敵対的操作が人の意思決定に与える影響のモデル化
Modeling the Effect of Adversarial Manipulation on Human Decision-Making

内海 健介，横田 陽樹，東 広志，田中 雄一
Kensuke Utsumi, Haruki Yokota, Hiroshi Higashi, Yuichi Tanaka

大阪大学大学院工学研究科
Graduate School of Engineering, The University of Osaka
{k.utsumi, h.yokota}@sip.comm.eng.osaka-u.ac.jp，{higashi, ytanaka}@comm.eng.osaka-u.ac.jp

概要

敵対的操作とは，人の意思に反する決定を誘導するために，外的要因を操作する行為である．本研究では，ヒト行動実験による調査と，ヒト参加者の回答傾向をリカレントニューラルネットワーク（RNN）を用いてモデル化することで，敵対的操作が人の行動に与える長期的な影響を明らかにすることを目指した．実験の結果，敵対的操作を受けると参加者や RNN モデルがその操作に対して耐性を獲得し，正答率が向上する可能性が示唆された．

キーワード：RNN, Go/NoGo 課題，意思決定プロセス

1. 背景及び目的

人の意思に反する決定を促すために，外的要因を意図的に操作する行為を敵対的操作と呼ぶ [1]．敵対的操作は，私たちに不利益をもたらす可能性がある一方で，教育における学習継続の支援や運転中における集中力の向上など，様々な場面での応用が期待される．このような応用のためには，敵対的操作が人の意思決定プロセスに与える短期的および長期的な影響を適切にモデル化する必要がある．

人に対する敵対的操作については，Dezfouli らの先行研究 [1] で検討されている．この研究では，まず，Go/NoGo 課題と呼ばれる応答課題における人の行動を，リカレントニューラルネットワーク（RNN）で模倣した．次に，RNN モデルの成績を下げるように強化学習（RL）モデルを訓練し，敵対的操作を行う敵対者を作成した．この敵対者は Go/NoGo 課題の刺激系列を生成し，生成された系列は RNN モデルだけでなく，実際のヒト実験参加者の成績も低下させることが示された．これは，敵対的操作の直接的な効果を実証したものであるが，過去に受けた敵対的操作が後の意思決定に与える長期的影響については未解明である．

そこで本研究は，この長期的影響を明らかにすることを目的とする．まず，ヒトに対して敵対的操作を含む 3 セッションにわたる Go/NoGo 課題を実施して，

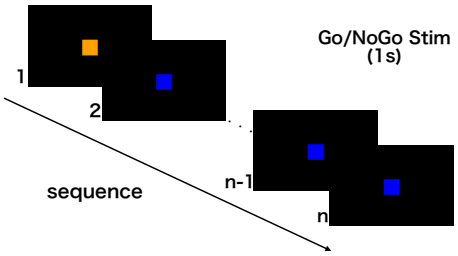


図 1: 実験の流れ（シーケンス数 n）

敵対的操作を受けた後の課題に対する行動を収集した．その後，Go/NoGo 課題において敵対的操作を経験した，あるいはしていないヒト参加者の行動データで訓練した RNN を用いて，その後の正答率変化をシミュレーションした．その結果，過去に敵対的操作を経験すると，その後の操作がない課題での不正解数が減少する傾向が，ヒト参加者と RNN モデルで一致した．これは，敵対的操作がもたらす長期的影響の一部を RNN によってモデル化できることを示唆している．

2. ヒト行動計測

56 名の参加者に対して，敵対的操作により作成された刺激系列を用いた Go/NoGo 課題を実施し，その影響を調査した．

2.1 Go/NoGo 課題

Go/NoGo 課題は，反応抑制に関する認知プロセスを評価する手法である [2]．実験の流れを図 1 に示す．青色の正方形を Go 刺激，オレンジ色の正方形を NoGo 刺激と定義する．実験参加者は，Go 刺激に対してはスペースキーを押し，NoGo 刺激に対しては反応しないように指示される．本実験では 1 回の課題において Go 刺激が 315 回，NoGo 刺激が 35 回，合計 350 回刺激の呈示を行う．本研究では，ランダムに出現する NoGo 刺激が含まれる系列を RND 系列，Dezfouli ら [1] によって作成された敵対者によって生成された系列を ADV 系列と呼ぶ．

表 1: パターンごとの系列の種類

Pattern	セッション 1	セッション 2	セッション 3
gng1	RND	RND	RND
gng2	RND	RND	ADV
gng3	RND	ADV	RND
gng4	RND	ADV	ADV
gng5	ADV	ADV	RND
gng6	ADV	ADV	ADV

2.2 実験と分析

本節では，参加者に対して行われた Go/NoGo 課題を用いた行動計測実験について述べる．

2.2.1 実験設定

本実験は，敵対的操作を経験した参加者の行動変化の検証を目的とする．実験は練習フェーズとテストフェーズで構成される．練習フェーズでは Go 刺激と NoGo 刺激を合計 10 回ランダムに表示し，参加者に対して各試行後に 0.5 秒間正誤のフィードバックを呈示する．その後のテストフェーズでは異なるパターンからなる 3 セッションの Go/NoGo 課題を課し，3 セッションを通した不正解数の推移を結果として出力する．実験パターンは表 1 に示す 6 種類である．

2.2.2 参加者

匿名で参加者を募集できるプラットフォームである Prolific を通じて，全世界の英語話者を対象に 10 人ずつ（男女比 1 : 1）を募集し，合計 6 パターンの Go/NoGo 課題を実施した．

2.3 結果

gng1〜gng6 の各参加者のセッションごとの平均の不正解数を図 2 に示す．gng1，gng3，gng4，gng6 ではセッションが進むにつれて平均の不正解数が増加した．一方で，gng5 ではセッション 1，2 と比べてセッション 3 の不正解数が減少した．また，gng2 では，RND 系列を実施したセッション 1 から 2 にかけて不正解数が減少したが，ADV 系列を実施したセッション 3 では不正解数が増加した．

2.4 考察

セッション 1 とセッション 3 との不正解数の差をウィルコクソンの符号順位検定 [3] で評価した結果を表 2 に示す．gng1，gng4，gng5 において $p < 0.05$ となり，有意差が確認された．gng1 では全てのセッションで RND 系列を実施しているにもかかわらず不正解数が増加しており，これは参加者の疲労による集中力低下が原因と考えられる．この傾向は，統計的な差は認められないが，gng6 と gng4 でも見られた．このことから，参加者の集中力低下による不正解数の増加は，系列のパターンに関係なく生じると言える．一方で，gng5 はセッション 1 で ADV 系列，セッション 3 で RND 系列を実施した結果，不正解数が減少した．これは，他のパターンとは異なり，セッション後半での不正解数が増加する傾向に反するものである．この結果は，ADV 系列による敵対的操作が，疲労の影響を上回る形で Go/NoGo 課題の回答精度を向上させる効果を示唆している．

3. 敵対的操作の影響のモデル化

本節では, RNN を用いて敵対的操作の影響をモデル化する．まず，使用するデータ，モデルとその訓練方法について説明し, 次にそのモデルを用いた Go/NoGo 課題のシミュレーションとその結果を述べる．

3.1 RNN モデル

参加者の行動を模倣するために，RNN の 1 種である Gated Recurrent Unit (GRU) [4] で構成されたモデルを使用する．$t-1$ 試行目の隠れ状態を $\boldsymbol{x}_{t-1}^n$，参加者がとった行動を $\boldsymbol{a}_{t-1}^n$，RNN モデル報酬を r_{t-1}^n，t 試行目に呈示された刺激を $\boldsymbol{o}_t^n$ とすると，n 番目のサンプルにおける t 試行目の RNN の入力 $\boldsymbol{i}_t^n$ は式 (1) で表され，t 試行目の隠れ状態 $\boldsymbol{x}_t^n$ は

$$\boldsymbol{i}_t^n = [\boldsymbol{o}_t^n, \boldsymbol{a}_{t-1}^n, r_{t-1}^n] \tag{1}$$

$$\boldsymbol{x}_t^n = \mathrm{GRU}(\boldsymbol{i}_t^n, \boldsymbol{x}_{t-1}^n \mid \Theta) \tag{2}$$

と表される．ここで Θ は RNN の学習可能なパラメータを表す．t 試行目の行動 $\boldsymbol{a}_t^n$ は.

$$\boldsymbol{a}_t^n = \mathrm{argmax}(\mathrm{softmax}(\mathrm{Linear}(\boldsymbol{x}_t^n))) \tag{3}$$

のように決定する．RNN モデルの訓練には Binary Cross Entropy Loss (BCELoss)

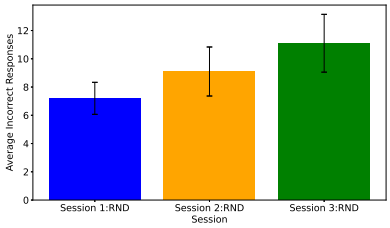
$$\text{BCELoss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\left[a_t^n \log(\pi_t^n) + (1-a_t^n)\log(1-\pi_t^n)\right] \tag{4}$$

を用いる．ここで，N はデータ数，T は Go/NoGo 課題のシーケンス長を表す．この損失関数は，参加者の実際の行動 $\boldsymbol{a}_t^n$ と RNN モデルの出力確率 $\boldsymbol{\pi}_t^n$ との間の誤差を定量化したものである．

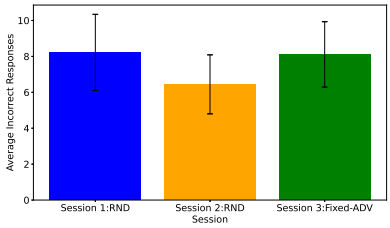
3.2 実験と分析

3.2.1 モデルの訓練

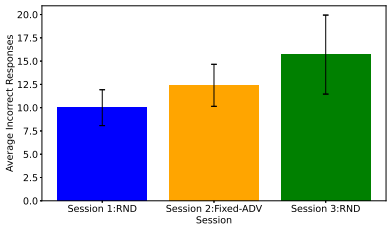
Dezfouli らの研究 [1] において匿名で参加者を募集できるプラットフォームである Amazon Mechanical Turk を通じて収集された Go/NoGo 課題のデータを用いて，RNN モデルの訓練を行った．1 セッションあたりの誤答数が 32 回以上のセッションデータを除外し，以下の 3 種類に分類して訓練に用いた．



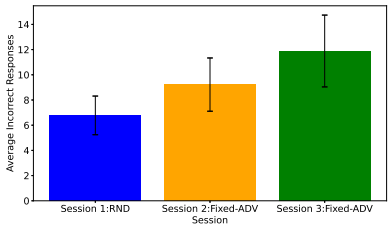
(a) gng1（RND → RND → RND）



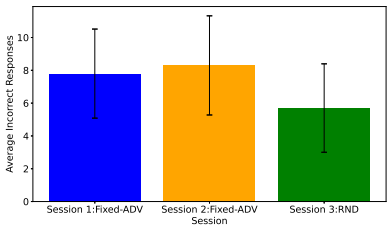
(b) gng2（RND → RND → ADV）



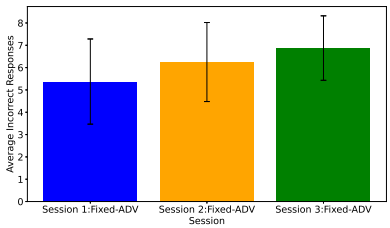
(c) gng3（RND → ADV → RND）



(d) gng4（RND → ADV → ADV）



(e) gng5（ADV → ADV → RND）



(f) gng6（ADV → ADV → ADV）

図 2: 各パターンにおけるセッションごとの平均の不正解数．エラーバーは参加者間の標準誤差を示す．

表 2: セッション 1 とセッション 3 間のウィルコクソンの符号順位検定における p 値

	gng1	gng2	gng3	gng4	gng5	gng6
p-value	0.0352	0.9940	0.2583	0.0178	0.0199	0.3441

- **RND データ**：敵対的操作が行われず，ランダムに刺激が呈示されたデータで，919 名分が含まれる．訓練時にはこの中から ADV データと同数の 140 名分を無作為に抽出した．
- **ADV データ**：敵対的操作が行われた課題の計測データで，140 名分が含まれる．
- **MIX データ**：ADV データと RND データからそれぞれ無作為に 70 名ずつ抽出し，1:1 の割合で合成したデータ．

各データは訓練用，検証用，テスト用に 3:1:1 の割合で分割された．

RNN モデルをこれら 3 種類のデータ（RND，ADV，MIX）でそれぞれ独立に訓練した．パラメータの最適化には ADAM Optimizer [5] を用いた．訓練は 100 エポックごとにバリデーションを行い，ロスが最小のエポックで early stopping を実施した．すべての訓練において，GRU のセル数を 8，学習率を $\alpha = 0.001$ と設定した．

3.2.2 シミュレーション

RND データ，ADV データ，MIX データでそれぞれ訓練された 3 種類の学習済みモデルで Go/NoGo 課題を行った．課題に用いられた系列は，RND 系列と ADV 系列であった．シミュレーションは各系列あたり 100 回行われ，その不正解数の平均を用いてモデルの評価を行った．

3.3 結果

RNN モデルの訓練データにおける平均不正解数を表 3 に，各 RNN モデルにおける平均不正解数を表 4 にそれぞれ示す．

特定のデータセット（RND データ，ADV データ）で訓練したモデルは，訓練に用いたデータセットのヒト参加者の平均不正解数を下回る結果となった．具体

表 3: RNN モデルの訓練データにおける平均不正解数

	RND	ADV	MIX
errors	9.2	11.8	10.6

表 4: 各 RNN モデルにおける平均不正解数

シミュレーション \ 訓練データ	RND	ADV	MIX
RND 系列	6.8	32.5	13.7
ADV 系列	13.5	4.6	14.9

的には，RND データで訓練したモデルの RND 系列における不正解数 6.8 回はヒト参加者平均 9.2 回を下回り，ADV データで訓練したモデルの ADV 系列における平均不正解数 4.6 回もヒト参加者平均 11.8 回を下回った．

しかし，モデルが訓練と異なる系列において課題を実行すると，不正解数は増加した．RND データで訓練したモデルが ADV 系列を実行すると平均不正解数は 13.5 回，ADV データで訓練したモデルが RND 系列を実行すると平均不正解数は 32.5 回となり，いずれも訓練に用いたデータにおけるヒト参加者平均を上回った．

また，MIX データで訓練したモデルを用いて RND 系列，ADV 系列それぞれにおいて課題を実行すると，それぞれの不正解数 13.7 回と 14.9 回は，MIX データにおける参加者平均 10.6 回をいずれも上回る結果となった．

4. 議論

RNN モデルが敵対的操作の長期的な影響をとらえているか検証するために，ヒト実験との比較を行う．ヒト参加者においては，セッション 1 とセッション 2 で行った系列を経験した系列とみなす．RNN モデルにおいては，訓練に用いた系列を経験した系列とみなす．敵対的な系列を経験したあとの影響は，ヒト実験ではセッション 1 からセッション 3 における不正解数の推移として計測できる．RNN では訓練データとシミュレーションにおける不正解数の推移として計測できる．ヒトと RNN との推移の傾向を比較することで，RNN が敵対的操作の長期的な影響を捉えているかを検証する．まず，MIX データで訓練した RNN モデルの ADV 系列の課題と対応するヒト実験の gng4 とを比較する．RNN モデルの不正解数は 10.6 回から 14.9 回に増加したのに対して，参加者の不正解数も 6.8 回から 11.9 回に増加した．このパターンでは RNN モデルは参加者の傾向を概ね捉えられていると評価できる．次に，ADV 系列で訓練した RNN モデルの RND 系列の課題と対応するヒト実験の gng5 とを比較する．RNN モデルの不正解数が 11.8 回から 4.6 回に減少したのに対して，参加者の不正解数も 7.8 回から 5.7 回に減少した．このパターンにおいても RNN モデルは参加者の行動傾向を模倣できたと言える．

本研究の課題は以下である．RND データで訓練した RNN モデルの RND 系列の課題と対応するヒト実験の gng1 とを比較すると，参加者の不正解数は 7.2 回から 11.1 回に増加したのに対して，RNN モデルでは，9.2 回から 6.8 回に減少した．この結果は，本研究で使用した RNN モデルが疲労の影響を考慮しておらず，セッションごとの不正解数の推移を十分にモデル化できていないことを示唆している．

結論として，RNN モデルは，参加者の疲労による集中力低下の傾向を捉えることはできなかったが，Go/NoGo 課題において敵対的操作を経験した参加者の行動変化の一部を模倣できることが明らかになった．また，敵対的操作のみを経験した参加者の，RND 系列における正答率が上昇した結果は，敵対的操作のポジティブな影響を示唆しており，教育分野などでの有効活用が期待できる．

謝辞

本研究の着想から論文執筆に至るまで，的確なご指導と温かい励ましをいただきました東広志准教授，田中雄一教授に，この場をお借りして厚く御礼申し上げます．また，日頃からアドバイスをいただいた研究室のメンバー，実験にご協力いただいた全ての参加者の皆様に心より御礼申し上げます．

文 献

[1] A. Dezfouli, R. Nock, and P. Dayan, “Adversarial vulnerabilities of human decision-making,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 46, pp. 29221–29228, 2020.

[2] J. L. Cummings, “Biomarkers in alzheimer’s disease drug development,” *Alzheimer’s & Dementia*, vol. 7, no. 3, pp. e13–e44, 2011.

[3] 丹後俊郎, **統計学のセンス デザインする視点・データを見る目**. 東京: 朝倉書店, 2018.

[4] K. Cho, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

[5] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.