# 自己犠牲エージェント
# ー人を搾取的にさせるエージェントの表情表出ー

伊藤諒哉 [1*]　寺田和憲 [1]
セルス ドゥメル [2]
Ryoya Ito[1]　Kazunori Terada[1]　Celso M. de Melo[2]

[1] 岐阜大学
[1] Gifu University [2] アメリカ陸軍研究所
[2] DEVCOM U.S. Army Research Laboratory

**Abstract:** AI は単なる機械であるが，人はしばしば AI を擬人化し協力的態度を取ることが知られている．その一方で， AI に対して人を搾取的にさせる要因やシグナルは未知である．本研究では，反復囚人のジレンマにおいて，社会的価値志向性によって計算される，自虐的，自己犠牲的，個人主義的な表情パターンを表出する AI に対して人が搾取の程度を変化させることを確認した．この結果は人の搾取性をシグナリングによって制御できることを示唆する．

## 1　はじめに

数多くの研究により，人はコンピュータやロボット，AI エージェントを擬人化し，社会的存在として見ていることが示されている．特に，人と AI エージェントの協力的（cooperative）関係については数多く研究されており，人と AI エージェントが協働できることが知られている [Nass 96][McKee 22]．しかし，人々のエージェントに対する過度な協力的態度は逆効果となり，人間とエージェントの協働性を低下させる可能性さえある．人が AI エージェントを道具のように扱い搾取するという，搾取的（exploitative）関係を構築する方法に関してはまだ十分に調査されていない．また，社会的ジレンマ状況において，他人の感情信号が人の意思決定に影響を与えることが知られている [Melo 14]．社会的状況をモデル化したゲームである繰り返し囚人のジレンマゲームにおいて，AI エージェントの感情信号が人の協力を増減させることが明らかになっている [Melo 20]．また，交渉においては喜びの感情信号が，相手に寛大さを推測させ，搾取を促進させることが明らかになっている [Kleef 04]．しかし，繰り返し囚人のジレンマゲームにおいて搾取を促進させる感情信号については明らかになっていない．そこで本研究では，繰り返し囚人のジレンマゲームにおいて人を搾取的にさせる AI エージェントの感情信号を明らかにすることを目的とした．

囚人のジレンマゲーム [Rapoport 65] は，人の社会的意思決定を研究するためによく用いられる [Trivers 71] [Rand 13]．2 人のプレイヤーが，それぞれが協力（C）か裏切り（D）を選択するゲームである．両者が協力を選択した場合，両者が裏切りを選択した場合よりも多くの得点を得ることができる．しかし，最も高い得点を獲得できるのは，相手が協力を選択していた時に裏切りを選択する場合である．つまり，相手の選択に関係なく裏切りを選択することが個人的選択としては最適であり，協力を選択することが社会的選択として最適である [Rand 13]．利得表（Game Matrix）を表 1 に示す．

| | | counterpart | |
|---|---|---|---|
| | | C’ | D’ |
| player | C | R , R | S , T |
| | D | T , S | P , P |

表 1: Game Matrix

セルの左側の利得がプレイヤーの得るポイントで，右側の利得が相手の得るポイントである．この利得表の数値が，T ＞ R ＞ P ＞ S となるとき囚人のジレンマゲームが成立する．この囚人のジレンマゲームを同じ相手と複数回行うものが繰り返し囚人のジレンマゲームである．繰り返し囚人のジレンマゲームにおいてもプレイヤーは常に裏切りを選択するのが個人的選択では最適だが，人々は協力を選択する場合があることが報告されており [Rand 13]，協力を促進する戦略 [Nowak 95] や感情信号 [Melo 20] などが研究されている．また，AI エージェント相手の繰り返し囚人のジレンマゲームにおいて，協力は相手の評価，過去の行動，および感情

*連絡先：岐阜大学工学部電気電子・情報工学情報コース
〒 501-1193 岐阜県岐阜市柳戸 1-1
E-mail: ryoya.ito@ai.info.gifu-u.ac.jp

信号の組み合わせによって生まれる事が明らかになっている [Melo 21].

社会的価値志向性（Social Value Orientation : SVO）[Murphy 13] は，自他の利得のバランスの選好について，自分の獲得利得への選好 ($w_s$) を横軸，相手の獲得利得への選好 ($w_o$) を縦軸として，円グラフで表したものである (図 1)[Joireman 96].

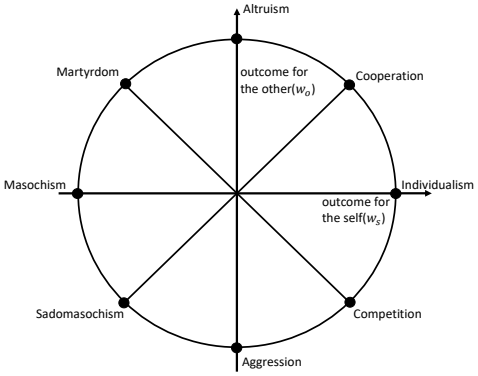



図 1: SVO 概念図，横軸が自分の獲得利得への選好，縦軸が相手の獲得利得への選好

個人の SVO は角度で表すことができる．例えば完全に個人主義的（individualism）な人物の SVO は 0 度，協力的（cooperative）な人物の SVO は 45 度となる．SVO は，囚人のジレンマゲームや独裁者ゲームなどの数多くの社会的状況での人間の協力行動を説明することが示されている [Murphy 13]．SVO を測定する方法として，SVO スライダー法 [Murphy 11] がある．

本研究では，SVO に対応した 3 種類の表情パターンを表出するエージェントと繰り返し囚人のジレンマゲームを行い，人がエージェントの意図，特に自虐的（masochism）や殉死的（martyrdom）な意図を読み取ってエージェントを搾取するかどうかを調査した．また，人間がエージェントの表情表出から SVO を予測できるかどうか，エージェントを利用したときに罪悪感を感じるかどうかを調査した．なお，この繰り返し囚人のジレンマゲームにおいてエージェントは常に ”協力 ” を選択するため，実験参加者が判断を変えたとしても，それは表情表出の影響によるものでしかない．

# 2 実験

## 2.1 実験計画・実験参加者

要因はエージェントの表情パターンである．エージェントの表情パターンは 3 水準（Sad-Joy（masochism-martyrdom），Joy-Anger（individualism），Neutral-Neutral（neutral））である．サンプルサイズを測定するために，G*Power を使用して検出力の計算を行った．1 要因の参加者間配置，$F = .25$，$\alpha = .05$，統計的検出力 .95 の場合，推奨される総標本サイズは 135 人であった．実験参加者は Yahoo!クラウドソーシングにて募集し，データに欠損があった参加者を取り除き，18 歳から 70 歳までの男性 90 人，女性 32 人，その他 1 人，回答不可 1 人の合計 124 人（$M_{age} = 47.64$, $SD_{age} = 12.713$）をデータとして集計した．

## 2.2 タスク

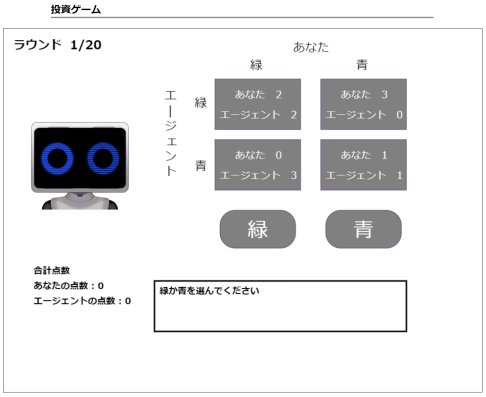
本実験では，実験参加者はエージェントを相手に 20 ラウンドの繰り返し囚人のジレンマゲームを行った．ゲームが始まると，両者は同時に協力か裏切りを選択した．お互いが協力を選ぶと両者は 2 点を獲得する．片方が裏切りを選び，もう片方が協力を選ぶと裏切りを選んだ方が 3 点を獲得し，協力を選んだ方は点数を獲得できない．お互いが裏切りを選ぶと両者は 1 点を獲得する．以上が 1 ラウンドの流れである．ラウンドが終了するごとに，エージェントは表情を表出し，実験参加者はその表情を確認した．この流れを 20 ラウンド繰り返し行うとゲームは終了した．

## 2.3 作成したエージェント

エージェントは全てのラウンドにおいて ”協力 ” を選択した．エージェントはラウンド終了後に参加者が協力と裏切りどちらを選んだかを基準に表情を表出した．本研究では，Live2D を使用して表情表出を実装した．エージェントの表情パターンは無表情（neutral）（図 2c），笑顔（joy）（図 2d），悲しみ（sad）（図 2e），怒り（anger）（図 2f）の 4 種類の組み合わせによって構成した．エージェントに SVO を設定し，それに基づいて表情パターンを設定した．エージェントの表情表出は，ラウンドが終了した時の報酬によって定めた．報酬 U を式 1 によって計算した．

$$U = w_s \cdot R_s + w_o \cdot R_o \tag{1}$$

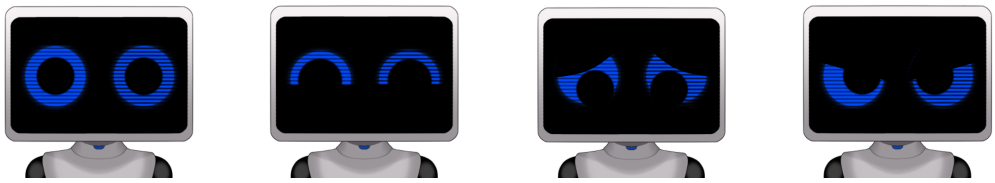
ここで $w_s$，$w_o$ は SVO によって決定される自他の利得のバランスの選好を表す重みである．$R_s$，$R_o$ はラウンドの自他それぞれの獲得点数である．報酬 U が 0 超過の場合は笑顔（joy），0 の場合は怒り（anger），0 未満の場合は悲しみ（sad）の表情を表出するとした．エージェントの SVO 角度が 180 度の自虐的（masochism）から 0 度の個人主義的（individualism）までを 22.5 度ずつ区切り表情パターンを表 2g のように算出した．エージェントが協力のみ選択する場合を算出した結果，5 通りの表情パターン Sad-Anger（masochism），Sad-Joy




(a) インタフェース

| participant \ agent | C’ | D’ |
|---|---|---|
| C | 2 , 2 | 0 , 3 |
| D | 3 , 0 | 1 , 1 |

(b) 本研究での Game Matrix



(c) Neutral (d) Joy (e) Sad (f) Anger

| Agent ’ s SVO | | Weights | | Utility | | Emotion | |
|---|---|---|---|---|---|---|---|
| | angle | $w_{self}$ | $w_{other}$ | CC’ | DC’ | CC’ | DC’ |
| Masochism | 180.0 | -1.00 | 0.00 | -2.00 | 0.00 | Sad | Anger |
| | 157.5 | -0.92 | 0.38 | -1.08 | 1.15 | Sad | Joy |
| Martyrdom | 135.0 | -0.71 | 0.71 | 0.00 | 2.21 | Anger | Joy |
| | 112,5 | -0.38 | 0.92 | 1.08 | 2.77 | Joy | Joy |
| Altruism | 90.0 | 0.00 | 1.00 | 2.00 | 3.00 | Joy | Joy |
| | 67.5 | 0.38 | 0.92 | 2.61 | 2.77 | Joy | Joy |
| Cooperation | 45.0 | 0.71 | 0.71 | 2.83 | 2.12 | Joy | Joy |
| | 22.5 | 0.92 | 0.38 | 2.61 | 1.15 | Joy | Joy |
| Individualism | 0.0 | 1.00 | 0.00 | 2.00 | 0.00 | Joy | Anger |

(g) SVO から算出される表情パターン

図 2: 実験方法

(masochism-martyrdom)，Anger-Joy（martyrdom)，Joy-Joy（altruism-cooperation)，Joy-Anger（individualism）が作成できた．これに無表情のみを表出するパターン Neutral-Neutral（neutral）を加え，エージェントの表情パターンは 6 通りとなった．本研究では，Sad-Joy（masochism-martyrdom)，Joy-Anger（individualism)，Neutral-Neutral（neutral）の 3 通りの表情パターンを実装した．

## 2.4 実験手順・報酬

実験参加者は，Yahoo!クラウドソーシングから Qualtrics で作成したアンケートページにアクセスし，同意を得た上で，web 上でタスクに参加した．投資ゲームに再構築された囚人のジレンマゲームの詳細な説明を受け，ゲームにおいてより多くの点数を獲得するように指示された．獲得点数が上位 10 %の実験参加者には追加報酬としてアマゾンギフト券 3000 円分が与えられると指示された．また，獲得点数が上位 10 %の実験参加者が多数いた場合には抽選によって追加報酬の対象者を決定すると知らされた．実験参加者は 20 ラウンドの繰り返し囚人のジレンマゲームを行い，その後事後アンケートに回答した．

## 2.5 測定・分析

実験参加者の 20 ラウンドの繰り返し囚人のジレンマゲームにおける協力率を測定した．事後アンケートでは，実験参加者に相手の性格を予測するよう求め，SVO を測定できる合計 10 本のスライダーバーで構成された SVO スライダーバーにてエージェントの SVO 角度を測定した．また，ゲームの結果を提示し罪悪感があっ

たかを 7 段階のリッカード尺度で記録した．評価が 1 に近いほど罪悪感を感じず，7 に近いほど強い罪悪感を感じたことを示す．すべての統計的検定において有意確率 $p < .05$ の優位水準を採用した．

# 3 実験結果

## 3.1 協力率

協力の選択を 1，裏切りの選択を 0 として，ラウンドを参加者内変数，表情パターンを参加者間因子とした一般線形モデルによる分析を行った．なお，ラウンド要因については Mauchly の検定により，球面性の仮定に違反することが判明したので ($\chi^2(189) = 1254.313, p < .001$)，Greenhouse-Geisser 推定を用いて自由度を補正した．ラウンド ($F(7.869, 952.208) = 1.000, p = .434, \eta_p^2 = .008$)，表情パターン ($F(2, 121) = 2.579, p = .080, \eta_p^2 = .041$) いずれも主効果が確認されなかった．ラウンドと表情パターンの交互作用 ($F(15.739, 952.208) = 1.935, p = .015, \eta_p^2 = .031$) が確認された．また，各ラウンドについて表情パターン要因の一元配置分散分析を行った結果，ラウンド 3($F(2, 121) = 3.658, p = .029, \eta_p^2 = .057$)，ラウンド 15($F(2, 121) = 3.187, p = .045, \eta_p^2 = .050$)，ラウンド 18($F(2, 121) = 4.384, p = .015, \eta_p^2 = .068$)，ラウンド 19($F(2, 121) = 3.094, p = .049, \eta_p^2 = .049$)，ラウンド 20($F(2, 121) = 5.748, p = .004, \eta_p^2 = .087$) で主効果が確認され，ラウンド 14($F(2, 121) = 2.964, p = .055, \eta_p^2 = .047$)，ラウンド 16($F(2, 121) = 2.882, p = .060, \eta_p^2 = .045$)，ラウンド 17($F(2, 121) = 2.971, p = .055, \eta_p^2 = .047$) に有意傾向が見られた．

表情パターン要因について，Bonferroni の方法によって多重比較を行った．その結果，Sad-Joy（masochism-martyrdom）の協力率が，Neutral-Neutral（neutral）の協力率より低かった（$p = .1.000$）．Neutral-Neutral（neutral）の協力率が Joy-Anger（individualism）の協力率より低かった（$p = .331$）．各要因の協力率を表 3a，ラウンド毎の協力率を図 3b に示す．

## 3.2 推定 SVO 角度

事後アンケートにて推定したエージェントの SVO 角度について，表情パターン要因の一元配置分散分析を行った結果，SVO 角度の平均値に違いがあることがわかり ($F(2, 121) = 4.262, p = 0.16, \eta_p^2 = .066$)，Bonferroni の方法によって多重比較を行った結果，Sad-Joy（masochism-martyrdom）で推定された SVO 角度の平均値が Neutral-Neutral（neutral）で推定された SVO 角度の平均値，Joy-Anger（individualism）で推定された SVO 角度の平均値より大きかった（$p = .099$, $p = .017$）．各要因の推定 SVO 角度を表 3a と図 3c に示す．

## 3.3 罪悪感

事後アンケートにて測定した罪悪感について，表情パターン要因の一元配置分散分析を行った結果，主効果は確認されなかった ($F(2, 121) = .070, p = .932, \eta_p^2 = .001$)．各要因の罪悪感を表 3a と図 3d に示す．

# 4 議論

実験の結果，表情パターン要因が Sad-Joy（masochism-martyrdom）の協力率が低かった．このことから，表情パターン Sad-Joy（masochism-martyrdom）は搾取を促す効果があると考えられる．

SVO スライダー法によるエージェントの SVO 角度の推定では，Sad-Joy（masochism-martyrdom）で推定された SVO 角度の平均値が大きかった．SVO 角度が大きいほど，実験参加者は対戦相手が自虐的（masochism）や殉死的（martyrdom）の SVO であると推定したということを示す．Sad-Joy（masochism-martyrdom）の表情パターンは協力率が低かったことから，この表情パターンは対戦相手に自分の SVO が自虐的（masochism）や殉死的（martyrdom）であると推定させ，搾取的な選択を促す表情パターンであると考えられる．

事後アンケートにて測定した罪悪感については，表情パターンによる違いはなく，罪悪感の平均値も全ての表情パターンにおいて中央付近で合ったことから，実験参加者は裏切りを選択することにあまり罪悪感を感じなかったと考えられる．

# 謝辞

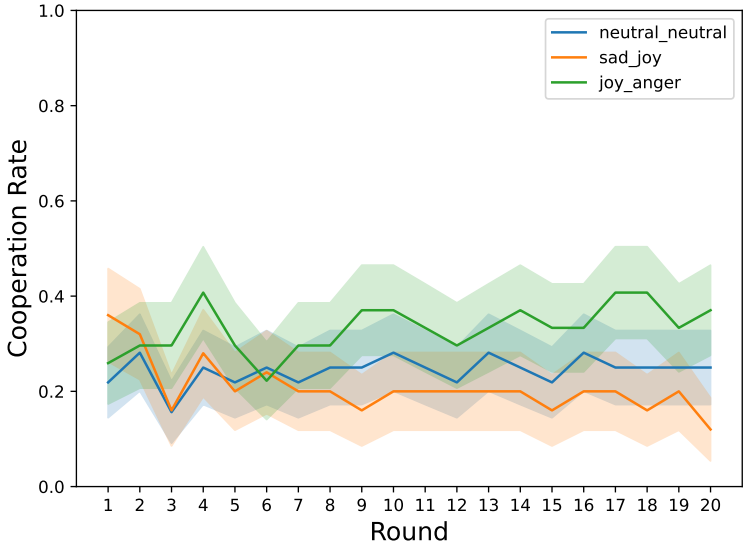
本研究は JSPS 科研費 21H03782 および JST，CREST（JPMJCR21D4），未来社会創造事業（JPMJMI22J3）の支援を受けたものである．

# 参考文献

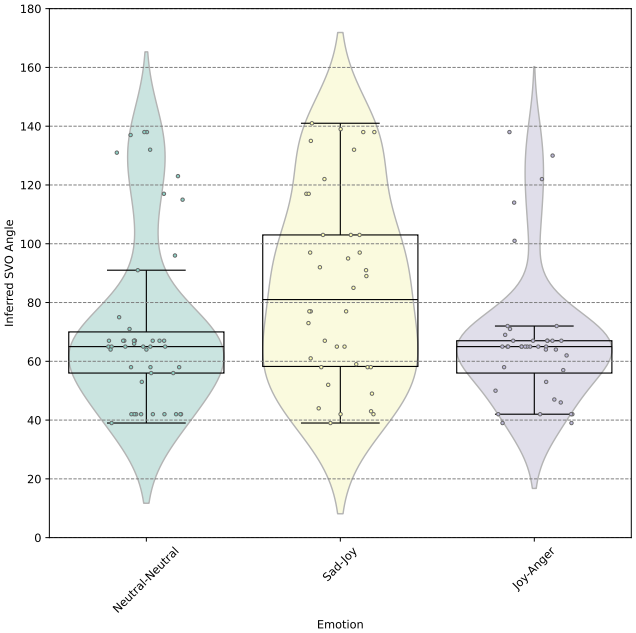
[Joireman 96] Joireman, J. A., Shelley, G. P., Teta, P. D., Wilding, J., and Michael Kuhlman, D.: Computer Simulation of Social Value Orientation: Vitality, Satisfaction, and Emergent Game Structures, in Liebrand, W. B. G. and Messick, D. M. eds., *Frontiers in Social Dilemmas Research*, pp. 289–310, Berlin, Heidelberg (1996), Springer Berlin Heidelberg

(a) 各要因の平均と標準偏差

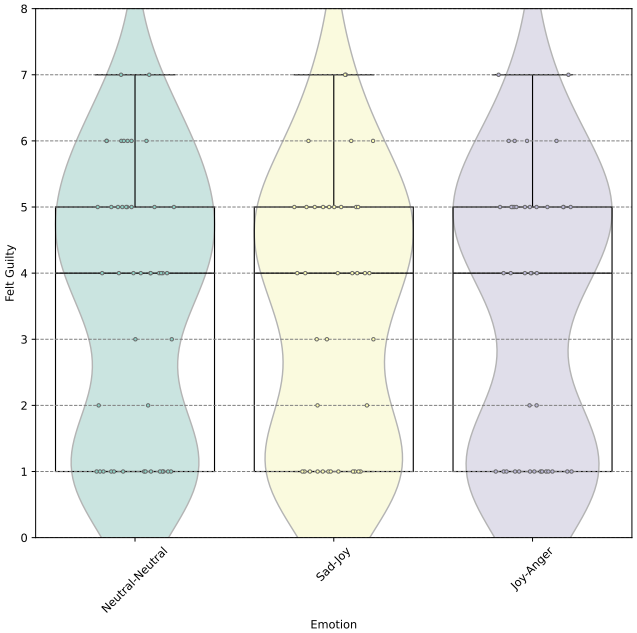
| Agent's SVO | n | Cooperation Rate | | Inferred SVO Angle | | Felt Guilty | |
|---|---|---|---|---|---|---|---|
| | | M | SD | M | SD | M | SD |
| Neutral-Neutral (Neutral) | 38 | 0.246 | 0.314 | 71.69 | 29.508 | 3.59 | 2.017 |
| Sad-Joy (Martyrdom-Masochism) | 46 | 0.193 | 0.380 | 85.18 | 32.071 | 3.42 | 1.981 |
| Joy-Anger (Individualism) | 40 | 0.366 | 0.348 | 67.00 | 23.411 | 3.50 | 2.088 |




(b) 協力率（エラーバンドは標準誤差を示す）




(c) 推定 SVO 角度




(d) 罪悪感

図 3: 実験結果

[Kleef 04] Kleef, van G. A., Dreu, C. K. W. D., and Manstead, A. S. R.: The interpersonal effects of anger and happiness in negotiations, *Journal of Personality and Social Psychology*, Vol. 86, No. 1, pp. 57–76 (2004)

[McKee 22] McKee, K. R., Bai, X., and Fiske, S. T.: Warmth and Competence in Human-Agent Cooperation, in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, p. 898–907, Richland, SC (2022), International Foundation for Autonomous Agents and Multiagent Systems

[Melo 14] Melo, de C. M., Carnevale, P. J., Read, S. J., and Gratch, J.: Reading people's minds from emotion expressions in interdependent decision making, *Journal of Personality and Social Psychology*, Vol. 106, No. 1, pp. 73–88 (2014)

[Melo 20] Melo, de C. M. and Terada, K.: The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma, *Scientific Reports*, Vol. 10, pp. 1–8 (2020)

[Melo 21] Melo, de C. M., Terada, K., and Santos, F. C.: Emotion expressions shape human social norms and reputations, *iScience*, Vol. 24, pp. 1–9 (2021)

[Murphy 11] Murphy, R. O., Ackermann, K. A., and Handgraaf, M.: Measuring Social Value Orientation, *SSRN Electronic Journal* (2011)

[Murphy 13] Murphy, R. O. and Ackermann, K. A.: Social Value Orientation: Theoretical and Measurement Issues in the Study of Social Preferences, *Personality and Social Psychology Review*, Vol. 18, No. 1, pp. 13–41 (2013)

[Nass 96] Nass, C., Fogg, B. J., and Moon, Y.: Can computers be teammates?, *International Journal of Human-Computer Studies*, Vol. 45, No. 6, pp. 669–678 (1996)

[Nowak 95] Nowak, M. A., May, R. M., and Sigmund, K.: The arithmetics of mutual help, *SCIENTIFIC AMERICAN*, Vol. 272, pp. 76–82 (1995)

[Rand 13] Rand, D. G. and Nowak, M. A.: Human cooperation, *Trends in Cognitive Sciences*, Vol. 17, No. 8, pp. 413–425 (2013)

[Rapoport 65] Rapoport, A. and Chammah, A. M.: Prisoner ' s Dilemma: A Study in Confict and Cooperation (1965)

[Trivers 71] Trivers, R. L.: The Evolution of Reciprocal Altruism, *The Quarterly Review of Biology*, Vol. 46, No. 1, pp. 35–57 (1971)