特　集「道徳判断の自動化をめぐる問題：規範の選択と協力の進化」

進化的機械倫理のあらまし

Evolutionary Machine Ethics Synopsis

Han The Anh　イギリス・ティーズサイド大学計算・メディア・アート学部
School of Computing, Media and the Arts, Teesside University, UK.
T.Han@tees.ac.uk, https://www.scm.tees.ac.uk/t.han/

Pereira Luís Moniz　ポルトガル・新リスボン大学情報学・計算機科学研究所
NOVA Laboratory for Computer Science and Informatics（NOVA LINCS）, Departamento de Informática, Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa, Portugal.
http://userweb.fct.unl.pt/~lmp/

岡田　勇（訳）　創価大学経営学部
Isamu Okada　Faculty of Business Administration, Soka University.
okada@soka.ac.jp

Keywords: machine ethics, intention recognition, commitment, apology, forgiveness.

1. はじめに

感情をもたない道徳の計算モデルは論理プログラミングの技術を用いて分析されており，個々のエージェントに組み込まれるモデルの精緻化が主要な関心となっている [Han 12d, Pereira 09, Pereira 15, Pereira 16a, Pereira 16b, Pereira 17, Saptawijaya 15a, Saptawijaya 15b, Saptawijaya 18]．そこでは，仮説的推論や整合性の制約，選好，論証や反事実的条件文，更新といった道徳判断を対象としているものの，単一エージェントにおける道徳認知のダイナミクスをモデル化することに力点が置かれるため，個人間の差異が考慮されていない [Prinz 16]．一方，集団的な文脈では，進化ゲーム理論の技術を用いた計算論的な道徳の創発が報告されている [Han 15c]．そこでは，多様な集団が想定され，意図認知，約束，復讐，弁明，寛容性や罪の意識といった認知能力の導入によって，それがない場合と比較して，協力の創発を強化することが示されている．

人間の道徳研究は，個人的な道徳かそれとも集団的な道徳かで分類されるものの，相互に関連する研究も多い．例えば，ある文脈では，個人の認知や熟慮や行動などが強調され，他の文脈では，集合的な道徳やそれが進化的にどのように創発するのかなどが強調されるなど，重複的な議論もなされている [Pereira 15, Pereira 16b]．個人的な認知能力と集団への展開とをどのように架橋するかには，多くの点を考慮しなければならない．例えば，意図を認知する能力，つまり，我々の意図を他者がどのように認識するかは，個人と集団の両方から考慮される．それ以外にも，契約を作成する能力，それを許諾したり拒否したりする能力，補完的手段を採用する能力，集団を監視したり，そのプロセスを外部に委託する能力，協力したり裏切ったりする能力，そして，謝罪したり，報復や寛容さを醸成する能力などを考慮する必要がある．

本稿では，我々がこれまで行ってきた集団における道徳に関する研究を紹介する．特に，複雑な進化ゲームにおける協力・非協力行動のモデルを紹介する．詳細は [Han 18] を参照されたい．

2. 進化ゲーム理論

ゲーム理論は1940年代に開発された．数学者フォン・ノイマン（1903 ～ 57）と経済学者オスカー・モルゲンシュテルン（1902 ～ 77）の「ゲーム理論と経済行動」[von Neumann 44] がこの分野における最初の著作となっている．彼らはそれを経済学の一分野と認識していたが，冷戦構造の理解に適用されるなど幅広い応用分野を有することになった．ここでは，複雑な状況を精錬された数学的手法で解析し，さらに計算機シミュレーションを用いることで複雑な連立方程式系の解が探索されている．

ゲーム理論が取り扱うテーマは複雑なものが多く，多様なニッチをもっている．基本的には，プレーヤ間の相互作用やそれぞれの行動戦略（これは文化的遺伝子と呼ばれることもある），各プレーヤの生存や勝利が要素になっている．また，行動戦略の進化，特にゲームのルールや環境などが多様な状況における，行動戦略の突然変

異なども扱われる．ゲームによっては不確実性を含むものも存在し，その場合，存在可能な戦略は確率によって入れ替わることもある．行動戦略が変化するときに，その変化をプレーヤの利得に基づいて決定するとした場合，それは進化ゲームと呼ばれる．これもまた抽象モデルがつくられ，数学的な方法で分析される．

ゲームの分類の一つとして，ゼロサムゲームか非ゼロサムゲームかというのがある．前者では，あるプレーヤが勝利すればほかのプレーヤは敗者となるといったように，全プレーヤの利得の合計は常にゼロとなるゲームである．一方，非ゼロサムゲームはそうではない．自然選択のメカニズムの中にはこの非ゼロサムになるものもある．そこでは全員が勝者になったり，全員が敗者になったりする．例えば，文化や市民社会の進化を非ゼロサムゲームで表現し，協力によって利得が生じれば，利他主義が成立することを示している．戦略が共存するときは，短期的な均衡を実現することもある．捕食者と被食者のゲームを考えよう．捕食者は被食者を完全に食べつくすことも望まなければ，被食者が無限に増殖することもない．なぜなら，無限の増殖は環境資源が消費し尽くされてしまうからである．

プレーヤが何度ゲームをするのかといった点も重要な要素になる．相手と一度きりしかゲームしない場合や，同じ相手と別の機会に再びゲームする状況など，もしくは，ゲームの相手として指定された相手を拒否することが可能かどうかでゲームの挙動は変わる．このことについて，具体例をあげてより詳細に検討してみよう．有名な「囚人のジレンマ」ゲームを，利他主義のパラドックスの例として始めてみる．このゲームでは，共犯である二人の囚人 A と B が自白するか黙秘するかを選択する状況に直面している．

上記の 2 × 2 の利得行列において，各行は A の行動として黙秘か自白のいずれかを，各列は B の行動として黙秘か自白のいずれかを示している．A の自白の行と B の自白の列とが交差する場所は，8 年の刑期と書かれており，これが両囚人の「利得」となる．もし A が自白し，B がそうではない場合，A だけ 2 年の刑期になる一方，B は 10 年となる．刑期をなるべく短くしたいと思う場合，両者とも自白を選択する誘因が存在し，黙秘し続けるのは実際のところ有利には働かない．なぜなら一方が自白という裏切り行為をし，もう一方が黙秘し続けた場合，10 年の刑期に服する必要があるので，自白して 8 年に縮めたいと思うからである．このように裏切りへの誘惑は魅力的であるが，これには内在的なリスクがある．ともに黙秘し続ければさらに短い 6 年の刑期となるからである．

囚人達はこの利得関係を知っているにもかかわらず，相手の行動を知ることはできない環境に置かれているとする．黙秘し続けことは有利であるにもかかわらず，彼らは相手が自白したかどうかを知ることはできない．どちらかが自白してしまえば，黙秘は 10 年の刑期となる．ここにジレンマが存在する．黙秘し続ければ 6 年で済むのに，相手が裏切るリスクのゆえ，それを避ける行動を両者が取り，結果的に最悪のシナリオ—両者ともに 8 年の刑期—が実現される．古典的なゲームでは，そのため自白するという合理的「解」を導出する．相手と対話するチャンスがないばかりに，またそのようなチャンスがあったとしても裏切るリスクを認識するがゆえに，たとえ黙秘し続けられ刑期が短縮されるにもかかわらず，一致協力した行動をとることができないのである．

もし，A と B がこのゲームを何度もプレイできるとすると，「解」に変化が生じる．複数回の対戦によって相互に信頼したり不信頼したりといった関係を構築することができる．一方が一度裏切ったとすると，もう一方が復讐として，もしくは単純に非寛容として将来の対戦において裏切ることとなるかもしれない．コンピュータシミュレーションを用いれば，どのような戦略が生き残っていくのかを視覚的に表現することができる．より現実的な状況にするために，ゲームに社会構造を導入させることもある．

戦略はどのように更新されるのだろうか．はじめに，最も勝利する戦略を選択し，それをまねる戦略学習方式があげられる．複製はコストがかかるため，なるべく成功する戦略を模倣したい．そのため，勝者は複製されやすくなり，敗者はそうはならない．ほかにも，利得の高い戦略が比例的に選択されやすくなる学習方式や，ランダムに模倣する学習方式などがある．これは模倣によって集団の中に存在する戦略が絶滅するのを防ぐ目的があるときなどに採用される．

進化的文脈での重要な問いは，もし，集団が協力的であったら，プレーヤ達は最終的に利益を得ることができるのかというものである．これは，自分は協力のコストを支払わずに多くの利得を得ようとするフリーライダをどのように防げるかという問いを含んでいる．集団的な種の進化においては，必ずこのようなご都合主義者と協力とのバランスをどのようにとるべきかが問題であり，進化心理学の重要なテーマとなっている [Pereira 12]．これは通常，数学モデルと解析的なシミュレーションの実行とによって分析される．このとき，長期の進化的安定に焦点を当てるため，戦略には突然変異が提供されるのが一般的である．

3. 協力ジレンマを解決する新たな視点：意図認知と約束

協力の進化研究で他の研究分野との統合を試みるものはそれほど多くない．[Nowak 06] や [Sigmund 10] のサーベイでは，いくつかの研究で協力行動を触媒として用いているものがある．たいていそれらは進化ダイナミクスやゲーム理論に基づいているものの，行動進化にお

ける意図認知の役割を考慮していないものが多い [Han 13c]．我々は，協力の進化における意図認知の役割を明示的に扱った研究を実施した [Han 11, Han 12a, Han 12b, Han 13e]．それによれば，意図認知は繰返し囚人のジレンマの文脈において，それまで知られていた主要な戦略（Win-stay-lose-shift 戦略や Tit-for-tat 戦略）よりも優れており，ノイズのある状況でも，また意図認知にはコストがかかると設定したとしても，有意に高いレベルの協力率を促進することがわかっている．これらの知見は，意図認知が行動の進化の複雑性にとって新たなアプローチになることを示している．

意図認知の研究はこの数十年，人工知能分野で重要視されてきた [Charniak 93, Han 13a, Han 13c, Han 13f, Sadri 11]．それは人間と計算機の相互作用をどう改善するかや，生活の援助，道徳の推論，チームワークといったさまざまな分野にわたる [Han 12d, Han 13b, Pereira 11]．意図性（Intentionality）もまた，道徳判断の形成に決定的に重要な役割を有する．いわゆる「二重結果論（ある行為の帰結を，意図された帰結と予見される帰結に区別し，行為者に責任があるのは，前者のみであるとする考え方：訳者注）」や三重結果論などがそれである [Hauser 06, Mikhail 07]．我々の解析とエージェントシミュレーションの分析結果は，他者の意図を認知する能力を有し，道徳的な決定の判断を行うエージェントを設計するにあたり重要な知見となり得る．主要な研究成果としては，意図認知を導入することで，社会における道徳エージェントは高いレベルの協力行動を維持できることがわかっている．

今の社会通念では，すべての参加者の潜在的なフラストレーションを避けるため，共同作業に先立って明確な合意がなされる必要があるといわれている．[Nese 11] でも議論しているが，我々はこの行動が自然選択によって生成されることを示した [Han 13d]．その研究では，取引結果の帰結を尊重するような先行的な合意（これをコミットメントと呼ぼう）は，たとえその合意にコストがかかるとしても，協力行動を促進（単に非協力行動を罰するよりも）することが明らかとなっている．通常社会通念では，誰かと共同作業を伴う新たなプロジェクトを始めようとするとき，前もってその相手がどの程度それに関与しようとしているのかを明確にすることにコストを払うものである．一方，取引において，強いコミットメントレベル―相手に誓約書を書かせたり，将来起こり得ることに対する保証を明記したりすること―を徹底的に追及することは歓迎されない．

我々の進化ゲーム理論を用いた研究では，コミットメントのおぜん立てのコスト（例えば，契約締結のための弁護士を雇う費用）は共同事業の便益（例えば，家を買うなど）に関して正当化される．もし保証金額が十分大きければコミットメントの提案者は満足し，それゆえ高いレベルの協力行動を引き出すことができる．そのような提案者は嘘の協力者（協力すると同意しておきながらそのような行動をしない者）を排除することができるので，協力コストを搾取しようとする意図をもつ悪人との交流を回避することができる．興味深いことに，我々の研究では，補償金額が一定額以上である（おおよそ，コミットメントのおぜん立てコストと協力の便益の和に一致する額以上である）ときはいつでも，さらなる補償額の増加による改善は見込めない．この結果は，法的契約の場面において，小事における法外な罰金がなぜ要求されないかなどに示唆を与える．

さらに興味深いことに，意図認知とコミットメントのシナジー効果によって，協力がさらに促進されるという研究 [Han 15a, Han 15b] がある．これはエージェントシミュレーションを用いて進化ゲーム理論を適用したものであるが，さまざまな社会における協力メカニズムを示唆している．もし信頼できる合意がなされれば高いレベルの協力が実現される．契約といった正式なコミットメントは，もしそれが十分な強制力をもつとき協力的な社会行動を促進する．そのため，その合意にかかるコストや時間は相互利益となる．

一方，他者の意図を評価できる能力は協力の創発を促進する際に重要な役割を有している．実際，この能力は経験に基づいたものであり，観察は契約といった正式なコミットメントなしに協力行動を促進させられる．すなわち，意図認知とコミットメントのシナジーは意図認知能力の信頼性と正確性に強く依存していることがわかる．高いレベルの協力を実現させるために，意図評価が十分な信頼性や正確性を担保できないときはいつでも，コミットメントは不可欠である．そうでなければ，コストのかかるコミットメントを避けるために意図認知を巧みに使うことは有利となるであろう．

4. コストのかかる懲罰とコミットメントの結合による裏切り防止

我々は，コミットメントとコストのかかる事後的懲罰（コミットメントなしに事後的に非協力者を罰すること）とを比較した．他者を罰するのが十分強力であれば，利己的なプレーヤからなる集団であっても協力は促進される [Fehr 02, Han 16a]．しかしながらそういった研究は，協力を維持させるためにかなりの懲罰コストを必要とするものも少なくない．我々は，事前の同意を行うことで懲罰コストを大幅に抑えつつ協力を維持できることを示した．

より興味深いことに，我々は裏切り行為に対してコミットメントや事後懲罰という2種類の異なる戦略を互いに統合する方法を発見した．はじめに，その二つのメカニズムの単純な確率的結合は，そのどちらか一方のみのときと比べて協力が達成しやすい [Han 16b]．コストのかかる懲罰は，コミットメントに対するフリーライダ

（コミットメントを避けることで，コミットメント戦略者とプレイすることで発生するコストから逃れる人々）に対して効果をもつ．我々の研究結果によれば，両者の結合戦略は協力レベルをかなり改善する．さらにこの協力レベルは，懲罰コストが十分大きく，また罰金があるしきい値を超えているときですら，有意に高い．つまりその結合戦略は，両戦略の弱点を同時に克服することが可能なのである．

両者を結合したほかの研究 [Han 16a] もある．そこでは，興味深いことに，反社会的懲罰（協力者を罰すること）[Powers 12, Raihani 15] を防ぐ新たな解を提供している．一度のみの囚人のジレンマの文脈において，もし懲罰に加えて相手と相互作用する前に協力への同意も提案できるとしたら，反社会的懲罰が実行可能な状況においてさえ，社会的懲罰と協力がともに進化することができる．プレーヤの提案にコミットするというオプションを導入すると，コミットメントを避けるプレーヤに対し補償の支払いを強制させることができ，そのため反社会的懲罰者を有意に抑制することができる．一方，コミットメントのコストがかかるときは，そのコストを払わない社会的懲罰者が優位となる．つまりこの研究では，関与と懲罰という戦略的オプションがともにあるとすると反社会的懲罰者よりも社会的懲罰者のほうが優勢となり，そのどちらかがない場合と比べて協力率が有意に高くなるのである．これは，コミットメントをアレンジすることそれ自体が非常に強力な協力メカニズムになるという知見なのである．懲罰戦略は反社会的懲罰や裏切りに対してぜい弱であるが，それに対するコミットメントに追加コストを支払うことで，協力が達成できる．つまり，コミットメントメカニズムは社会的懲罰と協力が創発する触媒として機能するのである．

5. コミットメントは集団の協力ジレンマを解決する

食料の共有や社会保障システムといった公共財は，すべての参加者が貢献や実施に関してコミットするという契約を事前に結ぶことができればうまくいくだろう．しかし，フリーライダはそのような同意を搾取する [Han 13d] がゆえ，十分な数の参加者がコミットメントに魅力を感じなければいつでもコミットメントの要求は公共財を育てない．この決定はフリーライダの利益をなくすだけでなく，協力して利益をつくろうとする人々も消滅させる．[Han 13b] では，一度きりの公共財ゲームを進化ゲーム理論の枠組みで分析し，非道義的なフリーライダの利益を画する政策を実装することはしばしばより望ましい社会的な帰結をもたらすことを示した．特に大規模集団で公共財の利益が高い状況では，公共財ゲームは集団相互作用における協力の進化を分析する標準的な枠組みである [Sigmund 10]．そこでは，固定された集団サイズにおいてすべてのプレーヤは公共財に貢献するか貢献せずにさぼるかの選択に迫られる．すべての貢献量は定数倍されたのち，貢献するとしないとにかかわらずすべてのプレーヤに均等に配分される．つまり，フリーライダは貢献コストを支払わない分，貢献者に比べて有利となる．このシナリオでは，コミットメントや同意などが協力行動を動機付けるときに本質的に重要な要素となり，それは自然界 [Nesse 01] や実験室実験などでも確認されている．

[Han 13b] では，公共財ゲームを拡張し，集団内相互作用におけるコミットメントベースの戦略を導入している．プレーヤは公共財ゲームをプレイする前に，相手に公共財への貢献にコミットすることを提案し，その同意のために提案者がコストを払うものとする．もしすべての提案が承認されたならば，提案者は全員貢献すると仮定する．コミットしつつ貢献しなかったものは提案者に補償を支払わなければならない [Han 13d]．我々の研究では，協力の利益に関するコミットメントのアレンジコストが正当化されるときはいつでも，特定の戦略は一度きりの公共財ゲームにおいて協力の創発を促進する [Han 13d] ことを明らかにした．

公共財ゲームの文脈でコミットメントに基づく行動に関する別のアプローチ [Han 17a] としては，コミットメントがコストになり，実行するのに追加的努力を要するためいつも実現できるとは限らないと考えるような戦略について考慮したものがある．つまり，集団事業に従事する前に，プレーヤはしばしばグループの他者から参加のレベルに応じて（つまり，他者の何人が集団努力に参加する意思をもつかどうかを決定するかという条件を付与した）先行的なコミットメントを行う [Nesse 01]．このアプローチは，参加者の多数派が公共財への貢献にコミットしているときに限り，その集団事業が着手されるということにインスパイアされたものである．

我々は，集団内の相互作用が協力行動を確保する最小の参加者を設定するようなコミットメントを検討した．その結果，もしそのアレンジコストが協力コストに比べて十分に小さいならば，コミットメントをアレンジすることはしばしば観察され，集団の高い協力率を引き出すことに成功する．さらに，ジレンマとコストの両者に依存して最適な参加率は決定されることもわかった．つまり，公共財のジレンマが厳しくなるにつれ，そして，コミットメントのアレンジコストが増加するにつれ，より多くの参加者が明示的にコミットすることが必要になる．

さらに別のコミットメントの研究 [Han 17b] では，プレーヤはコミットメントのアレンジや参加者の監視をコストのかからない中央集権や制度に委嘱することを検討している．その制度はそれ自身，集団の協力率を高め社会的厚生を増加させるので利益的である（例：政府によってアレンジされた公共交通機関，国連によって支援され

た国際的合意，クラウドソーシング）[Nesse 01]．それはまた，そのサービスを提供するためにすべてのコミットしたプレーヤから得た料金を支給することによる共同事業から直接的に利益を得るかもしれない．このコミットメントを中央集権化するアプローチは個人型コミットメント戦略を凌駕する．個人個人がイニシアティブを行うことに代わって，集団からコミットするアレンジを中央集権システムが肩代わりすることで，コミットメントのフリーライドを防ぎ，完全協力を実現できる [Han 13d, Han 17a]．参加率は合意が形式化されるかどうかについて決定的に重要な役割をもつ．つまり，合意を形式化する中央集権システムを厳格化すればするほど，一度それが機能すれば，協力率に関してより利益的となり，社会的厚生の到達可能レベルを向上させる．

6. 謝罪はなぜ困難か？ コミットメントは誠実さを生む

人は間違えたとき，例えばそれがコストのかかることであったとしても，さらなる協力を確保するために謝罪しようとするものである．同様に，協力行動が他者の一番の興味であることを保証し，かつ，コミットメントの失敗によるペナルティを避けるために個人はコミットメントをアレンジする．それゆえに，謝罪とコミットメントの両者はともに行動進化の文脈で捉えられるべきである．[Han 13b] では，繰返し囚人のジレンマの文脈でこれらの二つの戦略の結合に関する分析を行った．その研究によれば，コミットメントなしの相互作用において謝罪はまれであり，特に協力コストが高いときはほとんど見られなくなる．そのかわり，先行的なコミットメントは謝罪の頻度を有意に増加させる．また，コミットメントの有無にかかわらず，謝罪は真剣で十分かけたときに限り紛争を解決する．さらに興味深いことに，我々のモデルは，個人がコミットした関係者によりコストをかけて謝罪し，それゆえに嘘のコミットメントといったフリーライダをより明確に特定化させるのに役立っていた．

謝罪はおそらく，紛争解決において最も強力で普遍的な方法である [Abeler 10, Ohtsubo 09]．特に結婚をはじめとする長期的関係のある相互作用が存在する個人間において有効だろう．謝罪というものは第三者（教師・親・裁判所）による介在なしに紛争を解決できる．そのような第三者は紛争のあらゆる局面で実体的にコストがかかるものである．謝罪の有効性を支える証拠は医療ミスの状況から買い手と売り手の関係に至るまで無数にある [Abeler 10]．謝罪はプラスの感情や協力を促すものとして，人間機械混在系やオンライン市場といったさまざまな計算されたシステムに実装されている [Tzeng 04, Utz 09]．

[Han 13b] において，エラーが生じたときに明示的に謝罪を行う戦略を含むモデルを検討した．謝罪には，次のラウンド以降に相手が協力してもらえることを目的とした相手への適量の補償を含んでいる．その結果，謝罪するプレーヤのみからなる集団は完全な協力を維持できるものの，そのような謝罪を搾取する行動も創発することがわかった．つまり，他者からの補償を受け入れるものの必要なときに謝罪しない者（＝嘘の謝罪者）が侵入可能であり，彼らは謝罪の便益を破壊する．進化ゲーム理論 [Sigmund 10] を用いた分析によれば，プレーヤが相互作用前にコミットメントを求めることができるシステム [Han 12b, Han 12c, Han 13e] において謝罪をすることは，この搾取を防ぐことができる．この研究から四つの知見を導出した．(1) 謝罪だけでは高い協力率をもたらすことはできない，(2) 先行的なコミットメントによって支援された謝罪は高い協力率をもたらすことができる，(3) 謝罪はコミットメントされた間柄において正当に機能される必要がある [Ohtsubo 09]，(4) よりコストのかかる謝罪は，嘘の謝罪者といったフリーライダを匿名化させてしまうため，コミットメントがない関係よりもコミットメントのある関係において発現される．つまり，コミットメントは誠実さをもたらすのである．

我々の研究は，謝罪やコミットメントメカニズムの設計や運用にとって重要な知見を提供している．例えば，誤りが発生したときにどんな種類の謝罪が顧客に提供されるべきなのだろうか，とか，もし協力を引き出すコミットメントに補完されるならば謝罪（被害にあった顧客への補償）は強められるべきかどうか，といった点である．

7. 謝罪と寛容は協力的な合意の失敗を解決する

どのように振る舞うかについて合意することは，一度きりの社会的ジレンマにおいて進化的に実行可能な戦略として示される．しかし多くの状況において，合意は相互に長期的な互恵関係を築くために結ばれる．この点を最初に検討した我々の解析と数値研究 [Martinez-Vaquero 15, Martinez-Vaquero 17] によれば，繰返し囚人のジレンマの文脈において，合意が継続している採集に生じる誤りを検討し，どのような条件のもとで，復讐や謝罪，寛容が進化するかを示した．その結果，合意の失敗において参加者が，発生した衝突において裏切りによる復讐を好むことが示された．コストのかかる謝罪や寛容が認められるモデルでは，たとえ誤りがしばしば発生したとしても，誤りが合意の破壊をもたらさないような誠実さのしきい値が存在する．要約すれば，たとえ誤ったのが人間だとして，復讐や謝罪，寛容は進化的に実行可能な戦略であり，繰返しジレンマにおける協力を引き出す重要な役割をもっているのである．

我々は進化ゲーム理論の方法を用いて繰返し囚人のジレンマ [Axelrod 81, Axelrod 84] をモデル化し，繰返しのある社会的相互作用における解析的・数理的な知見と

して，コミットメント戦略の実行可能性について示してきた．これはかなりの行動的複雑性が必要とされる．まず，合意は繰返しの相互作用が終了する前に終わらなければならない．そのとき，合意を提案・承認・拒否することといった戦略について考えるべきことがあることを示している．第二に，直接互恵性 [Trivers 71] の文脈で示されていることだが，個人は相手や彼ら自身によってつくられる「震える手」や「ファジィな心」[Nowak 06, Sigmund 10] によって実体化されるような誤りに対処する必要がある．合意を続けるべきかどうかは考慮されるべき事項だ．

エラーが誤解やコミットメントの破壊さえも引き起こすように，個人は誤りを繰り返さないために，あるいは利益のある関係を継続させるための精錬された戦略を採用する．復讐や寛容はそういった状況に正確に立ち向かうために埋め込まれているのかもしれない [McCullough 08, McCullough 10]．懲罰や利益の取下げによってなされる復讐の脅威は人間関係に重大な影響を与える．にもかかわらず，しばしば他者の行動が意図的であるのか，たまたま偶然であるかについて，人は十分に区別することができない [Fischbacher 13, Han 11]．もし偶然だったとすると，寛容は害悪にもかかわらず利益のある関係を継続させるような回復的なメカニズムを提供することができる．我々の研究によれば，寛容の本質的な特筆は，コストのかかる謝罪のように見える点であり [McCullough 08]，これは [Smith 08] によって強調された点でもある．

合意を作成し他者にコミットを求めたりすることは社会のあらゆる階層における基本的な構造メカニズムとなっており，社会的相互作用において重要な役割を果たしている [Nesse 01, Sterelny 12]．我々の研究によれば，繰返しゲームを行う状況で，コミットメントのアレンジコストがかかりすぎることは，重要な役割を果たしている．いくつかのシナリオにおいて，もっと成功する個人はコミットメントを提案し，そのコストを負担する意思を示し，合意に従い，誤りが生じるまで協力する人々である．しかしもしコミットメントが破壊されてしまったら，これらの個人は復讐を行い，以降の対戦において裏切る．これは [McCullough 08] や [McCullough 10] においても解析的に確かめられている．この結果は，興味深い知見を引き出す．つまり，違反した罪人から利益を取り上げることによる復讐が，繰返し囚人のジレンマにおいて協力行動をより好ましく引き出すことができるのである．これはしっぺ返し戦略といったよく知られた互恵的行動とは反対のことである．寛容者のみがコスト便益費が十分大きいときによく振る舞うことができるのである．

そのうえ，長期にわたる関係において誤りが実際は避けられない．誤りが発生したとき，個人は合意をやめるのが価値的か，あるいは補償を求めるべきかどうか，はたまた相手を許し相互の利益的合意を継続するべきかどうかを決める必要がある．この点を確かめるために，コミットメントモデルを拡張し，謝罪・寛容メカニズムを導入してみた．ここで謝罪は外生変数あるいは個人の変数として定義されている．両方のケースにおいて，相手から謝罪を受けたのちに寛容になることは効果的であると示された．しかし，協力を促進する役割として謝罪は真剣さを必要とする．つまり，謝罪のコストは大きくなければならない（ただし大きすぎてはいけない）．これは最近の実験研究 [McCullough 14] によっても裏付けられている．このコミットメントモデルに対する拡張は復讐をベースとしたモデルよりも高い協力率を引き出すことができる．反対に，嘘の契約者（コミットメントは受け入れるのに裏切りは謝罪を繰り返すというシステム的な有利さを意図する人々）が集団を占める．この状況では，謝罪・寛容メカニズムの導入は，彼ら自身によるコミットメントを生成するような協力を破壊する．先行研究 [Martinez-Vaquero 13, Martinez-Vaquero 14] では，誤りは社会の中で不正を惹起し不誠実を冗長する．ただ厳格な倫理のみがそれを防ぎ得る．

繰返しゲームではコミットメントは忠誠の形をとるかもしれない [Schneider 13]．つまり，後で生じる補償に関する契約とは異なる．忠誠を誓うコミットメントは，個人はこれまでの相互作用の長さに基づいてとどまるか相手を選択するかといった傾向をもつという考えに基づいている．我々は，たとえパートナー選択がなくても，特に誤りが生じたとき，あるいは誠実な謝罪や寛容に伴ったときに，コミットメントは協力や長期関係を助長できることを示した．

8. お わ り に

道徳的な意思決定能力を有する自動化エージェントが社会に組み込まれていくにつれ，機械倫理は学際的に重要で萌芽的な研究分野となる．その研究は，道徳判断能力を有するエージェントの実装という現実的問題にとってのみならず，倫理を計算するモデルの構築や検証によって，道徳とは何かという根本的命題にとっても示唆を与えることになる．機械倫理の計算モデルは通常，よく定義され，動学が観察可能であり，探索的にさまざまな別のモデルと置き換えることができる．本稿では，機械倫理研究のサーベイと，我々自身が行ってきた進化ゲーム理論を用いたモデルや実験を紹介し，集合的分野における協力の創発と進化が機械倫理研究にもたらす知見について概観する．ここで機械倫理とは，道徳機械の設計，マルチエージェントシステム，約束アルゴリズム，そしてそれらの人間–機械系への応用を想定している．

謝　辞

引用した論文の共著者 Ari Saptawijaya, Francisco C.

Santos, Luis Martinez-Vaquero, Tom Lenaerts の各氏に感謝する．Pereira は研究資金 FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013 の，Pereira と Han は研究資金 Future of Life Institute grant RFP2-154 の支援を受けた．

◇ 参 考 文 献 ◇

[Abeler 10] Abeler, J., Calaki, J., Andree, K. and Basek, C.: The power of apology, *Economics Lett.,* Vol. 107, No. 2, pp. 233-235（2010）

[Axelrod 81] Axelrod, R. and Hamilton, W. D.: The evolution of cooperation, *Science,* Vol. 211, No. 27, pp. 1390-1396（1981）

[Axelrod 84] Axelrod, R.: *The Evolution of Cooperation,* p. 5145, Basic Books（AZ）（1984）

[Charniak 93] Charniak, E. and Goldman, R. P.: A Bayesian model of plan recognition, *Artificial Intelligence,* Vol. 64, No. 1, pp. 53-79（1993）

[Fehr 02] Fehr, E. and Gächter, S.: Altruistic punishment in humans, *Nature,* Vol. 415, No. 6868, pp. 137-140（2002）

[Fischbacher 13] Fischbacher, U. and Utikal, V.: On the acceptance of apologies, *Games and Economic Behavior,* Vol. 82, pp. 592-608（2013）

[Han 11] Han, T. A., Pereira, L. M. and Santos, F. C.: Intention recognition promotes the emergence of cooperation, *Adaptive Behavior,* Vol. 19, No. 4, pp. 264-279（2011）

[Han 12a] Han, T. A., Pereira, L. M. and Santos, F. C.: Corpus-based intention recognition in cooperation dilemmas, *Artificial Life,* Vol. 18, No. 4, pp. 365-383（2012）

[Han 12b] Han, T. A., Pereira, L. M. and Santos, F. C.: Intention recognition, commitment and the evolution of cooperation, *2012 IEEE Congress on Evolutionary Computation,* pp. 1-8（2012）

[Han 12c] Han, T. A., Pereira, L. M. and Santos, F. C.: The emergence of commitments and cooperation, *Proc. 11th Int. Conf. on Autonomous Agents and Multiagent Systems,* Vol. 1, pp. 559-566（2012）

[Han 12d] Han, T. A., Saptawijaya, A. and Pereira, L. M.: Moral reasoning under uncertainty, *Int. Conf. on Logic for Programming Artificial Intelligence and Reasoning,* pp. 212-227（2012）

[Han 13a] Han, T. A. and Pereira, L. M.: Intention-based decision making via intention recognition and its applications, *Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security,* pp. 174-211, IGI Global（2013）

[Han 13b] Han, T. A., Pereira, L. M., Santos, F. C. and Lenaerts, T.: Why is it so hard to say sorry? Evolution of apology with commitments in the iterated Prisoner's Dilemma, *Proc. 23rd Int. Joint Conf. on Artificial Intelligence,* pp. 177-183（2013）

[Han 13c] Han, T. A. and Pereira, L. M.: State-of-the-art of intention recognition and its use in decision making, *AI Communications,* Vol. 26, No. 2, pp. 237-246（2013）

[Han 13d] Han, T. A., Pereira, L. M., Santos, F. C. and Lenaerts, T.: Good agreements make good friends, *Scientific Reports,* Vol. 3（2013）

[Han 13e] Han, T. A.: Intention recognition, commitments and their roles in the evolution of cooperation: from artificial intelligence techniques to evolutionary game theory models, *SAPERE Series,* Vol. 9（2013）

[Han 13f] Han, T. A. and Pereira, L. M.: Context-dependent incremental decision making scrutinizing the intentions of others via bayesian network model construction, *Intelligent Decision Technologies,* Vol. 7, No. 4, pp. 293-317（2013）

[Han 15a] Han, T. A., Pereira, L. M., Santos, F. C. and Lenaerts, T.: Emergence of cooperation via intention recognition, commitment and apology — A research summary, *AI Communications,* Vol. 28, No. 4, pp. 709-715（2015）

[Han 15b] Han, T. A., Santos, F. C., Lenaerts, T. and Pereira, L. M.: Synergy between intention recognition and commitments in cooperation dilemmas, *Scientific Reports,* Vol. 5（2015）

[Han 15c] Han, T. A., Pereira, L. M. and Lenaerts, T.: Avoiding or restricting defectors in public goods games?, *Journal of the Royal Society Interface,* Vol. 12, No. 103, 20141203（2015）

[Han 16a] Han, T. A.: Emergence of social punishment and cooperation through prior commitments, *Proc 30th AAAI Conf. on Artificial Intelligence,* pp. 2494-2500（2016）

[Han 16b] Han, T. A. and Lenaerts, T.: A synergy of costly punishment and commitment in cooperation dilemmas, *Adaptive Behavior,* Vol. 24, No. 4, pp. 237-248（2016）

[Han 17a] Han, T. A., Pereira, L. M. and Lenaerts, T.: Evolution of commitment and level of participation in public goods games, *Autonomous Agents and Multi-Agent Systems,* Vol. 31, No. 3, pp. 561-583（2017）

[Han 17b] Han, T. A., Pereira, L. M., Martinez-Vaquero, L. A. and Lenaerts, T.: Centralized vs. personalized commitments and their influence on cooperation in group interactions, *Proc. 31st AAAI Conf. on Artificial Intelligence,* pp. 2999-3005（2017）

[Han 18] Han, T. A. and Pereira, L. M.: *Evolutionary Machine Ethics,* in O. Bendel（ed., *Handbuch Maschinenethik,* Springer（2018）

[Hauser 06] Hauser, M.: *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong,* Ecco/Harper Collins Publishers（2006）

[Martinez-Vaquero 13] Martinez-Vaquero, L. A. and Cuesta, J. A.: Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation, *Phys. Rev. E.,* Vol. 87, No. 5, p. 052810（2013）

[Martinez-Vaquero 14] Martinez-Vaquero, L. A. and Cuesta, J. A.: Spreading of intolerance under economic stress: Results from a reputation-based model, *Phys. Rev. E.*, Vol. 90, No. 2, p. 022805（2014）

[Martinez-Vaquero 15] Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M. and Lenaerts, T.: Apology and forgiveness evolve to resolve failures in cooperative agreements, *Scientific Reports,* Vol. 5（2015）

[Martinez-Vaquero 17] Martinez-Vaquero, L. A., Han, T. A. and Pereira, L. M.: Lenaerts, T.: When agreement-accepting free-riders are a necessary evil for the evolution of cooperation, *Scientific Reports,* Vol. 7（2017）

[McCullough 08] McCullough, M. E.: *Beyond Revenge: The Evolution of the Forgiveness Instinct,* John Wiley & Sons（2008）

[McCullough 10] McCullough, M. E., Kurzban, R. and Tabak, B. A.: Evolved mechanisms for revenge and forgiveness, *Understanding and Reducing Aggression, Violence, and Their Consequences,* pp. 221-239（2010）

[McCullough 14] McCullough, M. E., Pedersen, E. J., Tabak, B. A. and Carter, E. C.: Conciliatory gestures promote forgiveness and reduce anger in humans, *Proc. National Academy of Sciences,* Vol. 111, No. 30, pp. 11211-11216（2014）

[Mikhail 07] Mikhail, J.: Universal moral grammar: Theory, evidence and the future, *Trends in Cognitive Sciences,* Vol. 11, No. 4, pp. 143-152（2007）

[Nesse 01] Nesse, R. M.: Natural selection and the capacity for subjective commitment, *Evolution and the Capacity for Commitment,* pp. 1-44, Russell Sage Foundation（2001）

[von Neumann 44] von Neumann, J. and Morgenstern, O.: T*heory of Games and Economic Behavior,* Vol. 1, Princeton: Princeton University Press（1944）

[Nowak 06] Nowak, M. A.: Five rules for the evolution of cooperation, *Science,* Vol. 314, No. 5805, pp. 1560-1563（2006）

[Ohtsubo 09] Ohtsubo, Y, and Watanabe, E.: Do sincere apologies need to be costly?, Test of a costly signaling model of apology, *Evolution and Human Behavior,* Vol. 30, No. 2, pp. 114-123（2009）

[Pereira 09] Pereira, L. M. and Saptawijaya, A.: Modelling morality with prospective logic, *Int. J. Reasoning-Based Intelligent Systems,* Vol. 1, No. 3-4, pp. 209-221（2009）

[Pereira 11] Pereira, L. M. and Han, T. A.: Intention recognition

with evolution prospection and causal Bayes networks, *Computational Intelligence for Engineering Systems,* pp. 1-33, Springer Netherlands（2011）
[Pereira 12] Pereira, L. M.: Turing is among us, *J. Logic and Computation,* Vol. 22, No. 6, pp. 1257-1277（2012）
[Pereira 15] Pereira, L. M. and Saptawijaya, A.: Abduction and beyond in logic programming with application to morality, In: Magnani, L.（Ed.）, *IfColog J. Logics and Their Applications, Special issue on Abduction,* Vol. 3, No. 1, pp. 37-71, London: College Publications（2015）
[Pereira 16a] Pereira, L. M. and Saptawijaya, A.: Bridging two realms of machine ethics, *Programming Machine Ethics,* pp. 159-165, Springer International Publishing（2016）
[Pereira 16b] Pereira, L. M. and Saptawijaya, A.: *Programming Machine Ethics,* Springer Berlin（2016）
[Pereira 17] Pereira, L. M. and Saptawijaya, A.: Counterfactuals, logic programming and agent morality, *Applications of Formal Philosophy,* pp. 25-53, Springer Berlin（2017）
[Powers 12] Powers, S. T., Taylor, D. J. and Bryson, J. J.: Punishment can promote defection in group-structured populations, *J. Theoretical Biology,* Vol. 311, pp. 107-116（2012）
[Prinz 16] Prinz, J.: Emotions, morality, and identity, *Morality and Emotion,* Vol. 13, No. 34, pp. 83-98, London: Routledge（2016）
[Raihani 15] Raihani, N. J. and Bshary, R.: The reputation of punishers, *Trends in Ecology & Evolution,* Vol. 30, No. 2, pp. 98-103（2015）
[Sadri 11] Sadri, F.: Logic-based approaches to intention recognition, *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives,* pp. 346-375（2011）
[Sigmund 10] Sigmund, K.: *The Calculus of Selfishness,* Princeton University Press（2010）
[Smith 08] Smith, N.: *I Was Wrong: The Meanings of Apologies,* Cambridge University Press（2008）
[Saptawijaya 15a] Saptawijaya, A. and Pereira, L. M.: Logic programming applied to machine ethics, *Portuguese Conf. on Artificial Intelligence,* pp. 414-422（2015）
[Saptawijaya 15b] Saptawijaya, A. and Pereira, L. M.: The potential of logic programming as a computational tool to model morality, *A Construction Manual for Robots' Ethical Systems,* pp. 169-210, Springer International Publishing（2015）
[Saptawijaya 18] Saptawijaya, A. and Pereira, L. M.: *From Logic Programming to Machine Ethics,* In O. Bendel（ed.）, *Handbuch Maschinenethik,* Berlin: Springer（2018）
[Sterelny 12] Sterelny, K.: *The Evolved Apprentice,* MIT Press（2012）
[Schneider 13] Schneider, F. and Weber, R. A.: Long-term commitment and cooperation, *Tech. Rep., Working Paper Series*, Department of Economics, University of Zurich（2013）
[Trivers 71] Trivers, R. L.: The evolution of reciprocal altruism, *Quarterly Review of Biology,* Vol. 46, No. 1, pp. 35-57（1971）
[Tzeng 04] Tzeng, J. Y.: Toward a more civilized design: Studying the effects of computers that apologize, *Int. J. Human-Computer Studies,* Vol. 61, No. 3, pp. 319-345（2004）
[Utz 09] Utz, S., Matzat, U. and Snijders, C.: On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions, *Int. J. Electronic Commerce,* Vol. 13, No. 3, pp. 95-118（2009）

2019 年 1 月 15 日　受理

著者・訳者紹介

Han The Anh
ベルギーのブリュッセル大学 AI 研究所での 2 年間の FWO ポスドク研究員を経て 2012 年にポルトガルの新リスボン大学において計算機科学分野で Ph. D. を取得．2014 年よりイギリス・ティーズサイド大学計算・メディア・アート学部上級講師．研究関心は AI や横断的分野において，人間の協力動学，AI 認知モデル，進化ゲーム理論，エージェントモデル，行動経済学，意図認知，知識表現や推論など多岐にわたる．

Pereira Luís Moniz
新リスボン大学名誉教授，NOVA-LINCS 連携研究員，EurAI フェロー（2001 年より），2006 年ドレスデン大学名誉博士号．2006 年よりマドリード高等ソフトウェア研究所理事ならびに科学アドバイザー．1984 年にポルトガル人工知能学会（APPIA）初代会長．知識表現，推論，論理プログラミング，認知科学，進化ゲーム理論研究に従事．

岡田　勇
創価大学経営学部准教授．1995 年創価大学工学部卒業，2000 年電気通信大学大学院情報システム学研究科にて博士（学術）号を取得．電気通信大学助手を経て，2007 年より現職．2014 ～ 18 年本学会編集委員として本特集号の企画を担当．社会情報学会理事（2014 ～ 18），事務局長（2018 ～）．専門は計算社会科学，進化ゲーム理論，社会的ジレンマなど．