同調者と同調的承認者のみが社会の構成員となる状態では，社会規範の衝突は存在しないにもかかわらず，相手の財の承認において常に一般規範を参照する区別的同調者が一定割合存在することである。イメージスコアモデルでは，正当化の与え手となる一般規範の圏域は，区別的同調者の割合が $\frac{S_i(I_i(D_{ip}))}{S_i(I_i(D_{jq}))}-1$ を上回る実定的領域の社会状態に限定される。

Ⅲ　個別規範ダイナミクス

Ohtsuki and Iwasa（2004）［6］（2006）［7］は，プレイヤの評判を導入した進化ゲームモデル（以下，評判動学）を構築し，各人が対戦相手の評判を第三者を通じて獲得できる社会において，各人の評判と行動の進化的安定性について論じた。本論文では，評判動学における評判を承認理論における個別規範と読み替え，規範的領域と実定的領域が相互に影響しあうモデル

5）この割合が，どの程度 $S_i(I_i(D_{ip}))/S_i(I_i(D_{jq}))-1$ から離れるかは1世代内で，同じ相手と何度財の相互承認を行うかに依存する。

における一般規範の圏域について論じる。

評判動学は，前章のイメージスコアモデルと同様の利得表を仮定した２人進化ゲームモデルである。イメージスコアと同様に，世代間でレプリケータダイナミクスによる戦略の淘汰と突然変異が生じる。しかし，１世代の中で一人のプレイヤは同じ対戦相手とただ一回のゲームしか行わない[6]。イメージスコアモデルと大きく異なるのは，評判動学では，プレイヤの行動（ゲームの相手を助けるか否か）が善いか悪いかの判断に自由度がある点である。

プレイヤの行動が善いとみなされるか悪いとみなされるかは，１期前のゲームにおける，（相手の行動，相手の評判，自分の評判）の３要素に応じて決まる。各要素は，２つの値を持つ（行動集合｛助ける，助けない｝，評判集合｛善い，悪い｝）ので，プレイヤの評判は$2^3=8$通りの入力に対して善いか悪いかを判定する。従って，各人の評判の決まり方（評判ルール）は，$2^8=256$通り存在する。各人は，自分と相手の評判の組み合わせ$2^2=4$通りに応じて，相手を助けるか否かを決める。従って，自分の評判と相手の評判に対するプレイヤの行動方法（行動ルール）は$2^4=16$通り存在する。各人の評判ルールと行動ルールの組み合わせの数は$2^8\times 2^4=4096$となる。

評判動学において，ゲームをプレイする各人の評判は，第三者によって決定され，他のプレイヤ全体に共有される。ゲームをプレイしている２人は，自分が相手に対して持つ評判から自分の行動ルールに従って，相手を助けるか否かを決定する。その結果をゲームに参加しない第三者が観察し，その２人の評判を，自分の評判ルールにもとづいて，善いか悪いかを決める。２人の評判に対するこの第三者の判定は，他の全てのプレイヤに正確に伝達される。従って，全てのプレイヤのゲームをプレイしている２人に対する評判は，第三者の評判ルールに応じて唯一つに定まる[7]。

Ohtsuki and Iwasa（2004；2006）は，ESSの概念を拡張して，進化的に安定な評判ルールと行動ルールのセットをESSペアと定義した。ESSペアとなる評判ルールと行動ルールのペアは，それぞれ，現在の行動ルールの下で評判ルールが進化的安定となると同時に，現在の評判ルールの下で行動ルールが進化的に安定になる。評判動学には，利得表の構造に応じて多数のESSペアが存在する。更に，非常に広い利得表領域で定常的に観察される。特に適合度の高い（全てのプレイヤが相互に助けあう）ESSペアが存在する。８種類からなるESSペアは，leading eightと名づけられた。

以下では，このleading eightが承認理論における一般規範に対応し，一般規範が複数の個別規範を包摂し，財所有を正当化する一般規範の圏域について論じる。

1　評判と規範

評判動学の結果を承認理論に適用するために，まず，評判動学における評判を規範とみなすことができることを説明する。前章と同様に，評判動学における相手を助ける行動が実定

6）Ohtsuki and Iwasa（2004）では，各プレイヤはただ一人の対戦相手と１回のゲームしか行わず，そのゲームの結果で適合度が決まる。Ohtsuki and Iwasa（2006）は，１世代の中で，各プレイヤは同じ対戦相手と１回しか対戦しないが，対戦相手を組み替え，複数回のゲームを行い，合計利得で適合度が決まる。しかし，この２つの設定の差は結果に大きな違いをもたらさない。

7）この評判の伝達方法は間接観察方法と呼ばれる。Leimar and Hammerstein（2001）［８］は，ゲームに参加しない第三者全てが，ゲームをプレイする２人の行動を観察し，各々の評判ルールに応じて２人の評判を決める直接観察方法を採用した。直接観察方法では間接観察方法と異なり，全てのプレイヤのあるプレイヤに対する評判は必ずしも一致するわけではない。

的領域における同調的承認に対応し，相手を助けない行動が自律的承認に対応すると解釈して議論を進める。承認理論の規範は以下の３つの性質を持つ概念である。

１．規範は実定的領域における財の所有を正当化する。
２．規範は実定的領域から，財の所有の正当化の要請を受ける。
３．規範どうしはそれぞれが正当化する財の所有が矛盾する場合，いずれの規範が正当であるかの正義を争う衝突が起きる。

評判動学の評判は１．２．３．を満たすことを説明する。１．は評判の定義から直ちに条件を満たすことが分かる。評判は，ゲームでの各人の行動を，そのゲームの参加者２人の評判とそのゲームにおける行動から，善か悪かを判定する。これは，承認理論の実定的領域での二つの承認方法の正当化の与え手となる規範の機能に対応する。

２．は，ある評判ルール（個別規範）が行動ルール（実定的領域での財の承認方法）から決まることを意味する。評判動学では，各人は評判ルールと行動ルールを同時に持つ。この個人が社会に存在し続けるためには，自分の行動ルールを善いものとみなす，すなわち正当化する評判ルールが存在しなければならない。実際に，評判動学において，ある行動ルールが進化的安定になるためには，その行動ルールとペアとなる評判ルールが同時に進化的安定になっていなければならない。このことは，実定的領域における承認方法が規範による正当化を要請する仕組みと同様の仕組みが評判動学に含まれていることを意味する。

３．は，評判動学において，ある行動ルールが ESS である場合，それに対応する複数の評判ルールが進化的安定となるか否かの競争が生じることを意味する。善いものは助け，悪いものは助けない行動ルールと ESS ペアとなる評判ルールは複数あるが，それらは全体の一部であり，評判ルール間での競争が生じることを示そう。例えば，イメージスコアはこの行動ルールとの組み合わせで leading eight に含まれないが，Sugden（1986）［11］の Standing 規範は悪者に対して助けない行動を正当化するので，leading eight に含まれる。承認理論の言葉で言い換えれば，正当化されない財所有に対して自律的承認を行い，正当化される財所有には同調的承認を行う実定的領域での財の所有構造に対して，イメージスコア規範と Standing 規範の間で規範の衝突が生じ，Standing 規範がその衝突に勝利し，正義となる構造が評判動学に含まれている。

１．２．は規範的領域と実定的領域の間の垂直循環を規定する条件で，３．が規範的領域内での水平循環を規定する条件である。評判動学における各人は，評判ルールと行動ルールをセットで持っているので，評判を規範とみなす場合，垂直循環の運動を通じてしか規範間の動学を説明できないかのように思われる。しかし，実際にはそうではない。なぜならば，行動ルールと評判ルール（規範）が用いられる場合のプレイヤの役割が異なるためである。行動ルールは，そのルールを持つ個人が実際にゲームをプレイしている場合に採用される戦略である。一方，評判ルールは，自分がゲームをプレイしていない場合に，自分と関係のない第三者である２人のプレイヤの行動を評価するときに用いられる。従って，各人の行動ルールは相手の評判にもとづいて行動を決めるが，評判ルールが行動ルールの中に含まれるわけではない[8)]。あくまで，評判ルールと行動ルールは異なる領域に属する要素である。同時に，評判ルールと行動ルールは，それらをセットで持つ個人の適合度に応じて進化する。各人の適合度は自分の行動ルールと第三者の評判ルールに応じて決まる。

このように評判ルールと行動ルールは，それぞれ各人の適合度を決定する異なる役割を果たす2つの領域で定義される。このモデルの特徴が，評判動学を承認理論を表現するモデルに相応しいとみなす根拠である。評判動学における評判ルールと行動ルールの動学は相互依存するが，それぞれの要素が定義される領域はプレイヤに用いられる場合の違いという質的差を持つ。承認理論における実定的領域と規範的領域が，各人の財所有という1つの事象に対する感情的承認を規定する領域と規範による正当化を規定する領域であるのと同様に，評判動学における行動ルールと評判ルールは，プレイヤの適合度を直接的，間接的に規定する2つの領域である。

以上の議論により，以下では，評判ルールを個別規範とし，行動ルールを財所有の承認方法として承認理論の動学を評判動学の結果を用いて論じる。

2　一般規範と個別規範のズレ

Ohtsuki and Iwasa（2004；2006）は，全ての個別規範と承認方法の中から，常に非常に高い適合度を達成する8つの評判ルールと行動ルールの組み合わせを見つけ，leading eight と呼称した。評判を G が善い，B が悪いとし，行動を C が助ける，D が助けないと表す。以下，評判ルールを d（自分の評判，相手の評判，自分の行動）で表し，行動ルールを p（自分の評判，相手の評判）として表す。leading eight は，以下の表2に示される。

leading eight は全ての行動ルールと評判ルールの中で最も高い適合度を達成する。leading eight が承認理論における一般規範とそれに対応する財の承認方法として，多数の個別規範と財の承認方法の中から選抜されることを示そう。

まず，評判動学には4096もの個別規範と，各規範に対応する財の承認方法が存在する。この規範と承認方法の組み合わせの中でESSペアとなり，淘汰を生き延びる個別規範と承認方法

表2

評判ルール d:	自分の評判，相手の評判				
		G, G	G, B	B, G	B, B
自分の行動	C	G	*	G	*
	D	B	G	B	*

行動ルール p:		C	D	C	**

Ohtsuki and Iwasa（2006）より，leading eightの評判ルールと行動ルールの組み合わせを示す表。* は G, B のどちらでもよい。** は $d(B, B, C)=G$ と $d(B, B, D)=B$ のときのみ C を表し，他の組み合わせでは D を表す。* と ** への入力に応じてこの表は d, p の8種の組み合わせを示しており，それぞれがleading eightを表す評判ルールと行動ルールのペアである。

8）進化ゲームの戦略として，相手の行動以外の入力を持つ戦略を無限に想定することができる。例えば，パブロフ戦略などは相手の戦略以外に自分の戦略も入力に含む。しかし，評判動学における評判ルールは，戦略である行動ルールの入力として捉えることはできない。行動ルールは評判にもとづいて自分の行動を決めるが，その評判はゲームに参加しない第三者の評判ルールにもとづいて決まる。従って，各人は行動ルールと評判ルールをセットで持つが，自分の評判ルールを自分の行動ルールの入力とすることはできない。

は，自律的承認が全てを占めるケースを除いて25ペア存在する。しかし，これら25種の規範と承認方法のペアが全て一般規範とそれに対応する承認方法に分類できるわけではない。なぜならば，個別規範が一般規範となるためには第Ⅰ章の(2)式を満たす持続可能な規範でなければならないためである。この(2)式を満たす規範は，leading eightのみである。

実際，leading eightのみからなる社会では常に各人は互いに同調的承認を行い，その行動が規範によって常に正当化される。小さな確率で，承認行動や規範を誤ったとしても，すぐに訂正が行われ，「社会状態」が持続する。更に，その社会状態に他の規範と承認方法を持つ個人が侵入しても，その個人が社会に持続的に存在することはできない。この意味で，leading eightからなる社会は，一般規範が全ての個人の承認方法を正当化し，互いに相手の承認方法を尊重することによって類型的利害が最大化される社会である。leading eightが「社会状態」を支える仕組みと，そこで成立する一般規範の圏域について，次の節で述べる。

3　一般規範としてのleading eightとその圏域

leading eightに共通する特徴は以下の4つである。

1．協力関係の維持
この性質は，行動ルール$p(G, G)=C$と評判ルール$d(G, G, C)=G$に表現される。各人は，自分とゲームの対戦相手がよい評判を持っている場合には協力行動をとり，また，その行動は善いとみなされる。このことは，相互協力関係が，行動ルールと評判ルールで共に維持されることを示す。

2．逸脱者の特定
この性質は，評判ルール$d(G, G, D)=B$，$d(B, G, D)=B$に表現される。よい評判を持つものに対して非協力的行動をとったものは直ちに悪とみなされる。

3．罰則と罰則の正当化
この性質は，行動ルール$p(G, B)=D$と評判ルール$d(G, B, D)=G$に表現される。自分が善いものであるとき，悪い相手に対しては非協力という制裁を加え，なおかつこの制裁によって自分の評判は損なわれないことを意味する。

4．謝罪と許容
謝罪は，行動ルール$p(B, G)=C$で表現される。間違って，善いものに対して非協力的態度をとったものは，善い相手に対して協力を行うことによって，過去の自分の行動を謝罪できる。許容は，評判ルール$d(B. G, C)=G$で表される。これは，謝罪によってプレイヤの評判が回復することを意味する。

leading eightの特徴から，承認理論における一般規範の圏域と，対応する財の承認方法について考察を行う。まず第1に，互いに同調的承認を行うことは一般規範によって正当化され，かつ，同調的相互承認は一般規範を強化する。

第2に，自分の承認方法が正当化されているか否かに関わらず，一般規範によって財の承認が正当化されている相手に対する自律的承認は正当化されない。相手の財所有が正当化されているならば，常に相手の規範に従わなければならない。

第3に，相手の財の所有が正当でないならば，自律的承認が正当化される。正当でない財の所有をする相手の規範に同調することは許されない。

最後に，正当でない財の所有をしている各人は，正当な財の所有をしている相手に対して同調的承認を行うことによって，財所有を正当化できる。正当な財所有をしている個人の規範に従うことによって，自分の財所有を正当化でき

る。

この性質を備える規範は，その名のとおり８種存在する。この８種の規範と承認方法のペアは，どのような組み合わせでも全ての個人の同調的承認を保障する「社会状態」を形成できる。いま，leading eight ８種全てが社会に存在する状態を考える。leading eight に含まれるある実定的領域の承認方法を所与とした場合，その承認方法の正当化の与え手となる一般規範が複数存在する。

例えば，（自分の評判，相手の評判）の組み合わせが，（正当，正当）（正当，非正当）（非正当，正当）の場合，それぞれ（同調的承認）（自律的承認）（同調的承認）という承認行動は，８種全ての規範で正当化される。すなわち，これらの承認行動の正当化の与え手は複数存在し，各人が８種のうち異なる規範を用いていても矛盾は生じない。leading eight は，その規範と承認方法のペアの間で適合度に差がないため，このペアの間の進化は中立的であるが，自律的承認とペアになる規範の侵入によりその割合は変化する[9)]。leading eight は，８種それぞれが一般規範であり，どの一般規範が実際に各人に採用され，正当化の圏域がどの範囲になるかは決定論的には決まらない。

一方で，自分と相手の財所有が正当か否かの組み合わせが，（非正当，非正当）の場合に同調的承認を正当化する一般規範は，leading eight の内，２種のみである。この承認行動は，leading eight の内，６種の一般規範が正当化できない承認行動であり，６種の一般規範がその正当化の圏域に局所性（locality）を持つことを示している。同様に，自分と相手の財所有が正当か否かの組み合わせが，（非正当，非正当）の場合に自律的承認を正当化する一般規範は，６種であり，残りの２種の一般規範もまた正当化の圏域に局所性を持つ。

第１章(2)式は，ナッシュ均衡や ESS よりも厳しい条件であるにもかかわらず，この条件を満たす一般規範は８個もの候補を持つ。一般規範は，ある範囲内で各人の解釈に自由度を持つ柔軟な規範として成立しうることが分かった。逆に，一部の個人の財所有にしか正当化を与えることのできない規範は一般規範に淘汰される。従って，評判動学による承認理論の解釈における一般規範は全ての個人の財所有を正当化する圏域を持つと同時に，８パターンの解釈の自由度を持つ柔軟な規範として表現できる。ただし，共に財所有が正当化されていない個人間の相互承認の正当化については，一般規範の圏域に局所性が存在する。

Ⅴ　おわりに

本論文は，承認理論をイメージスコアモデルと評判動学によって解釈した。イメージスコアモデルは，承認理論における個別規範が１種類に限定されており，各人は全てこの規範にもとづいて財所有の正当化を行う。すなわち，イメージスコアモデルは承認理論において規範的領域で一般規範が成立していることを前提として，実定的領域における財の承認行動と一般規範の間の動学を表現するモデルである。イメージスコアモデルは，承認理論における規範的領域の水平運動を制約したモデルである。それに対して評判動学は，多数の個別規範が存在する場合を想定し，実定的領域における承認行動を通じて一般規範が生じるプロセスを描写する。評判動学は，実定的領域内と規範的領域内で，それぞれの要素が水平運動をすると同時に，２つの領域間でも垂直運動をする承認理論の一般的性質を備えたモデルである。

9）leading eight のどのペアの組み合わせも ESS であるため，このようなペアの侵入は一時的なものに過ぎないが，その後の leading eight 内の規範の割合は変化する。

イメージスコアモデルによる承認理論の解釈で，一般規範の圏域が実定的領域における類型的利害判断と個別的利害判断の差に依存して決まることが明らかになった。

評判動学による承認理論の解釈によって，多数の個別規範の中から「社会状態」の正当化の与え手となる一般規範が満たすべき性質が明らかになった。更に，この性質を満たす一般規範は８種類あり，「社会状態」にある各人の間で採用する一般規範が異なる場合があることが明らかになった。同時に，８種の一般規範の性質から，相互に財所有が正当化されない状態にある個人間の相互承認のあり方について，正当化の圏域が局所性を持つことが明らかになった。

これらの結果は，実定的領域における個人間の相互承認関係を囚人のジレンマで表現したことによって得られる。承認理論における「社会状態」は，規範的領域で正当化される各人の財所有が実定的領域における類型的利害判断によって支えられていることを条件とする。本論文では，この「社会状態」の実定的領域における成立条件を，各人の財所有が同調的承認によってのみ成立する条件に読み替えた。この条件は，ナッシュ均衡や ESS よりも厳しい条件であるにもかかわらず，イメージスコアモデルと評判動学において達成可能である。ただし，承認理論の想定する実定的領域での相互承認のあり方を制約している。この制約がどの程度厳しいものであるかを明らかにし，承認理論の想定する実定的領域と規範的領域の動学を完全に表現するための拡張方法を示して結びとする。

まず第一に，各人が同じ承認方法を採用する場合，自分の財の所有は保障されるものとした。各人が同調的承認を行う場合，それぞれの財の所有が保障されることは妥当であると思われるが，自律的承認を行うものどうしの場合，財の所有が承認されるか否かは規範的領域で決定されるとしたほうが承認理論を正確に表現できるだろう。第二に，同調的承認を行うものと自律的承認を行うものが相互承認を行う場合，自律的承認を行うものが全ての財を所有するとした。この仮定は，承認理論において，自律的承認主体が同調的主体と相対するとき，拡張的エゴイストになるという議論からなる。修正する必要があるのは，自律的承認主体と相対する同調的主体が，自分の財を相手が所有することに共感を感じるということである。主体が相手の財所有に対して共感を覚えるのは，互いに自分の財の所有が承認される同調的主体どうしの承認に限定するほうが自然であろう。修正後の実定的領域における利得表は表３のようになる。

このとき，実定的領域と規範的領域は２つの垂直運動を持つ。１つは，本論文のモデル解釈にもある実定的領域の承認方法を規範的領域が正当化するという運動である。もう１つは，実定的領域において自律的相互承認が起きる場合の財所有のあり方を規範的領域が直接規定する運動である。後者の運動を追加することにより，規範的領域と実定的領域はより強く相互依存し，一般規範の圏域について更なる知見が得られると期待できる。なぜならば，評判動学による承認理論の解釈において，相互に財の所有が正当化されず，各人が自律的相互承認を行う場合のみ，一般規範の圏域の局所性が観察されたためである。自律的相互承認をする場合の財所有の効用が規範的領域において決定されるとした場合，一般規範の圏域の局所性が規範的領域内の水平運動で説明される可能性がある。

拡張されたモデルは，実定的領域を表現するゲームのルールを，規範的領域のゲームの結果が決めるという構造を持つ。このモデルにおける規範的領域は，規範間の同調的承認のみを仮定した評判動学の拡張でもある。このとき規範間の均衡条件はより厳しいものになる。特に，規範的領域で自律的承認を行うものどうしの間の規範の運動は，直ちに実定的領域における相互承認のルール変化を生む。規範の変化が直ちに実定的領域に影響し，実定的領域でのルール

表3

	同調的タイプ	自律的タイプ
同調的タイプ	$S_i(I_i(D_{ip}))+S_i(I_i(D_{jq}))$	0
自律的タイプ	$S_i(I_i(D_{ip}))+S_i(I_i(D_{iq}))$	X

Xは，規範によって正当化される場合に$S_i(I_i(D_{ip}))$，正当化されない場合0。

変化は，各人の適合度を変化させる。このとき，規範的領域と実定的領域の間の垂直運動は，本論文で考察した2つのモデルよりも各領域の変化に対して敏感なものになる。実定的領域と規範的領域の相互作用を簡略化しても，2つの領域の要素の動学は非常に複雑な軌跡を描く可能性があり，第Ⅰ章(2)式の成立は厳しくなると予想される。モデルの拡張により，承認理論を一般規範の圏域についてだけでなく，ある程度の圏域を持つ個別規範が常に入れ替わるルール変化の激しい社会の分析に適用できる可能性がある。

2つの領域のゲームからなる大きなゲームのルールはその均衡によって決まる。均衡における規範と承認方法のペアは，青木（2001）［1］のいうゲームのルールと均衡の両方の性質を持つ制度の機能的条件を満たす。すなわち，一般規範は規範的領域と実定的領域の一般的認知均衡を満たす制度として描写できる。この観点から，拡張されたモデルは制度分析への適用が期待できる。

参考文献

［1］青木昌彦著，滝沢弘和・谷口和弘訳『比較制度分析に向けて』NTT出版，2001。

［2］八木紀一郎『講座　経済体制論　1』東洋経済新報社，1977。

［3］八木紀一郎「制度経済学と市民社会論：社会状態と市場状態」『進化経済学論集』第2集（進化経済学会第2回大会報告集）1998年3月，213–221ページ。

［4］八木紀一郎代表「平成8―10年度科学研究費補助金（基盤研究B1）研究成果報告書　制度の政治経済学の体系化」，1999年3月，35–45ページ。

［5］Alexander Richard D., *The biology of Moral Systems : Foundations of human behavior Evolutionary Foundations of Human Behavior Series* Aldine Transaction, 1987

［6］Ohtsuki Hisashi and Iwasa Yoh, “How Should We Define Goodness ? ― Reputation Dynamics in Indirect Reciprocity,” *Journal of Theoretical Biology* Volume 231, 2004, pp. 107–120.

［7］Ohtsuki Hisashi and Iwasa Yoh, “The Leading Eight : Social Norms That Can Maintain Cooperation by Indirect Reciprocity” *Journal of Theoretical Biology* Volume 239, 2006, pp. 435–444

［8］Leimar O. and Hammerstein P., “Evolution of Cooperation through Indirect Reciprocity,” *Proceedings of Royal Society London* Volume 268, 2001, pp. 745–753.

［9］Nowak Martin A., Sigmund Karl, “Evolution of Indirect Reciprocity by Image Scoring,” *Nature*, Volume 393, Issue 6685, 1998, pp. 573–577.

［10］Nowak Martin A. and Sigmund Karl, “The Dynamics of Indirect Reciprocity,” *Journal of Theoretical Biology*, Volume 194, Issue 4, 1998, pp. 561–574.

［11］Sugden R., *The Economics of Rights, Cooperation and Welfare*, Oxford, Blackwell. 1986.

［12］Trivers R., *Social evolution*, CA : Benjamin/Cummins Publishing Co., Menlo Park. 1985.

［13］Yagi K. “Trust and Sympathy in the Social and Market Order” in *Competition, Trust, and Cooperation*, ed. by Yuichi Shionoya and Kiichiro Yagi, Berlin, Heidelberg, Springer, 2001, pp. 20–41.