『道徳感情論』における社会秩序と腐敗
－ゲーム理論による定式化－*

An Evolutionary Model of *the Theory of Moral Sentiments*
－the Order and the Corruption－

菅　　隆彦
Takahiko Kan

1．序論

本論文は『道徳感情論』[1]における社会秩序と「道徳感情の腐敗」をゲーム理論によって定式化する。同書においてアダム・スミスは社会秩序の形成について論じた。スミスの論じた社会秩序は人々の相互の同感に基づく。同書は各時代の文脈に応じて様々な観点から解釈され，その現代的意義が見出されてきた。その中の一つの潮流として，主に1990年代以降における他分野の観点からの同書の解釈がある。各分野の研究者が同書から知見を得て現代的問題の解決に生かそうと試みてきた。例えば同書が行動経済学の研究成果を予見していることを示すもの（Ashraf et al., 2005），脳科学の観点から解釈するもの（Kiesling, 2012），経験的な道徳的判断手法の観点から解釈するもの（Konow, 2012），合理的選択理論の観点から解釈するもの（Khalil, 2017），等々が存在する。

特にゲーム理論の観点から同書を解釈した先行研究として，大きく二つが存在する。一つがMeardon&Ortmann（1996）であり，もう一つがSmith&Wilson（2017），Smith（2018），Smith&Wilson（2018）らの一連の研究である。前者は同書中の「自己規制」を繰り返しゲームモデルによって定式化した。後者の一連の研究は同書中の記述をゲームモデルによって定式化し，それが実証実験の結果と整合的であることを明らかにした。しかしこれらの先行研究においては同書の秩序形成論の根幹をなす「良俗の一般的諸規則」の形成過程が未だにモデル化されていない。このため同書の知見を現代的秩序問題の解決に十分に生かすことができていない。本論文は新たにゲームモデルによって秩序形成のメカニズムを解明することで，その意義と限界を明らかにする。また先行研究においては「道徳感情の腐敗」という重要問題が未だにモデル化されていない。道徳感情の腐敗はスミス自身が言及した社会秩序の破綻例であり，重要視されている。腐敗による社会秩序破綻のメカニズムを解明することで現代的秩序問題の解決に寄与する潜在性があるのにも関わらず，その解明が行われていない。秩序破綻の原因が明らかとなれば，破綻を予防することが容易になる。

本論文は以下のように構成される。第2節では一般的諸規則形成と道徳感情の腐敗について説明する。第3節では一般的諸規則の形成過程を定式化する。第4節では道徳感情の腐敗を定式化する。第5節では前節までの議論を総括しさらなる研究の余地に言及する。

2.『道徳感情論』と道徳感情の腐敗

2．1　『道徳感情論』と良俗の一般的諸規則

『道徳感情論』においてスミスは一つの社会秩序の形成について論じた[2]。同書における基礎的な概念が「同感」（sympathy）である。同感は，何らかの事柄に際して他者が抱く感情に，主体が接する際に生じる。同感とは，主体が想像の中で当事者の境遇に身を置き，当事者と同じ感情を共有することである。もし自分の感情が他者に同感されたとすれば，それは自分の感情が他者に一定程度正しいと認められたことを意味する。この意味での同感を，スミスは「是認」（approbation）と呼ぶ。社会秩序が形成されるためには，感情のみならず言動が重要になるわけだが，言動に対してもまた是認は行われる。ある言動を是認できるかは，その言動に伴う感情（意図）を是認できるかによって決まる。例えばある他者が暴力をふるっていたとすれば，その暴力が是認できるかは暴力に伴う感情（意図）に同感できるかによって決まる。正当防衛の意図の暴力であれば是認されやすいであろうし，悪意に満ちた暴力であれば是認されづらいであろう。他人に是認されるような言動のみを人々がとるとすれば，何らかの社会秩序が形成されるであろう。しかしスミスは人々がただ他者に是認される言動をとるだけでは，十分な社会秩序が形成されないとした。他者といっても様々な価値観の人間が存在するのであり，言動がある一人に是認されたとしても，その言動は社会秩序を破壊してしまうような言動かもしれない。十分な社会秩序が形成されるには，「中立的な観察者」（impartial spectator）という想像上の理想的な観察者に是認されるように，人々が言動をとる必要がある。そのために不可欠なのが，「良俗の一般的諸規則」（general rules of morality，以下，一般的諸規則）である（TMS：III.4.全体）。

一般的諸規則は集団の外部から与えられる規則ではなく，人々の相互作用によって形づくられる。一般的諸規則の形成のきっかけは他者の観察である（TMS：III.4.7）。他者の行動を繰り返し観察する中で，主体は中立的な観察者に是認されない行動を見て衝撃を受ける[3]。主体はこのような行動を見苦しいと感じる。やがて主体はその行動に対して周りの皆が自身と同様の嫌悪感を抱いているのを知る。他者が自身と嫌悪感を共有することを知った主体は，自身の感情が正当だという思いを強くする。この経験が繰り返されることで一般的諸規則は形成される。一般的諸規則は「継続的な観察」（TMS：III.4.7）によって形成されるのであり，主体の「経験にもとづいている」（TMS：III.4.8）。主体が自身の感情の正当性を確信するに至ると，その次の段階として，自身がその行動をとった際に他者が自分をどのように見るのかを想像する。他者は自分を見苦しいと当然みなすだろうと主体は想像し，それを避けるためにその行動はとるまいと決意する。かくして主体はある行動を取ってはならないとする一つの一般的諸規則を形成する。他の主体も同様の過程を経るのであり，同様の一般的諸規則を形成する。

2．2　道徳感情の腐敗

道徳感情の腐敗論は，『道徳感情論』第 6 版において追加された議論であり，同書を理解するにおいて無視することができない重要性を持つ。田中（2000, p.123）によればスミスが第 6 版を出版した最大の動機は道徳感情の腐敗論である。道徳感情の腐敗はスミスの思想の中で解決するのが最も困難な問題の一つみなされることもある（田島, 2003, p.256）。

徳の道（the road to virtue）と財産の道（that [the road] to fortune）という，二つの異なる倫理基準を，スミスは区別する。徳の道に進むことは，「英知の研究と徳の実行」によって尊敬を得ようとすることを意味し，財産の道に進むことは「富と地位の獲得」によって尊敬を得ようとすることを意味する（TMS：I.iii.3.2）。二つの道は全く異なり，財産の道は徳の道とは整合しない悪徳と愚行をしばしば伴う。例えば富の獲得に執着する人間は非道徳的な行動を躊躇することが少なくない。しかし二つの道は後に述べる理由により混同されてしまうことが多い（TMS：I.iii.3.3）。

中位・下位の生活階級と，上位の生活階級の，二つの分離した集団を，スミスは想定する。中位・下位階級においては二つの道はほとんど一致するが，上位階級においては両者は全く異なったものになる（TMS：I.iii.3.5- 6 ）。中位・下位階級においては人々に尊敬を得られるような努力をすることが財産の獲得に繋がる。しかし，上位階級においては，「成功と昇進は，理解力があり豊富な知識をもった同等者たちの評価にではなく，無知高慢で誇りの高い上位者たちの，気まぐれでばかげた好意に依存する」（TMS：I.iii.3.6）。その好意は，例えば財力がありそうだというような，外面的な品位や身だしなみから得られるのであって，徳のある言動から得られるのではない。

上位階級においては，徳の道が軽んじられ財産の道が重んじられる。この問題は，中位・下位階級が上位階級に憧れることによって，全階級に波及してしまう（TMS：I.iii.3.7）。スミスによれば，「富裕な人びと，有力な人びとに感嘆し，ほとんど崇拝し，そして，貧乏でいやしい状態にある人びとを，軽蔑し，すくなくとも無視するという，この性向は，……道徳感情の腐敗の，大きな，そしてもっとも普遍的な，原因である」（TMS：I.iii.3.1）。富裕層が感嘆や崇拝されやすく貧困層が軽蔑や無視されやすいということはありふれた現象であるが，これが徳の道の軽視に繋がってしまう。人々は上位階級に憧れを持ち彼らの言動を模倣する。

3．一般的諸規則形成のモデル

前節で述べた一般的諸規則の形成過程は主体の試行錯誤学習によって進行するとみなし，一般的諸規則形成過程を定式化する。主体はまず他者の見苦しい行動を観察し嫌悪感を持つ。その上で他者が自身と同様の感情を抱くことを知ると，自身の感情が正当だという思いを強くする。これが継続的に繰り返されれば，一般的諸規則が形成される。一方で，他者が自身と同じ感情を抱かないことが観察された場合には，感情の正当性が確認されることはない。主体は他者と感情を共有するか否かを判断基準として自身の感情の正当性を試行錯誤的に学習し，一般的諸規則を形成する。

本節では菅（2020）に修正を加える形でモデルを再構築する。厳しい紙幅の都合上モデルの要点の記述に留める。モデルの詳細な説明は菅（2020）や今後の筆者の著作を参照されたい。

試行錯誤学習による感情の正当性の上昇過程を，各主体がその感情を持つべきとみなす確率の上昇過程とみなす。本モデルにおけるプレイヤーは，二つの戦略の中から一つの戦略を確率的に選択する。プレイヤーの集合を$I=\{1,2...,n\}$とし，$n=2m(m\in \mathrm{N})$とする[4]。$S_i=\{A,B\}$をプレイヤー$i\in I$の純粋戦略の集合とする。プレイヤー$i\in I$が戦略Aを選ぶ確率を$x_{iA}\in[0,1]$，戦略Bを選ぶ確率を$x_{iB}\in[0,1]$とする。プレイヤー$i\in I$の混合戦略を$x_i=(x_{iA},x_{iB})\in[0,1]^2$とする。$x_{iA}+x_{iB}=1$が成り立つ。$\Delta_i=\{x_i\in[0,1]^2:x_{iA}+x_{iB}=1\}$とする。$\Theta=\times_{i\in I}\Delta_i$とする。$x=(x_1,x_2,...,x_n)\in\Theta$をプレイヤー集団の混合戦略プロファイルとする。プレイヤー$i\in I$は，$t\in \mathrm{N}$期に戦略Aを確率$x_{iA}(t)\in[0,1]$で選択し，戦略Bを確率$x_{iB}(t)\in[0,1]$で選択する。本モデルでは，プレイヤーがどちらの戦略をとるべきと考えているかが，プレイヤーの混合戦略において定量的に表されている。プレイヤーiは，t期に戦略Aを確率$x_{iA}(t)$で選択するべきと考えており，戦略Bを確率$x_{iB}(t)$で選択するべきと考えている。

本モデルにおいては，毎期，n人のプレイヤー集合の中で，全プレイヤーが二人一組ずつランダムマッチングされる。マッチングされた二人のプレイヤーは対戦において相手の選んだ戦略を知る。$s_i\in S_i$をプレイヤーiの純粋戦略とし，$S=\times_{i=1}^{n}S_i$とする。$s=(s_1,...,s_n)\in S$を純粋戦略プロファイルとする。プレイヤーiの利得は，任意の純粋戦略プロファイル$s\in S$について，純粋戦略利得関数$\pi_i:S\to \mathrm{R}$によって決まるとする。混合戦略プロファイル$x\in\Theta$がプレイされるとき，純粋戦略$s=(s_1,...,s_n)\in S$が採用される確率を，$x(s)=\Pi_{i=1}^{n}x_{is_i}\in[0,1]$とする。関数$u_i:\Theta\to \mathrm{R}$を次のように定義する。$u_i(x)=\Sigma_{s\in S}x(s)\pi_i(s)$。関数$u_i$は混合戦略$x\in\Theta$がプレイされるときの，プレイヤー$i$の期待利得である。

プレイヤーiが純粋戦略s_iをとることは極端な混合戦略をとることに等しい。このときの，混合戦略を$e_i^{s_i}\in\Delta_i$，xから第i成分を除いたものをx_{-i}とし，プレイヤーiの期待利得を$u_i(e_i^{s_i},x_{-i})$と表す。混合戦略xは，各期の各戦略の期待利得に応じて変化する。各プレイヤーiについて$x_{iA}+x_{iB}=1$であるから，xの挙動を知るには$x_A=(x_{1A},...,x_{nA})\in[0,1]^n$の挙動を知るだけで十分である。$x_A$を「社会状態」と呼ぶ。社会状態$x_A$の各成分である，$x_{iA}$のダイナミクスを，$\dot{x}_{iA}=[u_i(e_i^A,x_{-i})-u_i(x)]x_{iA}$，とする。$u_i(x)=x_{iA}\cdot u_i(e_i^A,x_{-i})+x_{iB}\cdot u_i(e_i^B,x_{-i})$であるから，$u_i(x)$は混合戦略$x$がプレイされる場合の，プレイヤー$i$の期待利得である。この期待利得よりも$u_i(e_i^A,x_{-i})$が大きければ，$x_{iA}$は増加し，逆に$x_{iB}$は減少する。

各プレイヤーiの利得は下の表の通りであり，プレイヤーごとに異なる。全てのiについて，$a^i\in \mathrm{R}^+,b^i\in \mathrm{R}^+,c^i\in \mathrm{R}^+,d^i\in \mathrm{R}^+$とする。

i ／ 相手	戦略A	戦略B
戦略A	a^i	b^i
戦略B	c^i	d^i

表1：プレイヤーiの利得

ここで，$a_i = a^i - c^i$, $\beta_i = b^i - d^i$ とする。各プレイヤーiの利得は，$\alpha_i = a^i - c^i > 0$, $\beta_i = b^i - d^i < 0$, を満たすとする。これは一般的諸規則の形成過程と整合する仮定である。各プレイヤーは，相手のとる戦略を所与として自分の戦略が一致した場合の方が，一致しない場合に比べて利得が大きくなる。相手と戦略が一致するということは，相手と感情を共有することを意味する。

本ダイナミクスの連立微分方程式に，利得を代入する。まず，$u_i(e_i^A, x_{-i}) = a^i(\Sigma_{j \neq i} x_{jA}/n-1) + b^i\left(1-(\Sigma_{j \neq i} x_{jA}/n-1)\right), u_i(e_i^B, x_{-i}) = c^i(\Sigma_{j \neq i} x_{jA}/n-1) + d^i(1-(\Sigma_{j \neq i} x_{jA}/n-1))$ である。これらを基に各 $\dot{x}_{iA}$ が導出される。

$$\dot{x}_{1A} = \left((\alpha_1 - \beta_1)\left(\Sigma_{j \neq 1} x_{jA} \Big/ n-1\right) + \beta_1\right) x_{1A}(1 - x_{1A}),$$

$$\vdots$$

$$\dot{x}_{iA} = \left((\alpha_i - \beta_i)\left(\Sigma_{j \neq i} x_{jA} \Big/ n-1\right) + \beta_i\right) x_{iA}(1 - x_{iA}),$$

$$\vdots$$

$$\dot{x}_{nA} = \left((\alpha_n - \beta_n)\left(\Sigma_{j \neq n} x_{jA} \Big/ n-1\right) + \beta_n\right) x_{nA}(1 - x_{nA})。$$

この連立微分方程式より，全ての端点（社会状態の各成分が0か1になる点）が平衡点であることがわかる。

二つの戦略について以下のように解釈する。ある文脈において戦略Aは中立的な観察者に是認されるような戦略であり，戦略Bは是認されないような戦略である。例えば，荷物を持つのが困難なお年寄りを見た際にお年寄りの荷物を持つ戦略と，お年寄りを馬鹿にする戦略の二つがあり，二者択一である。**一般的諸規則**が形成されている状況は，各プレイヤーが確実に戦略Aを選ぶ状況とする。つまり，「社会状態 x_A が，$x_A = (1,1,...,1)$ となる状況」とする。

一般的諸規則がどのように実現しうるかを明らかにする。ダイナミクスにおいて重要となるのが，平衡点と漸近安定点である。平衡点は，全てのiについて $\dot{x}_{iA} = 0$ となる点である。漸近安定点は微小な変化に対して安定な平衡点であり，モデル外の要因による変動に対して頑健である。平衡点が複数存在しても，漸近安定点が一つであればその点が収束先とみなされる。複数の漸近安定点が存在する場合には，ダイナミクスの収束先は初期値とモデル外の変動に依存して決まる。本ダイナミクスにおける平衡点は，すべての端点である。各 α_i，各 β_i の具体的な値に応じて漸近安定点は変化する。各端点が漸近安定になりうる。また内点平衡点が存在しうる。しかし，各iの α_i，β_i が

どんな値をとっても必ず言えることがある。漸近安定となる端点が二つある。

命題：社会状態（0,0,...,0）と社会状態（1,1,...,1）は，各 i の α_i，β_i がどんな値をとっても漸近安定である[5]。

この命題から，一般的諸規則だけではなく社会状態（0,0,...,0）も漸近安定であることがわかる。社会状態がどちらに収束するかは初期値に依存して決まる。初期値が一般的諸規則に近い場合にはその点に収束し，社会状態（0,0,...,0）に近い場合にはその点に収束する。前者の収束は一般的諸規則形成のシナリオとまさに整合する。同感的な利得を持つ主体達がマッチングを通して相互作用することにより一般的諸規則を作り上げて行く。一方で，後者の社会状態（0,0,...,0）への収束は，同感的な利得を持つ主体達が謂わば集団で堕落して行く過程と見ることができる。このように一般的諸規則の形成は実は初期値に依存する。一般的諸規則の形成が不確かであることは言及されていないが，これは，スミスが十分に高い初期値を想定していたからかもしれない。この想定が正しければ一般的諸規則が実現することは確かである。

4．道徳感情の腐敗のモデル

4．1　モデル

道徳感情の腐敗を定式化する。道徳感情の腐敗が模倣によって進むことに着目し，進化ゲームモデルの一つである模倣のモデルを採用する。これにあたり前節のモデルから大きな変更がある。一般的諸規則のモデルでは，プレイヤー $i \in I$ の混合戦略を $x_i = (x_{iA}, x_{iB}) \in [0,1]^2$ として各 x_i の変化に着目したが，本モデルでは各プレイヤーが混合戦略を持つのではなくどちらか一つの戦略を選ぶ。道徳感情の腐敗は一般的諸規則の形成が完了した後に生じる問題であり状況が異なる。また，『道徳感情論』中の記述からして，道徳感情の腐敗は段階的な確率変化として捉えられる現象ではなく，上位の階級への憧れにより主体が急進的に腐敗に陥る現象と捉えられる。あるプレイヤー集団の中で，戦略Aをとるプレイヤーの割合を x_A^I，戦略Bをとるプレイヤーの割合を x_B^I と記述し，これらの変化に着目する。添え字の「I」は模倣（imitation）の頭文字に由来する。

$X^I = \{x^I = (x_A^I, x_B^I) \in [0,1]^2 : x_A^I + x_B^I = 1\}$ とする。前節のモデルでは戦略Bを単に中立的な観察者に是認されないような戦略としたが，本モデルでは戦略Bを道徳感情の腐敗に対応した戦略とする。戦略Bをプレイヤーが採用することは，プレイヤーが腐敗していることを意味する。腐敗のモデルにおいて**一般的諸規則**が形成されている状況は，各プレイヤーが確実に戦略Aを選ぶ状況であり，x^I で表現するなら，$x^I = 1$ となる状況である。

本モデルでは，ランダムマッチングされた二人のプレイヤーが対戦する。各プレイヤーは対戦において相手の選んだ戦略を知る。このマッチングと対戦が連続的に幾度も繰り返され，継続的な観察が行われると想定する。マッチングされた二人の戦略の組み合わせが戦略 i と戦略 j のとき，そ

の組み合わせを $s_{ij}=(s_i,s_j)$ と表す。$s_i=s_j$ となる場合もある。戦略の組み合わせの集合を S_{ij} とする。純粋戦略 $s_{ij}\in S_{ij}$ の組み合わせで対戦が行われる集団中の割合を，$x^I(s_{ij})=x_i^I\cdot x_j^I\in[0,1]$ とする。$\pi_i(s_{ij})$ を，戦略 j のプレイヤーと対戦するときに得る，戦略 i のプレイヤーの利得とする。

関数 $u:X^I\to\mathrm{R}$ を以下のように定義する。$u(e^i,x^I)=\Sigma_{j\in S}x^I(s_{ij})\pi_i(s_{ij})$。$u(e^i,x^I)$ は，集団が x^I の状態にあるときに，純粋戦略iから得られる期待利得である。利得関数は全プレイヤーで同じなので添え字はない。集団からプレイヤーをランダムに抽出するときの期待利得は $u(x^I,x^I)$ であり，以下のように求められる。$u(x^I,x^I)=x_A^I\cdot u(e^A,x^I)+x_B^I\cdot u(e^B,x^I)$。プレイヤーは前節のモデルと同様の利得関数を持つが利得表が異なる。モデルの簡潔さのためにプレイヤーは全員共通の利得表を持つとする[6]。道徳感情の腐敗が起こっている状況では，起こっていない状況と比べて主体達の喜びに変化が生じる。前節のモデルでは $\alpha_i=a^i-c^i>0$ を仮定したが，腐敗が起こるとこれは仮定できないであろう。財産の道を進む主体は，相手が徳の道を進む主体だった場合に優越感を感じて，利得が大きくなると考える。財産の道を進む主体は上位階級に感嘆する傾向が強く，上位階級の悪徳愚行さえ模倣して彼らに類似することを誇る（TMS：I.iii.3.7)。

優越感の喜びを加味した利得を $g\in\mathrm{R}^+$ とし，徳の道を進む主体の同感の喜びを $e\in\mathrm{R}^+$ とする。$r=e-g<0$ となる。相手が戦略Bを取った際の利得を，$f\in\mathrm{R}^+$，$h\in\mathrm{R}^+$ とし，$\delta=h-f$ とする。δ の符号がどうなるかは場合分けを行う。前節のモデルと同様に同感の喜びが大きい（$\delta=h-f>0$）とみなすことも可能であるし，腐敗したプレイヤー同士がマッチングされ競争の不安から喜びが減るともみなしうる。喜びが同じ大きさになる可能性もある。

i ／ 相手	戦略A	戦略B
戦略A	e	f
戦略B	g	h

表2：プレイヤー i の利得（腐敗）

プレイヤーは自分の戦略を常に見直す[7]。見直しの後に戦略を変更する確率を「選択確率」と呼ぶ。選択確率はプレイヤーが現在どの戦略をとるかによる。戦略Aをとるプレイヤーの選択確率を $p_A^B(x^I)$ とする。下の添え字は現在の戦略であり，上の添え字は変更後の戦略である。$p_A^A(x^I)$ は戦略Aをとるプレイヤーが，戦略を戦略Aに変更する（つまり事実上変更しない）確率である。$p_B^A(x^I)$，$p_B^B(x^I)$ は，戦略Bをとるプレイヤーの選択確率である。各 $i\in\{A,B\}$ ついて，$p_i:X^I\to X^I$ は，X^I を含む開領域でリプシッツ連続な関数とする。選択確率は x^I に依存して変わる。戦略の利得が高いほどその戦略へ変更される選択確率は高くなる。徳の道を進む主体は財産の道を進む主体に憧れて模倣するが，本モデルでは利得の高さへの憧れによってそれを表現する。ただし戦略Bをとるプレイヤーも模倣を行う可能性がある。片方だけが模倣を行うと想定すると，模倣される戦略が一方的に割合を高めるのは自明である。選択確率は，プレイヤーが各戦略に対して割り当てるウェイトに依存する。i 戦略家が純粋戦略 j に注目するウェイトを

$\omega_i(u(e^J,x^I),x^I) = \lambda+\mu u(e^J,x^I)$ とする。ここで $\lambda \in \mathrm{R}$，$\mu \in \mathrm{R}^{++}$ で，すべて x^I のとすべての純粋戦略 i に対して $\lambda+\mu u(e^j,x^I) > 0$ である。このとき以下が成り立つ[8]。

$$\dot{x}_i^I = \frac{\mu}{\lambda+\mu n(x^I,x^I)}[u(e^i,x^I)-u(x^I,x^I)x_i^I。$$

ダイナミクスの挙動は利得の値に依存して変化する。$r = e-g < 0$ となることを仮定した。しかし，$\delta = h-f$ の符号については，①$\delta > 0$，②$\delta = 0$，③$\delta < 0$ と，場合分けする。①$\delta > 0$ の場合，端点以外において $\dot{x}_i^I < 0$ が常に成り立つ。端点 $x^I = 0$ は漸近安定な平衡点であり，端点 $x^I = 1$ は不安定な平衡点である。②$\delta = 0$ の場合，端点以外において $\dot{x}_i^I < 0$ が常に成り立つ。端点 $x^I = 0$ のみが漸近安定な平衡点である。③$\delta < 0$ の場合，端点と内点 $x^I = {}^{\delta}\!/_{r+\delta}$ が平衡点となる。x^I がこの内点平衡点より小さい場合には，$\dot{x}_i^I > 0$ が成り立つ。x^I がこの内点平衡点より大きい場合には，$\dot{x}_i^I < 0$ が成り立つ。この内点平衡点のみが漸近安定である。三つの場合の中で，特にどの場合が道徳感情の腐敗と整合するだろうか。道徳感情の腐敗においては戦略Bをとるプレイヤーが一定数存在しその状況が一定期間続いているであろう。スミスの説明からはどちらかの戦略にプレイヤーが偏っているとは解釈できないし，腐敗の問題が一時的であるとも解釈できない。この考えに立つと道徳感情の腐敗は漸近安定な内点平衡点上の状況（場合③）とみなされる。この点においては各戦略のプレイヤーが混在し，そのような状況が一定期間続く。残りの場合においては内点平衡点がそもそも存在せず，端点のみが漸近安定となってしまう。

4．2　モデルからわかること

道徳感情の腐敗は場合③のダイナミクスによって捉えられるが，この結果から何がわかるであろうか。まず道徳感情の腐敗が現状のままでは解決しえないことである。内点平衡点は漸近安定であり微小な変化を想定しても頑健である。この点が内点であることから完全に腐敗した状況に陥る恐れはないが，一般的諸規則による秩序とは整合しない結果である。部分的な腐敗から脱して一般的諸規則を形成するためには，どのような策が考えられるであろうか。利得表を変化させることが根本的な問題解決に繋がる。一般的諸規則が漸近安定な状況とすることができればやがて道徳感情の腐敗は解決するが，三つの場合いずれにおいてもそのような状況にはならない。$r = e-g < 0$ の仮定を覆すことで，初めてそのような状況となる。$r = e-g < 0$ の仮定は誤った優越感が反映されたものであるから，その誤りを正して道徳感情の腐敗を解決する必要がある[9]。

『道徳感情論』においては一般的諸規則による社会秩序と道徳感情の腐敗が併存しているため，道徳感情の腐敗のモデルと一般的諸規則のモデルは一体的に解釈される必要がある。腐敗のモデルにおいては部分的に腐敗した内点が収束先であることがわかったが，このような点は前節のモデルにおいても利得表によっては収束先になりうる。この意味では併存を整合的に理解でき，定式化のメリットが見出せる。ただし前節のモデルにおいては腐敗した点を含む，中立的な観察者に是認されない点が必ず漸近安定となり，腐敗のモデルにはない悲観的な結論がありうる。腐敗以前にそも

そも同感による秩序は不確実なのである。だからといって『道徳感情論』の社会秩序論が無意味であることにはならない。前節のモデルでは一般的諸規則への収束がありうるが，これは一つの理想的な経路の提示とみなされる。この種の提示は経済学で意義が認められている。ただし，二つのモデルを整合的に解釈するならば部分的に腐敗した内点が収束先となるのであり，理想は実現しえないし，完全に腐敗してしまうわけでもない。

二つのモデルはどのような現実社会における意義を持つだろうか。前節のモデルからは，同感の可能性と限界がわかる。理想的な経路がある一方で真逆の経路もあり，部分的な腐敗に至る経路が存在する可能性がある。同感を利用した秩序の形成には，他の補足的な取り組みが必要とされる。腐敗のモデルからはやはり模倣の可能性と限界がわかる。完全な秩序（一般的諸規則）の実現には向かないが，悲観的な結論を避けられるメリットがある。

5．結論

本論文は『道徳感情論』における社会秩序形成を試行錯誤の過程と捉えて，道徳感情の腐敗を模倣の過程と捉えて，ゲームモデルを用いてそれぞれ新たに定式化した。これらにより同書における秩序形成と腐敗の併存を整合的に理解しうることを示した。また同感と模倣それぞれによるメカニズムの意義と限界を明らかにした。本論文は，『道徳感情論』の記述に忠実なモデルとするため，各モデルにおいて異なる仮定を置いた。このために両モデルの一体的な理解がしづらくなっていることは否めず，その点を改善したモデルの構築が今後期待される。

参考文献

- Ashraf, N., Camerer, C. F., & Loewenstein, G. (2005). Adam Smith, behavioral economist. *Journal of Economic Perspectives, 19(3)*, 131-145.
- Khalil, E. L. (2017). Socialized view of man vs. rational choice theory： What does smith's sympathy have to say?. *Journal of Economic Behavior & Organization, 143*, 223-240.
- Konow, J. (2012). Adam Smith and the modern science of ethics. *Economics and Philosophy, 28(3)*, 333-362.
- Kiesling, L. L. (2012). Mirror neuron research and Adam Smith's concept of sympathy： Three points of correspondence. *The Review of Austrian Economics, 25(4)*, 299-313.
- Meardon, S. J., & Ortmann, A. (1996). Self-command in adam Smith's theory of moral sentiments a game-theoretic reinterpretation. *Rationality and Society, 8(1)*, 57-80.

■ Smith, A. (1982). *The theory of moral sentiments* (D.D. Raphael, & A.L. Macfie, Eds.). Liberty Fund. (Original work published 1790). 水田洋訳（2003）『道徳感情論』（上－下）岩波書店。

■ Smith, V. L. (2018). Adam Smith, scientist and evolutionist： modelling other-regarding behavior without social preferences. *Journal of Bioeconomics, 20(1)*, 7-21.

■ Smith, V. L., & Wilson, B. J. (2017). Sentiments, conduct, and trust in the laboratory. *Social Philosophy and Policy, 34(1)*, 25-55.

■ Smith, V. L., & Wilson, B. J. (2018). Equilibrium play in voluntary ultimatum games： Beneficence cannot be extorted. *Games and Economic Behavior, 109*, 452-464.

■ Weibull, J. W. (1997). *Evolutionary game theory*. MIT Press. 大和瀬監訳（1999）『進化ゲームの理論』オフィスカノウチ。

■ 菅隆彦（2020）「『道徳感情論』における良俗の一般的諸規則の進化ゲーム理論的再解釈」『TERG Discussion Paper』, *433*, 1-26。

■ 田島慶吾（2003）『アダム・スミスの制度主義経済学』ミネルヴァ書房。

■ 田中正司（2000）『アダム・スミスと現代』御茶の水書房。

---

* 本論文は，2020年度公益信託山田学術研究奨励基金から助成された研究の成果である。

[1] これ以降，『道徳感情論』から引用する際には，TMSと記したうえでグラスゴウ版のパラグラフ番号を付し引用元を示す。訳文は水田訳を利用する。

[2] 本論文の主目的は『道徳感情論』における社会秩序と道徳感情の腐敗の定式化である。厳しい紙幅の都合上，前者について詳細な解説を施すことは回避する。例えば菅（2020）を参照せよ。

[3] 中立的な観察者に是認される言動についても，一般的諸規則は形成される。後に説明する嫌悪感を好意に読み替えることで，そのような諸規則の形成過程を説明できる。

[4] 本論文の定式化はウェイブル（1999）を参考にしており，変数や集合について，同書の記法を採用する場合が多い。

[5] この命題は連立方程式のヤコビ行列を調べることで証明できる。

[6] 前節のモデルではプレイヤーごとに利得表が別であるとした。これは利得表が同一と仮定して得られる結果が，一般化した場合にも簡単に求められるからである。

[7] これは，ウェイブル（1999）では，式4.32の仮定に相当する。

[8] このダイナミクスは，レプリケーターダイナミクスと解軌道が同一であるが，全く異なる状況設定から導出されたものである。

[9] 誤った優越感を正すための取り組みとして，スミスによる『道徳感情論』第 6 版の改訂を位置付けられるかもしれない。新第 3 部で個人の倫理観に訴える主張をしているようにもとれる。