# 二重過程理論に基づいた道徳判断の事例ベースモデリング
# Instance-based modelling of moral judgements based on dual process theory

佐々木健矢[†] 長島 一真[‡] 西川 純平[‡] 森田 純哉[§]

Kenya Sasaki, Kazuma Nagashima, Jumpei Nishikawa and Morita Junya

[†]静岡大学情報学部, [‡]静岡大学創造科学技術大学院, [§]静岡大学学術院情報学領域
Shizuoka University Faculty of Informatics, Graduate School of Science and Technology Shizuoka University
sasaki.kenya.21@shizuoka.ac.jp

## 概要

自動運転車などの自律した機械を社会実装するためには, 機械自体が直面する道徳的問題に対して, 人間と整合する判断を行う必要がある. 本研究では, 人間とコンピュータの価値観のすり合わせを達成するため, 二重過程理論に基づく人間の思考システムとしての道徳をモデル化する. この研究のステップとして, 言語モデルと認知アーキテクチャ ACT-R を組み合わせた事例ベースな道徳判断のプロトタイプモデルを構築し, ケーススタディとしてのトロッコ問題に適用する.

キーワード : 二重過程理論 (dual process theory) , 道徳 (moral) , 認知アーキテクチャ (cognitive architecture)

## 1. はじめに

自動運転車など自律した機械の社会実装を進めるためには, 人間と整合的なコンピュータによる道徳的判断が必要である. そして, 機械と人間の道徳判断を整合的なものとするためには, 道徳判断の背後に存在する思考システムを深く理解しなければならない. 人間が道徳を通じて複雑な社会をどのように認識し, どのように判断しているのかを検証することで, 人間とコンピュータの価値観のすり合わせが達成されると考える.

本研究では, 二重過程理論 [1] から示される 2 つの思考モード, システム 1 とシステム 2 [2] に基づき, 人間の道徳的判断のメカニズムを検討する. また, 道徳的判断に関する有名な思考実験の一つであるトロッコ問題 [3] を課題として用いて, 言語モデルと認知アーキテクチャ ACT-R [4] を統合するモデルを開発する. このモデルにおいては, トロッコ問題の問題文を手がかりとして検索された記憶事例に基づく判断が行われる. 人間の道徳判断の特性を, このモデル上で表現することで, 人間とコンピュータの価値観をすり合わせる一段階上の次元からとらえた道徳, メタ道徳の実現に踏み入ることが可能となる.

以下では, 本研究の背景として, 二重過程理論と道徳的判断についての考察を行い, この理論に関する認知モデルの先行研究を紹介する. そして, ACT-R を用いた先行研究を説明し, 本研究で提案するモデルへとつなげる.

## 2. 関連研究

二重過程理論とは, 人間は速い思考と遅い思考の異なる 2 つの思考モードを用いて意思決定を行っているとする理論である. 速い思考はシステム 1, 遅い思考はシステム 2 と呼ばれ, 前者はヒューリスティックに基づく無意識的, 直感的, 衝動的なもので, 後者はより意識的で処理に負担のかかる熟慮的なものであるとされている. 人間は基本的にシステム 1 を中心に思考しており, システム 2 がシステム 1 と競合し, 必要なときに介入してより複雑な意思決定を行うと言われている [5] . システム 1 とシステム 2 は明確に区別されておらず, 2 つのモードの間にはスペクトラムが存在し, システム 1 はより少ない「努力」, システム 2 はより多くの「努力」を必要とするものとして考えられる [6] .

システム 1 に基づくヒューリスティクスの典型として, 利用可能性ヒューリスティックスと呼ばれる表面的に類似する事例に基づく推論が見つかっている. このヒューリスティクスは, 倫理的, 政治的判断が問われる問題に対する類推による意思決定を扱う研究において観察される [7] . また, 社会問題を類似事例の検索によって推論するモデルも挙げられる [8] .

これらの政治的判断の類推モデルは, システム 1 とシステム 2 の切り替えの動的プロセスを考慮していないという問題が存在する. また, モデルに与えられる事例はハンドコーディングによって記述されており, モデルの一般性に疑問符がつくものである. これらの問

題点を解決するためには，時系列的なプロセスを含む統合的なモデルが必要である．認知の統合モデルは，認知アーキテクチャの概念によって実現され，ACT-R はその代表例である．

ACT-R を用いた記憶の想起に基づいた意思決定に関する研究例として，都市の人口規模を判断する再認ヒューリスティックスモデル [9] が挙げられる．しかし，このモデルが扱う課題は，道徳的判断におけるシステム 1 とシステム 2 の切り替えを説明するものとしては適切ではない．そこで本研究では，ACT-R を用いたモデルを導入し，システム 1 とシステム 2 の動的な変化を考慮した道徳的判断に対応するモデルを構築する．

## 3. 事例表現

関連研究 [9] で事前知識がモデルに与えられたように，本研究では，モデルが宣言的記憶に保持する事例として，livedoor ニュースコーパス [10] のニュース記事を用いる．このコーパスは，特徴の異なる複数のニュースサイトから構成されており，本研究ではこれらを，道徳的判断を行う個人の文化的背景とみなす．

事例のコーディングでは，利用可能性ヒューリスティックスを表現するために，問題文に対する感情と類似度を付与した．感情と類似度は，それぞれ言語モデルを用いた手法で計算した．感情の要素には，Google Cloud による Natural Language API [11] を使用した．Natural Language API は，文から認識された肯定的・否定的な感情を示す score と，感傷的な内容の量を示す magnitude を組み合わせて，文全体の感情的な要素を表現する．また，Sentence BERT [12] を用いて，問題文の各文と記事タイトルの類似度を算出した．

## 4. プロトタイプモデル

図 1 はモデルが行う道徳判断の過程を示すフローチャートと，歩道橋問題の文章，そしてニュース記事タイトルの例を示したものである．フローチャートが示すように，システム 1 による意思決定のプロセスにシステム 2 が介入する設計となる．モデルは，与えられた問題文を順番に繰り返し読み，各文に対して読んでいる文から連想されるニュースの事例を検索する．本研究で構築したプロトタイプモデルは，著者らの構築したプロトタイプモデル [13] における利用可能性ヒューリスティックスの活性値処理を修正したものとなっている．事例の検索処理では ACT-R のチャンク活性化メカニズムが使用されている．ニュース記事はそれぞれ宣言的記憶モジュールにニュースタイトルと感情分析の結果をパラメータとして持つチャンクとして格納され，各チャンクは活性値を保持している．ACT-R では，チャンクの活性値は，現在の状況との関連性（類似性）と経験から推定される効用（チャンクの使用頻度）の和として算出される．したがって，モデルは問題文が入力される度に，その問題文に対する関連性が高く，かつ効用の高いニュース記事を連想する．そして，事例の効用は連想に成功するごとに増加するため，問題文を読み進めることで，連想が特定の事例へ収束していくことになる．モデルに対して，最後の一文が入力されると，事例に付与された感情に従って判断を下す（POSITIVE は「押す」，NEGATIVE または NEUTRAL は「押さない」，MIXED は「考え直す」）．

上記のプロセスは，人間の利用可用性ヒューリスティックスにおける既知の特性と概ね一致すると考えられる．Kahneman は，システム 1 に基づく思考プロセスは，想起しやすい記憶によって決定され，感情に大きく影響されると述べている [2] ．これとは逆に，検索された感情が最後の一文で MIXED になったとき，システム 2 が介入する余地があると考える．このとき，モ

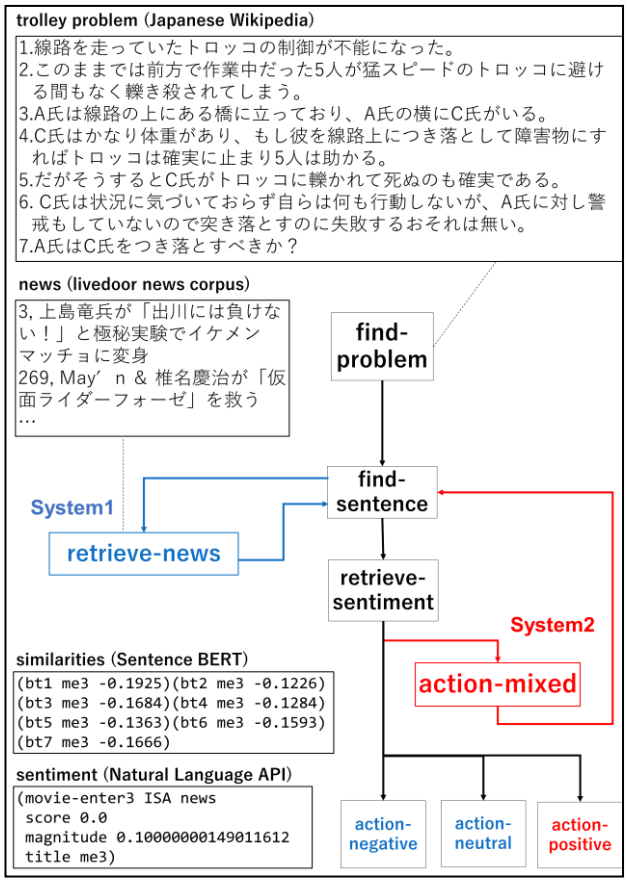


図 1 二重過程論に基づいた道徳判断

表 1 事例表現とモデルによる道徳判断 （p=positive の割合，res=「押す」を選択した割合，nil=再考を繰り返した割合，▲▽=カイ二乗検定における有意差）

| サイト名 | 記事数 | Positive | negative or neutral | mixed | p | $res_s$ | $res_b$ | $nil_s$ | $nil_b$ |
|---|---|---|---|---|---|---|---|---|---|
| トピックニュース | 770 | 96 | 653 | 21 | 0.128 | 0.124 | 0.113 | 0.018 | 0.026 |
| Sports Watch | 900 | 219 | 658 | 23 | 0.250 | 0.218 | 0.24 | 0.021 | 0.014 |
| IT ライフハック | 870 | 578 | 179 | 113 | 0.764 | 0.693 | 0.695 | 0.074▽ | 0.102▲ |
| 家電チャンネル | 864 | 430 | 322 | 112 | 0.572 | 0.514 | 0.53 | 0.043 | 0.049 |
| MOVIE ENTER | 870 | 395 | 421 | 54 | 0.484 | 0.463▲ | 0.404▽ | 0.036 | 0.052 |
| 独女通信 | 870 | 338 | 461 | 71 | 0.423 | 0.385 | 0.393 | 0.026 | 0.031 |
| エスマックス | 870 | 388 | 359 | 123 | 0.519 | 0.524 | 0.532 | 0.088 | 0.104 |
| livedoor HOMME | 511 | 249 | 248 | 14 | 0.501 | 0.487 | 0.496 | 0.086 | 0.099 |
| Peachy | 842 | 564 | 233 | 45 | 0.708 | 0.684 | 0.655 | 0.047 | 0.063 |

デルは再度問題文を読む．ここではシステム 1 と同一の処理を行って再考することで意思決定の試行回数が増え，「努力」を必要とすると解釈することができる．

## 5. シミュレーション

### 5.1 目的・方法

トロッコ問題の異なる表現としてスイッチ問題と歩道橋問題を用意し，異なるニュースサイトからニュース記事を記憶した，異なる文化的背景を持つ 9 つのモデルを準備した．これらのモデルでシミュレーションを行い，文化的背景によって道徳判断に差が生じることを検討した．Greene [5] は，スイッチ問題と歩道橋問題が人間を対象とした実験において異なる意思決定結果を導くことを指摘した．両表現について，シミュレーションを 1000 回実行した．表 1 の 7 列 ($=res_s$) と 8 列 ($=res_b$) は，各サイトに対応して構築されたモデルが，「押す」を選択した割合を，スイッチ問題と歩道橋問題に区別して示している．さらに，9 列 ($=nil_s$) と 10 列 ($=nil_b$) にはスイッチ問題と歩道橋問題で再考を繰り返すループへと突入し，時間内に決定を下すことができなかった割合を区別して示している．

### 5.2 結果・考察

ニュース記事の感情，つまり事前に記憶していた事例の感情的要素が意思決定に影響したことを確認するために，全てのニュースサイトにおいて$res_s$と$res_b$の差を検討した．結果，ニュースサイト「MOVIE ENTER」からニュース記事を収集したシミュレーション結果においてのみ，スイッチ問題と歩道橋問題のそれぞれで「押す」を選択した割合の間に有意な差が見られた ($res_s$ =0.463，$res_b$ =0.404，$p$<.05)．他のニュースサイトにおいては有意差が見られなかった．サイト間で問題文の影響が異なったことから，蓄積された事例による個人差の表現に成功したと言うことができる．つまり，Livedoor ニュースコーパスのサイト間での文化的背景によって道徳判断に差が生じることを示すことに成功したとも解釈できる．

また，再考を繰り返した割合 ($=nil$) は歩道橋問題においてスイッチ問題と比較して高い傾向 ($p$<0.009，$t$(8)=0.503) にあった．これは歩道橋問題において，スイッチ問題よりもシステム２が多く介入したことを示し

ており，トロッコ問題に対する人間による実験で示される傾向と異なる結果となる．

本研究のモデルでは，先行研究のプロトタイプモデル [13] とは異なり，システム 1 が機能している状態に対するシステム 2 の介入効果は明確に観察されなかった．要因としては，活性値処理が修正されたことによって同じ事例が連想されやすくなっていることが挙げられる．連想されるごとに同じ事例が連想される確率が上がることにより，システム 2 が介入することで選択する結果が変わるということが起こりにくくなっている．そのため，同じ事例のみを連想してループに陥るという事象が観測されることとなった．

## 6. まとめ

本研究では，言語モデルと ACT-R を用いて，二重過程理論に基づいた道徳的判断における利用可能性ヒューリスティックスのプロトタイプモデルを構築した．2 つの単純なトロッコ問題を課題としてシミュレーションを行った結果，モデルによる判断は宣言的記憶にある事例の感情に強く影響されることが示された．今回のプロトタイプモデルでは，システム 1 からシステム 2 への切り替えが感情に基づくものであったため，システム 2 の介入による判断結果の変更は観測されなかった．モデルの動作を人間の道徳的判断の傾向に対応させるためには，より複雑で熟慮的な，感情に基づかない論理的なシステム 2 の再設計が求められ，現時点では文章から人数の情報を取得して四則演算を行う システム 1 と並列競合するシステム 2 の設計を進めている．

## 文献

[1] Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate, Perspectives on psychological science, 8 (3), 223-241.

[2] Kahneman D., (2012). Thinking, Fast and Slow, Penguin

[3] Thomson J. J. (1984). The trolley problem. Yale LJ, 94, 1395.

[4] Anderson J. (2007). How can the human mind occur in the physical universe?, Oxford University Press

[5] Greene J. D., (2015). Moral Tribes: Emotion, Reason, and the Gap Between Us and Them. Penguin Press

[6] Brendan C. and Robert L. (2022). Clarifying System 1 & 2 through the Common Model of Cognition, Proceedings of the 20th International Conference on Cognitive Modeling

[7] Spellman B. A. and Holyoak K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. Journal of personality and social psychology, 62 (6), 913.

[8] Blanchette I. and Dunbar K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals, Memory & Cognition, 29 (5), 730-735.

[9] Schooler L. J. and Hertwig R. (2005). How forgetting aids heuristic inference. Psychological review, 112 (3), 610.

[10] RONDHUIT. (2012). livedoor news corpus. [Data set]. https://www.rondhuit.com/download.html#ldcc

[11] Google, “Cloud Natural Language API”, Google Cloud, https://cloud.google.com/natural-language/docs/reference/rest

[12] Reimers N. and Gurevych I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[13] 佐々木健矢，長嶋一馬，西川純平，森田純哉 (2023). 言語モデルと ACT-R を利用した道徳判断の事例ベースモデリング. HAI シンポジウム 2023, P-74