裏切り者の魅力的な微笑み：信頼ゲームを用いた再検討

大久保　街　亜

協調行動や社会的な交換は、ヒトの社会に見られる普遍的な特徴である。ただし、協調行動が進化し、社会的交換が公正に行われる社会を実現するためには、社会交換における裏切り者、つまり、利益を得るが対価を払わないフリーライダーを排除しなければならない。このようなアイデアに基づきCosmides (1989) は、ヒトが裏切り者検出メカニズムを進化の過程で獲得し、それが協調行動の進化を可能にしたという仮説を提唱した。彼女はこの仮説を検証するため、Wason 選択課題（Wason, 1966）を用い実験を行った。この課題では、４枚のカードから課題要求に合うカードを選択することが求められる。これら４枚のカードには、一方の面にはアルファベット、もう一方の面には数字が書いてある。そして、被験者は「カードの一方の面に母音が書かれているなら、もう一方の面には偶数が書かれている」という規則が正しいか確かめることを求められる。このとき制約があり、必ず確かめなければならないカードしか選択ができない。Wason (1966) が作成したオリジナルの選択課題は難易度が極めて高い。このような数字とアルファベットを使ったオリジナルの問題で、正答率はせいぜい 10%程度である。Cosimides は、この単純ではあるが難解な問題の正答率が、課題に裏切り者を検出する文脈を与えることで飛躍的に上昇することを示した。例えば、「ビールを飲んでいるなら、20 歳を越えていなければならない」というルールが守られているか確かめるという文脈を設定すると、たいていの被験者は正答することが可能だ（ただし、この例は Griggs & Cox, 1982 のものである）。これはビールを飲むという利益を得ているのに、成人年齢になっているという対価を払っていない裏切り者を見つけるという文脈が、われわれが有する裏切り者検出メカニズムと適合するからである。さらに、Cosmides は、論理的な正解を逆転させ

た場合も、被験者は裏切り者を検出する文脈に従って回答することを示した。これらの結果から Cosmides は、裏切り者を検出するメカニズムをヒトは生得的に備えており、それが協調行動の進化における基盤となっていると主張した。

この Cosmides による裏切り者検出仮説に基づいて、さまざまな研究が行われた。多くの研究は、ヒトが裏切り者に極めて敏感であることを示しており、Cosmides の仮説を支持するものである。例えば、この仮説に基づくと、人々は（社会的交換における）裏切り者の顔を良く記憶することが予測され、実際にそれを支持する実験結果が得られている（Mealey, Daood, & Krage, 1996; Oda, 1997; Yamagishi, Tanida, Mashima, Shimoma, & Kanazawa, 2003）。このような 裏切り者の顔に対する再認優位性は、再認課題を行う被験者が実際の裏切り行為を体験、あるいは目撃せずとも生ずることがわかっている（Mealey et al., 1996; Oda, 1997; Yamagishi et al., 2003）。Yamagishi et al. (2003) は、囚人のジレンマゲームを用い、このゲームにおいて、協調行動の回数が多い被験者を協力者、少ない被験者を裏切り者と定義した。その後、協力者と裏切り者の顔写真を用い、刺激人物を知らない別の被験者を用いて再認課題を行った。実験の結果、協力者に比べ、裏切り者の顔に対する再認成績が高かった。この結果は、顔写真として提示された顔情報のみから、社会的交換に関わる利益を不当に受益する「裏切り者」を見破ることができることを示している。この結果は Verplaetse, Vanneste, and Braeckman (2007) によっても追試され、頑健性が確認されている。

ヒトが非協力者に敏感なことは注意の側面からも検討されている。Vanneste, Verplaetse, Van Hiel, and Braeckman (2007) はドット・プローブ課題を使用した実験を行った。彼らは協力者と裏切り者の顔写真をプライム刺激として用いその効果を比較した。実験の結果、裏切り者の写真がプライム 刺激として提示されたとき、協力者のときと比べ、ターゲット課題の成績が向上した。この結果は、裏切り者の顔が注意を言わば自動的に奪取することを示唆する。

これらの実験は協力者と非協力者には外見上の違いがあり、ヒトはそれらを利用して対人コミュニケーションを行っていることを示唆する。実際、い

くつかの研究は、ヒトが外見から裏切り者と協力者を見分けることができることを示した。Franks, Gilovich, and Regan (1993) は、初対面の被験者同士に囚人のジレンマゲームを行わせた。また、ジレンマゲームを行う前に、30 分の会話を行い、ゲームのパートナーが行う選択を予測させた。パートナーの選択は、偶然よりも高い確率で正しく予測された。協力の選択については 80%、非協力の選択については 57%の正確さであった。さらに、Vanneste et al. (2007) は、事前のコンタクトを被験者同士で行わせず、Franks et al. (1993) の実験を追試した。その結果およそ 60%の確率でパートナーの選択が正しく予測できることがわかった。また、Brown, Palameta, and Moore (2003) は、モデルに短いストーリーを音読させた。その音読の様子をビデオ画像で観たあと、モデルと面識のない被験者は、モデルの利他性を見た目から評定した。その結果、被験者が評定した利他性とモデルが自己評定した利他性に正の相関があることが示された。Oda, Yamagata, Yabiku, and Matsumoto-Oda (2009) は、日常により近い状況を用い、この実験結果を再現した。利他性の低さは社会交換における利益を一方的に受ける裏切り者である可能性を示唆する。従って、Brown et al. (2003) や Oda et al. (2009) の結果は、ヒトが見た目に基づいて裏切り者を見分けていることを示している。

それでは、協力者と裏切り者には外見上どのような違いがあり、われわれはどのような側面に着目しているのだろうか？ Vanneste et al. (2007) は、囚人のジレンマのような社会的ゲームで定義された裏切り者と協力者には、感情の表出に違いがあることを指摘した。彼らの実験では裏切り者の顔は、協力者に比べ、攻撃的で恐ろしく見えると評定された。彼らはこのような感情表出の特性が、上述のドット・プローブ課題において裏切り者が注意を奪取した理由ではないかと考察した。Carré らは、テストステロンが過剰に分泌される男性の顔幅が広くなることに着目し、顔幅と攻撃行動の関連を検討した。実験の結果、実験室においても、日常場面においても、顔幅が広くなるほど攻撃的な行動をとることが示された（Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009）。さらに、Stirrat and Perrett (2010) は、顔幅の広くなるほど、信頼ゲームと呼ばれる経済ゲームにおける裏切り行為が増加し、さらに他者が評定する見た目の信頼感も下がることを示した。

Oosterhof and Todorov (2008) は、コンピュータ・グラフィックスにより顔画像を作成し、さまざまな表情をした顔画像の評定を行った。その結果、顔から受ける印象は信頼感と支配性の 2 軸にまとめられることがわかった。そして、信頼感の評定は表情と直接関連することがわかった。具体的には、信頼感は笑いの強度とともに増加し、怒りの強度とともに減少することが示された。これらの結果は、ヒトが、他者の顔から受ける攻撃的な印象や、表情として現れる怒りや笑いを用いて、裏切り者を検出している可能性を示唆する。すなわち、攻撃的な怒り顔は、信頼感が低く、裏切り者と判断されがちであるのに対し、協調的な笑顔は、信頼感が高く、協力者と判断されがちであると結論することができる。

このように表情や見た目の攻撃性から、裏切り者を検出できる可能性は示唆されてきた（Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009; Oosterhof & Todorov, 2008; Stirrat & Perrett, 2010）。また、実験室実験において、ある程度の正確さで裏切り者と協力者を見分けることが示された（Brown et al, 2003; Franks et al., 1993; Oda et al., 2009）。しかしながら、現実世界で裏切り者を検出するのは決して簡単なことではない。詐欺や背任など深刻な裏切り行為の結果、多くの犯罪が報道されていることを考えてもその困難さがわかる。また、実験室実験においても、裏切り者検出の正確さは、60%程度にすぎないことが多い（Brown et al, 2003; Franks et al., 1993）。つまり、裏切り者検出は可能であるものの、決して精度の高いものではない。

我々は、これまで、裏切り者検出の精度が低くなる原因を探るためにいくつか実験を行った（e.g., Okubo, Kobayashi, & Ishikawa, 2012）。怒りや攻撃性が、裏切り者検出のシグナルとなるなら（Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009; Oosterhof & Todorov, 2008; Stirrat & Perrett, 2010）、このシグナルを打ち消すあるいは弱めることにより、検出の精度が落ちると考えられる。そこで我々は、顔の信頼感が怒りとともに低下し、笑いとともに上昇すること（Oosterhof & Todorov, 2008）に着目した。すなわち、社会交換における裏切り者は、笑顔を巧みに操ることによって、裏切り者検出から逃れるという仮説をたてた。この仮説に基づくと、(1) 怒り顔について、裏切り者は、協力者よりも信頼感が低く評定されること（怒りシグナル

による裏切り者検出)、(2) 笑顔では、この信頼感の差が消失すること（笑いによる怒りシグナルの打ち消し)、(3) 裏切り者は、協力者より、強い強度の笑顔を作ることという 3 つの予測が導かれる。これら予測を検証するため、われわれは経済ゲームの一種である信頼ゲームの成績により、客観的指標から裏切り者と協調者を定義した。そして、彼らを写真モデルとして、信頼感と表情強度の評定を行った。実験結果は予測とすべて一致した。つまり、怒り顔について、裏切り者は、協力者よりも、信頼感が低く評定され、この信頼感の差が笑顔では消失した。さらに、裏切り者は、協力者より、笑顔の強度が高く評定された。この結果は、社会交換における裏切り者は、笑顔を巧みに操ることによって、裏切り者検出から逃れるという仮説を支持する。

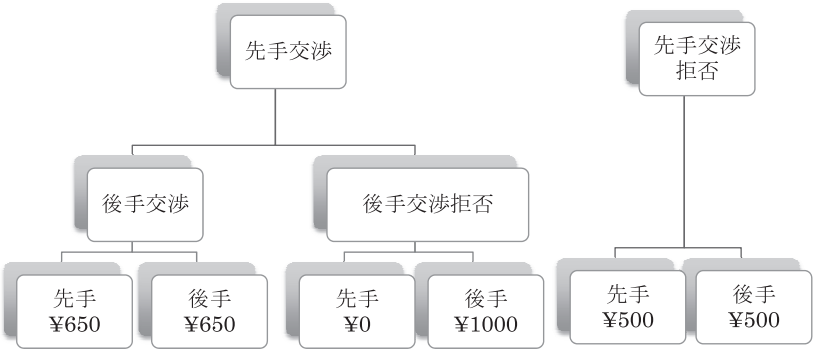

図 1：信頼ゲームにおける役割とルールの概要

本論文の目的は、Okubo et al. (2012) における方法論上の問題点について検討するとともに、その知見の頑健性を確認することである。Okubo et al. (2012) において、モデル候補者はコンピュータ上のパートナーと信頼ゲームを行った。その成績により裏切り者と協力者のモデルが選定された。Okubo et al. (2012) で採用した信頼ゲームにおける役割とルールの概要を図 1 に示した。信頼ゲームの各試行においてモデル候補者は、先手をとるか後手をとるかがランダムに決定された。このゲームにおいて、先手の役割を与えられたとき、モデル候補者は後手と交渉するか、交渉を拒否するか選択することが求めら

れた。一方、後手の役割となり、かつ、先手が交渉を選択したとき、モデル候補者は先手と交渉するか、交渉を拒否するか選択することが求められた。Okubo et al. (2012) では、後手となり交渉を拒否した回数を裏切り行為の指標ととらえ、その回数に基づいて上位 28%を裏切り者、下位 28%を協力者と定義した。

Okubo et al. (2012) が用いた裏切り者の定義は、Stirrat and Perrett (2010) の手続きに則ったものである。ただし、裏切り者を正確に判別できたか若干の疑義がある。図 2 に Okubo et al. (2012) における後手交渉における拒否回数（裏切り得点）のヒストグラムを示した。一見してわかるように、右端に大きな山がある偏った分布となっている。これは多くのモデル候補者が、裏切り得点の最大値をとったためである。このような偏りのある分布の場合、天井効果が生じ、裏切り得点に基づく分類が必ずしも正確になされていなかった可能性がある。

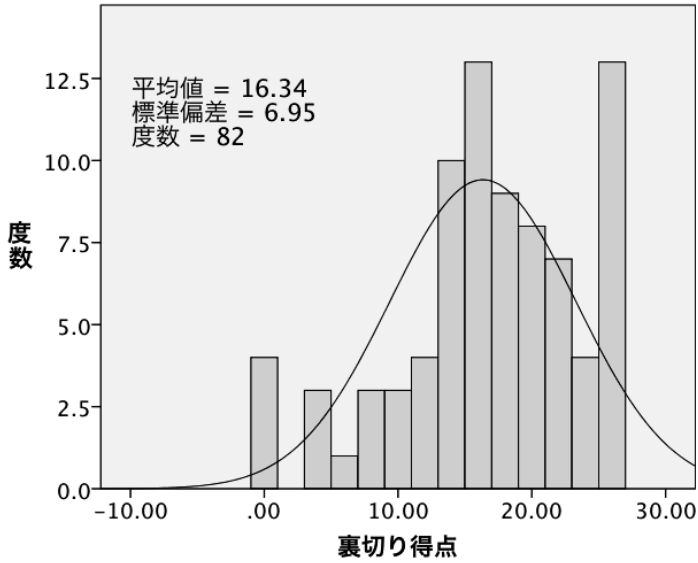

図 2：Okubo et al. (2012) における裏切り得点
（後手交渉における拒否回数）の分布

そこで、本研究では異なる裏切りの指標を用いて、Okubo et al. (2012) の仮説、すなわち、社会交換における裏切り者は、笑顔を巧みに操ることによって裏切り者検出から逃れるというアイデアを再検討する。それにあたり、裏切りの指標として、先手、後手を問わず、交渉を拒否した回数を用いることとした。図 1 に示した通り、先手として交渉を拒否した場合、先手が交渉した場合と比較して、後手が獲得する金額は低くなる。従って、後手が得る将来の利益を減ずるという点から、先手の交渉拒否も、パートナーに対する裏切り行為と位置づけることができる。そこでこのような定義のもと、Okubo et al. (2012) と同じモデル候補者から裏切り者と協力者を選択した。そして、彼らの顔写真を用いて信頼感と感情強度の評定を行った。Okubo et al. (2012) の知見が頑健なものであり、その仮説が支持されるならば、裏切り者の定義の違いに関わらず、実験結果が再現されると予測される。すなわち、(1) 怒り顔について、裏切り者は、協力者よりも、信頼感が低く評定されること（怒りシグナルによる裏切り者検出の成功)、(2) 笑顔では、この信頼感の差が消失すること（笑いによる怒りシグナルの打ち消し)、(3) 裏切り者は、協力者より、強い強度の笑顔を作ることが予測される。

方法

被験者

大学生 152 名（男性 134 名、女性 18 名）が実験に参加した。すべての被験者のうち、84 名の男性がモデル候補者であった。モデル候補者は専修大学に通う男子大学生が参加した。本研究の仮説の一部である攻撃性と裏切り者の関連は男性モデルにおいて確認され、その説明も男性ホルモンという男性に特に強い関連のある生物学的特質であったため（Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009; や Stirrat & Perrett, 2010)、本研究では男性モデルのみを採用した。

一方、顔写真に対する評定実験の評定者として、68 名（男性 50 名、女性 18 名）が参加した。評定者は上智大学に通う大学生であった。モデル候補者と評定者は、異なる大学に通っており、事前に知り合いではなかった。なお、

評定者には実験終了後に、既知の顔が評定実験で呈示されたか確認を行った。

実験計画

独立変数は、モデル（裏切り者 vs. 協力者）、評定項目（信頼感 vs. 感情強度）、刺激表情（笑い vs. 怒り）の 3 要因混合計画であった。モデルと評定項目は被験者内、刺激表情は被験者間で操作した。従属変数は、信頼感と感情強度の評定値であった。それぞれの評定値は 7 段階尺度で評定された（1=まったくない --- 7=とても強い）。

刺激

信頼ゲーム：裏切り者と協力者を選別するため、84 名のモデル候補者が信頼ゲームに参加した。信頼ゲームの各試行においてモデル候補者は、先手をとるか後手をとるかランダムに決定された。信頼ゲームにおける役割とルールの概要を図 1 に示した。このゲームにおいて、先手の役割を与えられたとき、モデル候補者は後手と交渉するか、交渉を拒否するか選択することが求められた。一方、後手の役割となり、かつ、先手が交渉を選択したとき、モデル候補者は先手と交渉するか、交渉を拒否するか選択することが求められた。（1）先手と後手の双方が交渉に合意した場合、双方が ¥650 を受け取った。（2）先手が交渉を拒否した場合、双方が ¥500 を受け取った。（3）先手が交渉に合意し、後手が拒否した場合、先手は何も受け取れず（¥0）、後手は ¥1000 を受け取った。信頼ゲームはすべてで 100 試行あった。実験終了後、モデル候補者は信頼ゲームで獲得した金額のおよそ 1/10 を受けとった。

信頼ゲームに参加した 84 名のモデル候補者について、先手と後手の交渉拒否回数に基づいて、裏切り者と協力者を定義した。先手と後手の交渉拒否回数を反社会性得点とし、モデル候補者ごとに集計した。図 3 にその結果をヒストグラムとして示した。裏切り得点の上位 25% ($n = 21$) を裏切り者、下位 25% (n=21) を協力者と定義した。裏切り得点は、裏切り者で ($M = 62.0$, $SD = 9.95$)、協力者よりも有意に高かった ($M = 18.8$, $SD = 9.32$), t (40) = 14.5, $p < .001$)。

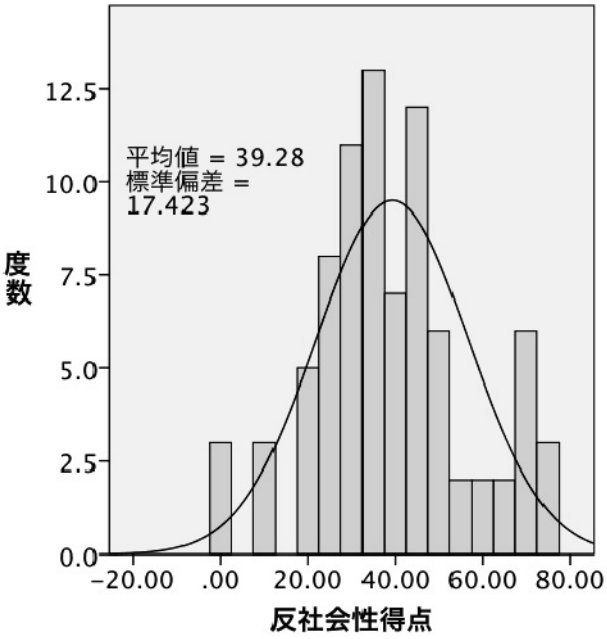

図 3：反社会性得点（交渉拒否回数）の分布

顔写真とその撮影：信頼ゲーム終了後、すべてのモデル候補者は、顔写真の撮影を行った。明るい照明の部屋で、無地の壁を背景にし、デジタルカメラ（Nikon Coolpix S610）を用いて、モデル候補者の撮影は行われた。各モデル候補者について、笑い、怒り、中立の 3 表情について撮影を行った。笑いと怒りの表情については、できるだけ強い感情強度で表情を浮かべることが求められた。また、すべての表情について、カメラはモデルの顔の正面におかれ、モデルは撮影時にカメラのレンズを見つめるよう求められた。

裏切り者と協力者それぞれの群 21 名、計 42 名が、モデルとして選出された。選出された 42 名の各モデルについて、笑顔および怒り顔の写真を評定課題に用いた。すべての写真でモデルは正面を向いており、顔は画像の中央に配置された。写真のサイズは視角にしておよそ 5 度の大きさであった。また裏切り得点の中央値から上位、下位それぞれ 5 名、計 10 名の無表情の刺激を

フィラー刺激としてもちいた。従って、評定課題で用いた写真の枚数はすべてで 94 枚であった。

手続き

刺激の提示と反応の取得にあたりパーソナルコンピュータと LCD ディスプレイを使用した。プログラムは Windows 版 Active Perl 5.8 で作成し、Internet Explorer 7 を用いて表示した。

各試行において、画面の上部に顔刺激、下部に評定項目を配置した。被験者は顔写真に対して表情強度と信頼感を評定した。評定は 1（全くない）-4（中くらい）-7（とても強い）の 7 段階で行われた。半数の被験者は笑顔の写真に対して、残りの半数が怒り顔の写真に対して評定を行った。

全試行数は 52 試行であった。裏切り者と協力者の写真それぞれ 21 枚に加え、無表情刺激 10 枚がフィラー刺激として追加された。裏切り者と協力者の写真のみを実験に用いると、笑顔条件では笑顔が、怒り顔条件では怒り顔のみが連続して呈示されることになる。これは表情を被験者間要因として操作したためである。また、顔写真の感情強度はどれも強い。なぜなら、撮影時にできるだけ強い感情強度を浮かべるようモデルに求めたからである。呈示

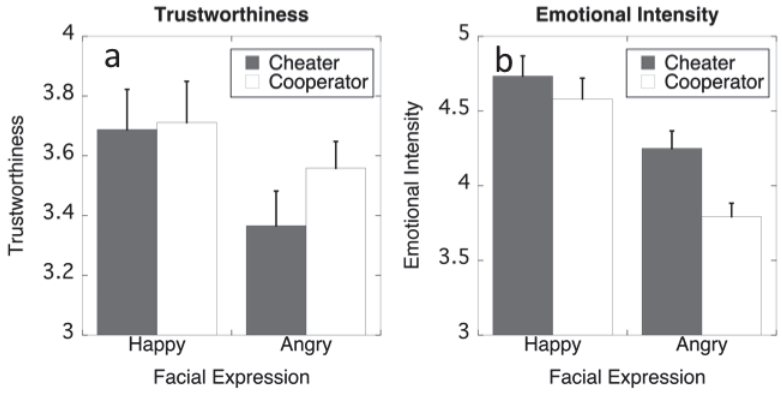

図 4：裏切り者と協力者の顔写真に対する(a)信頼感と(b)感情強度の平均評定値．エラーバーは標準誤差である。

される刺激全体の感情強度が均一になることをさけ、被験者の評定を容易にするためフィラー刺激を用いた。

結果と考察

各被験者について、信頼感と感情強度の評定値について平均値を求めた。笑顔条件に割り当てられた 2 名の被験者について、コンピュータ・プログラムのエラーにより一部の試行でデータを取得できなかった。そのためこの 2 名のデータは分析から除外した。図 4 に、信頼感と感情強度の平均値を示した。この信頼感と感情強度のそれぞれの平均評定値に対して、2 要因の混合計画の分散分析を行った。2 要因のうち、表情（笑顔 vs. 怒り顔）を被験者間要因としてモデル（裏切り者 vs. 協調者）を被験者内要因として操作した。まず、信頼感の分析結果、続いて、感情強度の分析結果について述べる。図 4a に信頼感の結果を示した。モデルの主効果が有意であり、裏切り者の信頼感が、協力者よりも低く評定された、$F\,(1, 64) = 4.30, p < .042, \eta_p^2 = .063$。これは裏切り者検出が成功したことを示している。モデルと表情の交互作用は有意水準に届かなかったが、$F\,(1, 64) = 2.62, p = .11, \eta_p^2 = .039$、笑顔と怒り顔のそれぞれの条件について、先験的な予測に基づき裏切り者と協力者の信頼感について、表情ごとに対比較を行った。予測通り、怒り顔で、裏切り者の信頼感は、協力者よりも高く評定された、$t\,(33) = 2.34, p = .025$。しかし、この差はモデルが笑顔を浮かべると消失した、$t\,(31) = .38, p = .709$。

図 4b に感情強度の平均値を示した。モデルの主効果が有意であり、裏切り者は協力者より強く感情を表出した、$F\,(1, 64) = 44.76, p < .001, \eta_p^2 = .412$。表情の主効果も有意で、笑顔の感情強度が怒り顔よりも高かった、$F\,(1, 64) = 16.16, p < .001, \eta_p^2 = .202$。さらにモデル属性と表情の交互作用も有意であり、$F\,(1, 64) = 11.07, p = .001, \eta_p^2 = .147$、怒り顔において、裏切り者における感情強度の優位性が、笑顔よりも大きくなった。ただし、対比較の結果、笑顔、怒り顔双方の条件で裏切り者の感情強度は協力者よりも高かった、笑顔：$t\,(31) = 2.29, p = .002$、怒り顔：$t\,(33) = 5.95, p < .001$。

今回の結果は、我々の予測を支持するものであった。すなわち、（1）怒り

表情について、裏切り者は、協力者よりも、信頼感が低く評定されること（怒りシグナルによる裏切り者検出の成功）、(2) 笑い表情では、この信頼感の差が消失すること（笑いによる怒りシグナルの打ち消し）、(3) 裏切り者は、協力者より、強い強度の笑顔を作ることという 3 つの結果が予測通り観察された。これらは、Okubo et al. (2012) の結果を再現するものである。裏切り者の定義を修正しても結果が再現されたことは、Okubo et al. (2012) の知見が高い頑健性を有することを示唆するものである。また、その高い再現性は研究結果の信頼性と仮説の妥当性を高めるものである。

謝辞

本稿は、平成 23 年度専修大学研究助成・個別研究「社会交換における表情認知に関する実験心理学的研究」の研究成果の一部である。

引用文献

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Frank, R. H. (2001). Cooperation through emotional commitment. *Evolution and the capacity for commitment*, *3*, 57–76.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407-420.

Mealey, L., Daood, C., & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology*, *17* (2), 119–128.

Oda, R. (1997). Biased face recognition in the prisoner's dilemma game. *Evolution and Human Behavior*, *18* (5), 309–315.

Okubo, M., Kobayashi, A., & Ishikawa, K. (2012). A fake smile thwarts cheater detection. *Journal of Nonverbal Behavior, 36*, 217-225.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA, 105,*

11087-11092.

Vanneste, S., Verplaetse, J., Van Hiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human behavior*, *28* (4), 272–276.

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: the sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28* (4), 260–271.

Wason, P. C. (1966). Reasoning. In Foss, B. M. *New horizons in psychology.* Harmondsworth: Penguin

Yamagishi, T., Tanida, S., Mashima, R., Shimoma, E., & Kanazawa, S. (2003). You can judge a book by its cover Evidence that cheaters may look different from cooperators. *Evolution and Human Behavior*, *24* (4), 290–301.