行動介入シミュレータ：
社会的ジレンマの解消支援ツールとして

工藤泰幸
(株)日立製作所 研究開発グループ

1. はじめに

個々人の利己的な行動の結果が社会全体に不利益をもたらす現象は，古くは共有財の枯渇を招くコモンズの悲劇 [1] や，共有財を無対価で享受するフリーライダー問題 [2] として提起され，現在では社会的ジレンマとして概念化されている．社会的ジレンマは，協力と非協力が選べる状況下において，個人としては他人の行動によらず非協力を選ぶ方が高い利得が得られる一方，全員が非協力を選ぶと全員が協力を選ぶよりも利得が低くなる状況を指す [3]．身近な例では，ごみのポイ捨てや放置自転車，広域的には気候変動やエネルギ不足，パンデミックの蔓延などが挙げられる．このように，社会的ジレンマは，多くの社会問題の発生メカニズムになり得ることが知られており，これを解消するためには，集団利益の獲得に向けた各人の協力促進，つまり行動変容が必要となる．

協力行動を促進するメカニズムの解明に向けた研究は，社会心理学や経済学を中心に発展してきた．その結果，協力行動を規定する要因として，インセンティブ構造，集団人数，コミュニケーション，サンクションなどの環境要因や，国籍，年齢，パーソナリティなどの個人・集団要因など，多くの変数が同定されている [4]．近年では，これらの知見に基づくフィールド実証も進展している [5] [6]．さらには，認知バイアスを利用した行動変容手法であるナッジに着目し [7]，公共政策としてこれを活用する取り組みが各国で進められている [8]．

一方，社会的ジレンマにおける協力行動の規定要因は，その種類の多さと共に，要因間の非線形な交互作用も報告されている [9] [10]．これは，ジレンマの発生現場ごとに協力行動の支配的な要因が異なることを意味し，効果的な介入が現場によって異なることを示唆している．このため，最適な介入施策を見極めるには，現場の特徴を踏まえた高度な分析が必要である．このことが，社会的ジレンマを解決したい介入実践者の意思決定を困難にしていると考えられる．したがって，分析の支援に対するニーズは極めて高いと予想され，例えば，現場で試してみたい施策の介入効果を事前に予測できるようなシミュレータの存在は有用であろう．本研究では，このような現場の意思決定を支援するツールの提供をゴールとしており，複雑な事象の予測に長ける機械学習モデルをツールの分析エンジンとすることで，その実現をめざす．

2. 方法

2.1 コンセプト

機械学習モデルを用いて高度な意思決定を支援する類似のアプローチとして，医療現場における治療支援が挙げられる [11]．このシステムにおいては，患者の情報を入力すると，その患者の病状や，適切な治療方針を出力することが可能である．出力結果は主に医師に提示され，この情報が医師の治療方針に関わる意思決定を支援する．本研究においても，これと同様のアプローチを実践したい．具体的には，患者を社会ジレンマの発生現場に置き換え，治療方針を介入施策に，医師を介入実践者と想定することで，介入実践者の意思決定を支援する仕組みである．

上記アプローチの実現にあたり，最も重要な要件は訓練データの確保である．医療分野では臨床データの十分な蓄積があり，これを機械学習モデルの訓練データとして活用することができる．言い換えれば，その環境がゆえに，機械学習モデルの実用化が進んだともいえる．一方，社会的ジレンマの現場においては，介入効果などの結果に対する蓄積が極めて少なく，共有化もされていない．それどころか，どのような変数を蓄積すべきか，その変数は測定可能か，他の現場でも使えるかなど，基礎的な事項から検討する必要がある．したがって，医療現場と同様のアプローチを試みることは容易ではなく，現場データの代替えとなるようなデータセットが必要である．

そこで本研究では，社会的ジレンマの状況をゲーム課題の形式で模擬し，その中で興味ある協力行動の要因を操作し，それによる行動の変化をアウトカムとして測定している数多くのラボ実験に着目した．ラボ実験に関する文献を網羅的に収集して符号化し，統一的なデータセットを構築できるのであれば，先に示した基礎的な検討課題を全て解決することができる．このことから，現場データの代替えとして，ラボ実験のデータを活用するコンセプトを策定し，以下の検討を進めた．

2.2 訓練データの作成

ある研究分野の文献を収集して符号化する作業は，メタアナリシスにおけるコーディングと呼ばれる工程に相当する．近年，この作業を集約的に行ってデータベースを構築し，その資産を利用してオンライン・メタアナリシスを実施するプラットフォームが台頭している [12] [13]．社会的ジレンマについても，ラボ実験の論文のコーディング・データに

基づく，オンデマンド・メタアナリシス向けプラットフォームが公開されている [14]．このプラットフォームは，Cooperation Databank (CoDa) プロジェクトが主導しており，構築されたデータベースには，約 2,000 論文から抽出された 1 万件以上の実験結果 (効果量) が，独立変数や統制変数と共に収録されている．したがって，このデータベースを活用することにより，多大な労力を要するコーディング作業が不要となる．したがって，本研究のコンセプトの実現が大幅に加速することが期待できる．

予備検討の結果，機械学習用の訓練データとして適用するには，データフォーマットに関する幾つかの変換を実施すれば良いことが分かった．具体的には，各種変数の合成，複数の実験条件が混在している場合の重み付け，欠損値の補間などである．図 1 は，一連のデータ変換処理を示したフローチャートであり，これらの処理を実施することで訓練データの作成を可能にした．

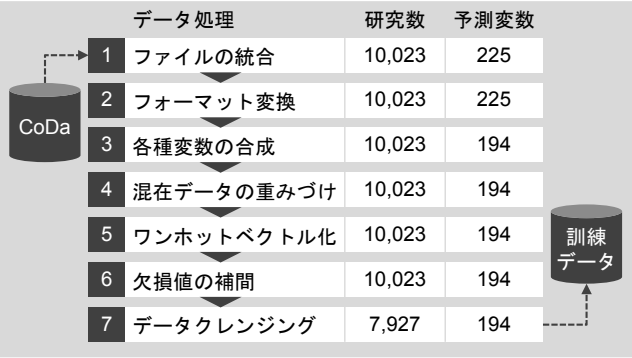

図 1　データ変換プロセス

生成された訓練データは，194 個の予測変数と 1 個のアウトカム (集団成員の協力率の平均値，以下，全体協力率と呼ぶ) を単位とする，7,984 個の実験結果によって構成された．訓練データに含まれる実験実施国は 48 カ国，1 実験あたりの平均サンプルサイズは 70.7 人 (延べ 56.4 万人) であった．また，アウトカムである全体協力率には，0%から 100%までの測定値が含まれており ($M = 49.5\%$，$SD = 19.0\%$)，実験のセッティングや介入内容に応じて，全体協力率が大きく変動していることが分かった．

2.3 機械学習モデリング

訓練データの機械学習においては，訓練データ全般に対し，高い予測精度を確保することが重要である．そこで，代表的な 3 種類の機械学習モデル (ニューラルネットワーク [15]，ランダムフォレスト [16]，勾配ブースティング [17]) を線形に組み合わせたアンサンブル学習を適用した [18]．また，機械学習モデルの特性により発生しやすい過学習については，各モデルのハイパーパラメータを決める際に，クロスバリデーション法を用いて未知データに対する予測精度を評価すると共に，パラメータ探索を早期に終了させるアーリ・ストッピング法 [19] の適用により抑制した．

なお，本研究における機械学習は，メタアナリシスの文脈においては，混合効果モデルを用いたメタ回帰分析に相当する [20]．そこで，メタアナリシスと同様に，観測された全体協力率の分散とサンプルサイズに基づいて重みを算出し，この重みを学習時に設定することで，各研究の信頼度を反映させた．予測精度の評価指標については，混合効果モデルであることを加味し，研究間と実験間の異質性分散aとモデルの予測残差との比によって示される疑似決定係数 R_*^2 とした．疑似決定係数R_*^2は，式 1 に示すに計算式を用いて算出した [21]．

$$R_*^2 = 1 - \frac{\tau^2_{\text{unexplained}}}{\tau^2_{\text{total}}} = 1 - \frac{\tau^2_{s\text{M}} + \tau^2_{t\text{M}}}{\tau^2_{s\text{R}} + \tau^2_{t\text{R}}} \qquad (1)$$

ここで，$\tau^2_{s\text{R}}$ と $\tau^2_{t\text{R}}$ は，全体協力率の測定データが有する研究間と実験間の異質性分散であり，$\tau^2_{s\text{M}}$ と $\tau^2_{t\text{M}}$ は観測データと予測データの残差が有する研究間と実験間の異質性分散である．なお，上述の 4 つの分散値は，制限付き最尤推定法 (REML) を用いて算出した．

2.4 予測性能

次に，構築した機械学習モデルに対して，予測性能を評価した結果を図 2 に示す．左図は既に学習された訓練データに対して予測を行った場合のフィッティング性能，右図は未知データに対して予測を行った場合の汎化性能である．汎化性能の評価にあたっては，10 分割クロスバリデーション法を適用した．この方法は，全体の 90%の訓練データを機械学習し，残りの 10%の未知データに対しての予測値を観測値と比較する．この動作を，10%の未知データを入れ替えつつ 10 回繰り返すことで，実質的に全ての訓練データを未知データとして評価した．

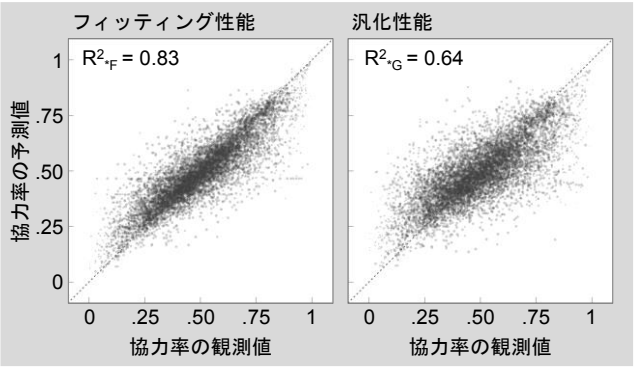

図 2　機械学習モデルの予測性能

a 1 つの研究に複数の実験条件 (統制群と介入群など) が含まれている，いわゆる入れ子構造を想定している．

双方の予測性能において，疑似決定係数の値はそれぞれ $R^2_{*F}=0.83$，$R^2_{*G}=0.64$ であったb．これは，研究間と実験間の異質性分散，つまり異なるセッティングや介入内容の違いに起因する全体協力率の分散の 83%および 64%が機械学習モデルによって説明できていることを意味する．この結果は，社会科学系のデータに基づく予測において実用的であると考えられる．さらに，既学習データと未知データに対する予測精度 (疑似決定係数 R^2_*) の平均差は 0.19 であり，過学習も抑制されていることが示唆された．なお，図 2 から分かる通り，各プロットの分布において，系統誤差と思われるような偏りは確認されなかった．このことは，全体協力率を予測するにあたり，全体協力率に影響する重要な予測変数の見落としが無いことを示唆している．

3. 行動介入シミュレータの開発

次に，構築した機械学習モデルをより実用的な形態として普及させるため，本機械学習モデルを分析エンジンに持つ行動介入シミュレータを開発した．シミュレータの入力パラメータは，前述した 194 個の予測変数に相当する．現場の状況や介入内容，個人・集団特性に応じたパラメータの値を入力することで，シミュレータは現状の全体協力率を予測する．この結果をベースラインとし，任意のパラメータの設定値を変えることにより，介入効果 (つまり，ベースラインとの全体協力率の差分) をシミュレートする．

3.1 エディタ機能

前述した通り，機械学習モデルの訓練データは，社会的ジレンマを模倣したラボ実験 (ゲーム課題) の結果である．このため，環境パラメータの中には，ゲーム課題の「ルール」が多く含まれる．これらのルールを現実世界の状況に置き換えるためには，一定の予備知識を必要とする．そこで，パラメータの設定を容易化するためのエディタ機能を用意することで，この問題に対処した．図 3 は，開発した行動介入シミュレータのエディタ画面である．

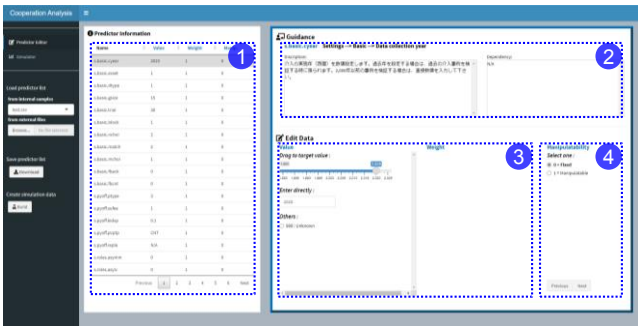

図 3　行動介入シミュレータのエディタ画面

エディタ画面では，パラメータリスト (図 3 の領域①) の中から設定したいパラメータを選択すると，これを現場のパラメータとして使用するための設定ガイドラインが提示される(領域②)．同時に，直観的な入力操作を支援するスライダやラジオボタンを設けることで (領域③)，ユーザが容易に設定値を編集できるようにした．さらに，設定したパラメータの値を変更した時の介入効果を知りたい場合は，変更後の設定値，あるいはその範囲を指定するためのエリアを設けた (領域④)．なお，現場によっては，特定のパラメータが観測困難または非該当となる可能性がある．そこで，各パラータには「不明・非該当」の項目を設け，それらが選択できる仕組みとした．

3.2 ダッシュボード機能

上記したように，パラメータ設定値を「変更する」設定により，複数の介入候補が生成され，各候補のベースラインとの差分に基づく個々の介入効果が算出される．この結果を効率的に視覚化するためのダッシュボードを開発した．図 4 は，開発したダッシュボードの画面である．

図 4 に示すように，本ダッシュボードでは，各介入施策を一つのカードに見立てている (領域⑤)．まず，ベースライン (グレーのカード) が左端に配置され，その右側から介入効果の高い施策が順番にソートされている．任意のカードを選択すると，それに連動して，現場の特徴を 6 次元で表現したレーダチャート (領域⑥)，全体協力率の経時変動予測 (領域⑦), パラメータの変更箇所 (領域⑧) に関する情報が，カードと同じ色調で表示される．図 5 の例では，領域⑤において，最も介入効果の高い「58.3%」のブルーのカードが選択された状態を示している．比較対象として，ベースラインである「44.2%」のカードの情報が，グレーの色調で同時に表示されるようにした．

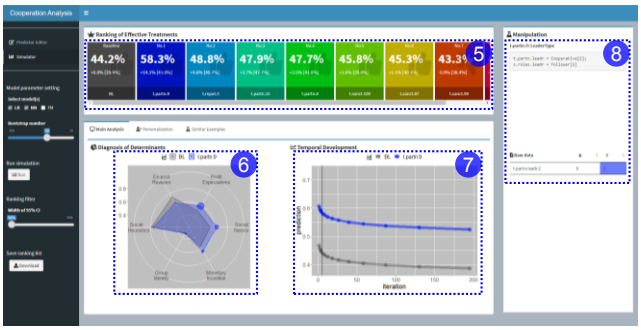

図 4　行動介入シミュレータのダッシュボード画面

次に，本シミュレータが複雑な予測機能を実装できているかを確認するため，2 種類のパーソナリティ (日和見タイプ，抜け駆けタイプ) を想定し，オフィスでの節電を想定したシミュレーションを実施した．その結果，用意した介入施

b R^2_*の後ろに記した F はフィッティング (Fitting) 性能，G は汎化 (Generalization) 性能を意味する．

策において，日和見タイプのグループには「効果と感謝の伝達」が最も効果的であり(全体協力率 46.7% → 67.5%)，抜け駆けタイプには「ポイントの付与」が最も効果的 (全体協力率 22.5% → 39.7%) との予測結果が得られた．以上のことから，有効な介入施策の順位がパーソナリティに連動して変化し，リコメンドされた介入施策についても，主観的に納得度が高い内容であることが分かった．以上の結果から，構築された機械学習モデルが，序論で述べたパラメータ間の非線形な交互作用をモデリングできていることが示唆された．

領域⑥にレーダチャートを用いた理由は，社会的ジレンマの発生現場の状態を把握するために，協力の阻害要因を可視化することが有効と考えたためである．この際，可視化の観点から，阻害要因は数個レベルであることが望ましい．そこで，訓練データから機械的に阻害要因を求める方法として，194 パラメータの主成分を求めつつ，各主成分の線形結合とアウトカムである全体協力率との誤差を最小化する，部分的最小二乗回帰モデルを適用した．訓練データを用いて本モデルを構築した結果，図 5 に示すように，全体協力率の予測誤差は主成分の数の増加とともに減少し，10 個でほぼ収束することが分かった．この結果を踏まえ，レーダチャートの評価軸には，主成分の意味の解釈が比較的容易であった上位 6 個の主成分を採用した．解釈された各主成分の意味は，主成分負荷量の大きい順に，協力への期待度，社会的規範，協力へのインセンティブ，コスト便益，集団へのコミットメント，構造的な動機付けであった．

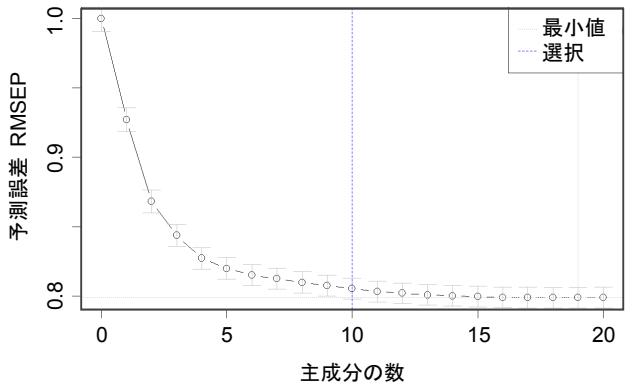

図 5　主成分の数と予測誤差との関係

領域⑦の経時変化の予測機能については，同じ状況または介入を繰り返し実施した時の，各回次の全体協力率を測定したラボ実験が多数報告されていることに着想を得た．これらの結果が訓練データに含まれており，回次を設定するパラメータも存在することを利用し，本研究では，回次の設定値を変更することで経時変化の予測機能を実現した．したがって，領域⑦の折れ線グラフの横軸については，時間の絶対値ではないことに注意する必要がある．

3.3　シミュレ―タの活用方法

行動介入シミュレータを活用するための条件として，まず，対象とする現場とその現場での協力行動が予め定まっている必要がある．例えば，「オフィスから退席する際に，こまめに消灯することを協力行動と定める」などである．その理由は，シミュレータのパラメータ値を設定するための前提として，現場と目標行動の明瞭さが必要となるためである．

次に，シミュレータを最も効果的に活用できるシーンとして，どのような介入施策が有効であるかの「方針決め」の段階を想定している．つまり，介入検討における早期フェーズである．その理由は，本シミュレ―タのパラメータには，介入内容そのものというよりも，アプローチに近いものが多く含まれるためである．また，パラメータに含まれない現場固有の制約等には対応できないことも理由の一つである．したがって，ユーザは，シミュレーション結果に基づき介入方針を決定した後，現場において実現可能な，具体的な介入施策の実装形態を検討する必要がある．

なお，人が行動を変える場合，「無関心期」「関心期」「準備期」「実行期」「維持期」の 5 つのステージを通過すると考えられている [22]．この中で，本シミュレータが出力する介入施策は，準備期以降のステージの該当者に適合すると考えられる．なぜなら，訓練データとなったラボ実験では，参加者がゲームのルール (すなわち，協力と非協力具体的な行動や，それによる利得構造) を理解していることが前提になっているからである．中でも「その行動をすることが良いと分かっているが，つい短絡的な利益を追ってしまう」という行動が現れがちな「実行期」や「維持時」の該当者に対し，より効果的であることが期待できる．

また，個人の意思の種類として，例えば「態度」「価値観」「信条」があるとし，この順番に意思が強固であると仮定するならば，介入施策によって変容が許容されるのは，態度レベルの範囲までと考えられる．特に，信条を操作する行為である洗脳については，介入の計画段階でそのリスクを慎重に検討すべきである．本開発のシミュレータは，ラボ実験で行われた比較的「ソフトな」介入施策を出力する．よって，介入によって意思が変容されるとしても，それは態度レベルの範囲に収まると考えるが，実際の介入の実施にあたっては，現場ごとの最終判断が必要である．以上の議論を踏まえ，行動介入シミュレータの適用範囲について整理したものを，図 6 に示す．

このように，行動介入シミュレータの活用においては，いくつかの条件や制約を伴う．しかし，介入施策の方針を決める工程を省力化できるため，そのインパクトは大きいものと期待している．先に述べた通り，この工程は介入方針の決定に高度な分析を必要とし，その分析精度が行動変容の成否を大きく左右すると予想されるためである．

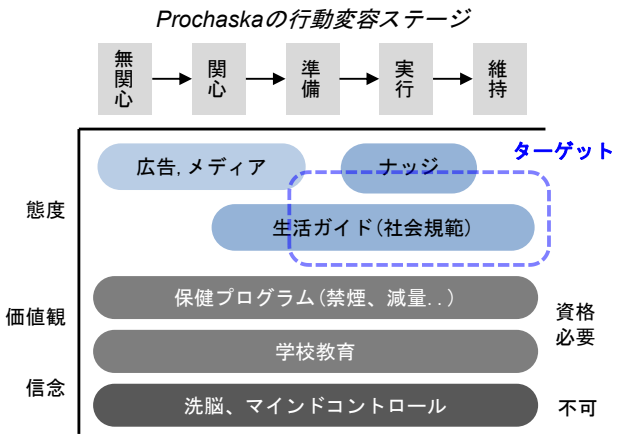

図 6　行動介入の実施ターゲット

3.4　今後の課題

以下，本シミュレータの今後の課題について述べる．機械学習に用いた訓練データは，ラボ実験データが主体である．ラボ実験は，参加人数が小規模であり，また参加者がゲーム課題の結果として被る損益も微小である．したがって，集団サイズが数千人規模以上となる問題，全員の非協力により深刻な脅威が生じる問題，集団利益の獲得に大きなタイムラグがある問題などの予測については，いわゆる外挿となる．そのため，予測自体は可能なものの信頼性が低下する可能性が高い．これらの問題をカバーするためには，参加者にその状況を想像してもらうようなラボ実験を実施するか，フィールド実証の結果を学習する必要がある．

同様に，協力行動が明確に定義できない問題，集団と個人の利害関係の方向性が同じ (つまり，非ジレンマ的な) 問題などについては学習サンプルが無いため予測困難である．このような問題については，この状況を模擬したラボ実験やフィールド実証のデータを別途収集し，学習することが必要となる．

4.　おわりに

本研究では，社会的ジレンマにおける協力行動の「予測」を目的に，ラボ実験の膨大なデータを利活用して訓練データとする機械学習モデルについて検討した．その結果，構築された機械学習モデルのフィッティング性能は非常に高く (R^2_{*F}＝0.83)，未知データに対する汎化性能も良好であった (R^2_{*G}＝0.64)．このことから，予測モデルを構築する手法としての有効性が確認された．

さらに，機械学習モデルを解析エンジンとするシミュレータの開発を通じ，ユーザの意思決定を支援するための各種機能が実現可能であることが分かった．これらの結果は，蓄積された社会的ジレンマに関する数々の知見を，シミュレータの形態で社会に還元可能であることを示唆している．本ツールの活用により，様々な社会的ジレンマの状況下において，適切な介入施策が実施され，協力行動が促進されることを期待したい．

今後の課題として，以下の 2 点を挙げる．1 つ目は，ラボ実験のデータで構築された機械学習モデルが，どこまでフィールドに適用できるかについての評価である．評価にあたっては，今回用意したパラメータを用いて様々な現場や介入内容が表現できるかの観点と，予測精度に対する観点からの検証が必要である．開発したシミュレータの試用を通じ，それらの実力を検証したい．2 つ目の課題は，シミュレータの提供に対する価値の評価である．想定するユーザは，現場の社会的ジレンマを解決したい介入実践者であり，この中には政策立案者，自治体の職員，企業の管理職などが含まれる．想定ユーザのニーズ評価と共に，シミュレータの支援によって自己効力感が向上するかなどの心理的効果についても明らかにしたい．

参考文献

[1] Hardin, G.: The tragedy of the commons: The population problem has no technical solution; it requires a fundamental extension in morality, *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[2] Olson, M.: The Logic of Collective Action: Public Goods and the Theory of Groups, Cambridge: Harvard University Press, 1965.

[3] Dawes R. M.: Social Dilemmas, *Annual Review of Psychology,* vol. 31, no. 1, pp. 169-193, 1980.

[4] Van Lange, P. A., Joireman, J., Parks, C. D. and Van Dijk, E.: The psychology of social dilemmas: A review, *Organizational Behavior and Human Decision Processes,* vol. 120, no. 2, pp. 125-141, 2013.

[5] Komatsu, S. and Nishio, K.: Applicability of 'Nudge' as information provision for energy and electricity conservation - Energy reports for the US households as a case example-, *CRIEPI Socio-economic Research Center Report,* vol. Y12035, 2013.

[6] Bruce, J.: TravelSmart: large-scale cost-effective mobility management. Experiences from Perth, Western Australia, *Proceedings of the Institution of Civil Engineers - Municipal Engineer,* vol. 151, no. 1, pp. 39-48, 2001.

[7] Thaler, R. H. and Sunstein, C. R.: Nudge: Improving decisions about health, wealth, and happiness, New Haven: Yale University Press, 2008.

[8] DellaVigna, S. and Linos, E.: RCTs to Scale: Comprehensive Evidence From Two Nudge Units,

Econometrica, vol. 90, no. 1, pp. 81-116, 2022.

[9] Brewer , M. B. and Kramer, R. M.: Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing, *Journal of Personality and Social Psychology,* vol. 50, no. 3, p. 543–549, 1986.

[10] Van Vugt, M., De Cremer, D. and Janssen, D. P.: Gender Differences in Cooperation and Competition: The Male-Warrior Hypothesis, *Psychological Science,* vol. 18, no. 1, pp. 19-23, 2007.

[11] Somashekhar, S. P., Sepúlveda, M. J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Kumar, C. R., Rauthan, A. N., Kuma A. and Ramya, Y.: Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board, *Annals of Oncology,* vol. 29, no. 2, pp. 418-423, 2018.

[12] Bosco, F. A., Steel, P., Oswald, F. L., Uggerslev, K. and Field, J. G.: Cloud-based Meta-analysis to Bridge Science and Practice: Welcome to metaBUS, *Personnel Assessment and Decisions,* vol. 1, no. 1, pp. 3-17, 2015.

[13] Bosco, F. A.: Accumulating Knowledge in the Organizational Sciences, *Annual Review of Organizational Psychology and Organizational Behavior,* vol. 9, pp. 441-464, 2022.

[14] Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Ten Teije, A. and Balliet, D.: The Cooperation Databank: Machine-Readable Science Accelerates Research Synthesis., *Perspectives on Psychological Science,* vol. 17, no. 5, pp. 1472-1489, 2022.

[15] Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors, *Nature,* vol. 323, no. 6088, p. 533–536, 1986.

[16] Breiman, L.: Random forests, *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[17] Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, USA,* pp. 785-794, 2016.

[18] Breiman, L.: Stacked regressions, *Machine Learning,* vol. 24, no. 1, pp. 49-64, 1996.

[19] Prechelt, L.: Early stopping-but when?, in *Neural Networks: Tricks of the trade*, Springer, Berlin, Heidelberg, 1998, pp. 55-69.

[20] Van Lissa, C. J.: MetaForest: Exploring heterogeneity in meta-analysis using random forests, p. PsyArXiv, 2017.

[21] Harrer, M., Cuijpers, P. T., Furukawa. A. and Ebert, D. D.: Doing Meta-Analysis with R: A Hands-On Guide., ISBN 978-0-367-61007-4 ed., Boca Raton, FL and London: Chapman & Hall/CRC Press, 2021.

[22] Prochaska, J. O. and DiClemente, C. C.: Stages and processes of self-change of smoking: Toward an integrative model of change, *Journal of Consulting and Clinical Psychology,* vol. 51, no. 3, pp. 390-395, 1983.