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

Analogical modelling of moral judgements using language models and ACT-R

佐々木健矢 [1]　長島 一真 [2]　西川 純平 [2]　森田 純哉 [3]

Kenya Sasaki[1], Kazuma Nagashima[2], Jumpei Nishikawa[2] and Morita Junya[3]

[1] 静岡大学情報学部
[1] Shizuoka University Faculty of Informatics
[2] 静岡大学創造科学技術大学院
[2]Graduate School of Science and Technology, Shizuoka University
[3] 静岡大学学術院情報学領域
[3]College of Informatics, Shizuoka University

Abstract: A fundamental problem in HAI is the reconciliation of values between humans and computers. To solve this, it is necessary to intervene in morality as a thought system. The study is based on the two modes of thinking proposed by Kahneman (System 1 and System 2) and the distinction between morality and utilitarianism by Greene, who succeeded him. As a concrete research step, a prototype of case-based modelling combining language modelling and ACT-R is presented, using the trolley problem as a case study.

1. はじめに

近年、自動運転車など自律的に動作する機械の研究開発が進展している。これらの技術の社会実装が議論される際、コンピュータによる道徳的判断がしばしば問題となる。この問題は、人間とコンピュータ間の価値のすり合わせ (alignment) という HAI (Huma Agent Interaction) の根本的な問題の一つと関連する。今後、高度に自律化が進展するに従い、思考システムとして人間の道徳に対する踏み入った理解が必要となる。人間は各自が道徳を重んじているにもかかわらず、対立が起きてしまう。そのため、人間は道徳を通して社会をどの様に捉えているのか、人間は道徳を通してどの様に判断を下しているのかを検討することが求められる。

本研究では二重過程理論[1]から示される思考モードのシステム 1 とシステム 2 、それを引き継いだ Greene による道徳と功利主義の区別[2]に基づいて人間の道徳的判断のメカニズムについて検討を行う。また、道徳的判断において有名な思考実験の一つであるトロッコ問題を課題とし、日本語の問題文を文章で与えることで記憶している事例に基づいて判断を行うモデルを、言語モデルと認知アーキテクチャの ACT-R (Adaptive Control of Thought-Rational)[3]を利用することで構築する。

2. 関連研究

本節では本研究の背景として二重過程理論と道徳判断に関する議論を述べ、この理論に関する過去の認知科学的な計算機モデリングの研究を示す。その後に ACT-R を利用した先行研究を述べ、本研究において提案するモデルと接続する。

2.1.二重過程理論

人間には速い思考と遅い思考の二つの異なる思考モードが両立して存在し、これらを通して意思決定を行っていると説明する理論を二重過程理論と呼ぶ。速い思考はシステム 1、遅い思考はシステム 2 とも呼ばれ、前者はヒューリスティックスによる無意識的、直感的かつ衝動的な思考であり、後者はより意識的で処理に負担のかかる熟慮思考であるとされている。通常の状態では基本的に負担のかからないシステム 1 が中心に情報の処理をしており、必要に応じてシステム 2 が介入することでより複雑な意思決定を行っているとされている[2]。Greene はシステム 1 による判断を道徳的判断、システム 2 による判断

を功利主義的判断と述べている[3]。また、システム 1 とシステム 2 は明確に区別されるものではなく、処理に必要なワーキングメモリのコストなどの“努力”が小さいものをシステム 1、大きいものをシステム 2 とし、二つの思考モードの間にはスペクトラムが存在する[4]。

2.2.類推による政治的判断のモデル

複雑な問題の解決手段の一つとして利用可能性ヒューリスティックス[1]が存在する。これは事例をベースとした思考の一種である。特に問題解決の時点で想起されやすい（利用しやすい）事例を利用することで目の前の複雑な問題に対処する思考を指す。事例の利用可能性は過去にその事例が思い出されてきた頻度や、その事例の持つ感情の強さに影響される。問題に直面したときに即座に機能し、感情的要素に大きく影響されるため、このヒューリスティクスは主にシステム 1 に対応づけられる。

利用可能性ヒューリスティクスは、表層的な類似に基づく類推と呼ぶこともできる。過去、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に関する研究はいくつか存在する。Spellman らの研究 [5]では、制約充足の考え方に立つ類推の計算機モデル ACME (Analogical Constraint Mapping Engine) [6]を用いた湾岸戦争と第二次世界大戦に関する類推のシミュレーションを示している。また、Blanchette らの研究[7]では、類似した事例の検索による社会問題への推論を対象に、表層的な類似性と構造的類似性、目的、感情を考慮したモデルを作成している。

これら類推による政治的判断のモデルは、システム 1（表層的類似、感情）をベースとした通常の思考とシステム 2（構造的類似）をベースとした遅い思考の区別をよく説明する。しかし、これらのモデルにおいては、システム 1 とシステム 2 が切り替わる時間的な過程は説明されていない。また、表層的類似と構造的類似の定義についても、ハンドコーディングに基づくものであり、モデルの汎用性には疑問がある。

2.3.事例ベース推論の ACT-R モデル

上記で述べた類推に関する研究の 1 つ目の問題を解決するためには、問題解決の時系列的な過程を含む統合的なモデルが必要である。認知の統合的モデルは、認知アーキテクチャの考え方により成し遂げられると考えられており、ACT-R はその代表とされる。

ACT-R を用いた利用可能性ヒューリスティクス（事例ベース推論）の研究として、Schooler らによる都市の人口の大きさ判断するモデル[8]が挙げられる。このモデルは外国における実在の都市の人口の大きさを都市名のみから判断させる実験を対象としている。実験では、都市の名称の呼び出されやすさ（利用可能性）に応じて、人口規模が過大評価される傾向が示されている。

この実験結果をシミュレーションするために、Schooler らは現実のニュースコーパスを利用することで実験参加者が有するであろう都市名に対する利用可能性を見積もった。ACT-R の記憶には活性値が付与され、この値が高い記憶が優先的に想起される。Schooler らのモデルは、事前にモデルにニュース記事を記憶させ、都市の記憶の活性値をニュースコーパスにおける都市の出現頻度によって定めた。そして、ACT-R の有するプロダクションシステムを利用することで、都市の人口の推定に関わる決定プロセスを構築し、実験結果の再現に成功した。

Schooler らのモデルは、利用可能性ヒューリスティクスを利用した社会的な意思決定の研究にコーパス内の頻度という具体的な説明変数を導入したという点で進展がある。しかし、このモデルも道徳的判断におけるシステム 1 とシステム 2 の切り替えを説明していない。そこで、本研究では Schooler と同様に ACT-R を使用し、トロッコ問題という道徳的判断に対応したモデルを構築する。

3. プロトタイプモデル

本節では本研究で構築を進める道徳判断のモデルを示す。以下でははじめに本研究で扱う課題について述べ、その後に本研究で提案する道徳判断のメカニズムを述べる。

3.1. 問題文

本研究は、トロッコ問題における道徳的判断に影響を与える要因をモデル化することを試みる。一つの要因は、トロッコ問題における表層的な文章表現の影響である [2][9]。

一般的に有名なスイッチ問題では、スイッチを押すことで 5 人の作業員に突き進むトロッコの路線を変えてその先の 1 人の作業員を死なせ、5 人の命を助けるか否かという判断をすることになる。一方、歩道橋のジレンマでは歩道橋から、大きな男を歩道橋から突き落として殺すことでトロッコを止め、5 人の命を助けるか否かという判断を強いられる。どちらの問題も 1 人か 5 人の片方が犠牲になるという構造になっている。しかし、スイッチ問題ではスイッチを押すと回答する割合が高いが、歩道橋問題では歩道橋から男を突き落とさないと回答する割合が

表 1: プロトタイプモデルにおける感情の定義

感情	score	magnitude
positive	>0	-
negative	<0	-
neutral	=0	=0
mixed	=0	>0

増える傾向にある。

本研究では、これら 2 つの問題を ACT-R によるモデルに提示し、上記の差異が生じる条件を探ることを意図する。それぞれの問題文をチャンクと呼ばれる ACT-R のデータ構造の形式にコーディングし、それを 1 文ずつモデルに提示する。その過程のなかで、モデルは利用可能性ヒューリスティックス、すなわち事例ベースによる推論から道徳的判断の意思決定をおこなう。

3.2. 事例表現

モデルに持たせる事例には Schooler らと同様、ニュースコーパスを利用した。特に本研究では、livedoor ニュースコーパスのニュース記事をモデルの記憶事例として使用する。このコーパスは、性質の異なる複数のニュースサイトから構成される。表 2 の 1 列にサイト名、2 列に各サイトの記事数を示す。時事問題に関連した記事を含むトピックニュースから家電製品に関わる IT ライフハック、特定の読者層に焦点をあてた毒女通信や Peachy など、多様な情報源からデータが構成される。これらを利用することで、道徳判断を行う個人の文化的背景をモデル化することができると考えている。

事例のコーディングにおいては、利用可能性ヒューリスティックスに基づき、感情と問題文との類似度を持たせた。感情と類似度はそれぞれ言語モデルを使用した手法で計算している。

3.2.1. 感情分析

事例の感情の要素には Google Cloud が提供する Natural Language API[10]による感情分析[11]を使用する。感情分析とは文に付随する感情的要素を抽出する手法である。Natural Language API では文から認識されるポジティブとネガティブな感情を示す score、感情的な内容の量を示す magnitude を組み合わせることで，文全体の感情的要素を表現する。これらの指標を組み合わせることにより，プロトタイプモデルは表 1 のように感情を定義した。

この定義に従い、livedoor ニュースコーパスの記事タイトル全てに対して感情分析を適用した。表 2 の 3 列 (positive)、4 列 (negative or neutral)、5 列 (mixed) にその結果をサイト別にまとめた。なお、今回のモデルでは negative と neutral を同様に扱うため、これらの記事数は合算している。さらに 6 列 (p) には各サイトにおけるポジティブな記事の比率 (positive / negative or neutral) を示している。この結果より、コーパスに含まれるサイトが多様な感情的な背景を有していることが示される。

3.2.2. 問題文と事例の類似度

問題文と記事タイトルの類似度の計算には Sentence BERT[12]を使用した。この手法は、本研究の実施時において、文の意味を抽出する State of the Art (SOTA) の手法とみなせる。現在の深層学習ベースの自然言語処理では、文内の論理構造の十分なコーディングが行われていない問題も指摘されるものの、表層的な類似を計算する本研究の目的において問題とはならないと考えられる。

Sentence BERT を適用することで、問題文の各文章と記事タイトル全てに対して文章ベクトルを生成し、生成されたベクトルからコサイン類似度を計算する。本研究ではこの手法で計算されたコサイン類似度をさらに ACT-R で使用する類似度へ変換している。

表 2 事例表現とモデルによる道徳判断

サイト名	記事数	positive	negative or neutral	mixed	p	r_s	r_b	$\bar{r}$	$\bar{r}$ − p
トピックニュース	770	96	653	21	0.128	0.121	0.124	0.122	-0.005
Sports Watch	900	219	658	23	0.250	0.237	0.251	0.244	-0.005
IT ライフハック	870	578	179	113	0.764	0.754	0.769	0.761	-0.002
家電チャンネル	864	430	322	112	0.572	0.589	0.581	0.585	0.013
MOVIE ENTER	870	395	421	54	0.484	0.456	0.490	0.473	-0.011
独女通信	870	338	461	71	0.423	0.429	0.432	0.430	0.007
エスマックス	870	388	359	123	0.519	0.560	0.532	0.546	0.026
livedoor HOMME	511	249	248	14	0.501	0.515	0.494	0.504	0.003
Peachy	842	564	233	45	0.708	0.690	0.693	0.691	-0.016

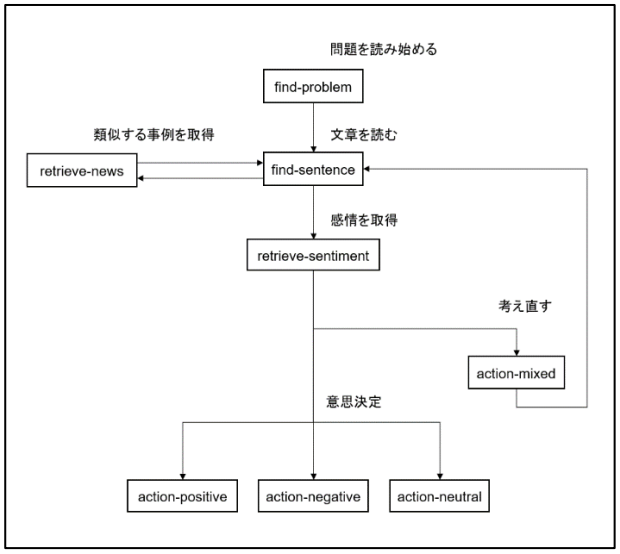

図 1: モデルの思考プロセス

3.3. モデルの思考プロセス

今回作成したプロトタイプモデルは、与えられた問題文を順番に読み、文章ごとに記憶から現在読んでいる文章と類似する事例を取得することを繰り返す。最後の文章を読んだ時点で取得した記憶の感情に従い、アクションを起こす。positive では対象を押すことを選択する。negative または neutral では対象を押さないことを選択する。

このように利用可能性ヒューリスティックスに従い、問題文と類似する事例を取得し、その感情に従って意思決定をする思考過程はシステム 1 にあてはめることができる。また、事例はチャンクとして宣言的知識モジュールに格納されており、それぞれのチャンクは活性値を保持している。ACT-R では想起されたチャンクの活性値が増加し、再度想起される確率が上がるため、過去に思い出してきた頻度が多いほどその記憶が思い出されやすくなるという利用可能性ヒューリスティックスの特性を再現している。感情が mixed の場合は最後の文章をもう一度読み返して考え直すということを行う。考え直すことで試行回数が増え、意思決定にシステム 2 が介入する設計となっている。以上の全体の思考プロセスを図 1 に示す。

4. シミュレーション

livedoor ニュースコーパスのサイトと対応する 9 つのモデルに対してスイッチ問題と歩道橋問題をそれぞれ 1000 回ずつ実行した。表 2 の 7 列と 8 列は、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 ($=\boldsymbol{r_b}$)と歩道橋問題 ($=\boldsymbol{r_s}$)に区別して示している。さらに、9 列にはスイッチ問題と歩道橋問題で「押す」を選択した割合の平均($=\bar{\boldsymbol{r}}$)を示している。これらより、2 つの問題間で「押す」の割合に差は認められず(t(16)=0.097, p=.399)、「押す」の割合はサイト中のポジティブな記事の割合に強く影響をうけていることがわかる($\bar{\boldsymbol{r}}$と pの相関: r=0.998, p < .001)。 ただし、サイトにおけるポジティブな記事の割合は必ずしも「押す」と直結するわけでもない。想起された記事タイトルの感情が mixed であった場合、モデルは考え直しを行う。このシステム 2 による介入の効果を検討するために、$\bar{\boldsymbol{r}}$と p の差（表 2 の 10 行目）を計算し、m との相関を検討した。結果、mixed な記事数と押す割合の相関が有意傾向となった(r=0.636, p=.066)。ここから、システム 2 的な介入が「押す」選択を増加させることが示唆される結果となった。 以上の結果から、プロトタイプモデルにおいては記憶している事例の感情の意思決定に対する影響が大きいと考えられる。トロッコ問題に関する人間を対象とした先行研究とは異なり、「スイッチ問題」と「歩道橋問題」で意思決定が変化する結果とはならなかったが、考え直おすというシステム 2 の介入により押すという功利主義的判断の割合が増加する傾向が見られた。

5. まとめと今後

本研究では言語モデルと ACT-R を利用した、事例ベースの意思決定を行うプロトタイプモデルを構築した。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、意思決定は事例の感情に強く影響され、システム 2 の介入による功利主義的な判断の増加傾向が見られた。

今後の課題として、ACT-R の活性値の処理をより効果的に利用する思考プロセスの導入、事例だけでなく与える問題によって回答の傾向が変わるモデルの作成、文章の類似度に関してより人間の解釈に近い言語モデルの検討が挙げられる。また、システム 2 の意思決定における効果を再度検討する

参考文献

[1] Kahneman D.，ファスト&スロー あなたの意志はどのように決まるのか？(上･下)，竹田円訳，早川書房，(2012)

[2] Greene J. D.，モラルトライブズ――共存の道徳哲学へ（上･下），村井章子訳，岩波書店，(2015)

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

[4] Brendan Conway-Smith. and Robert L. West.: Clarifying System 1 & 2 through the Common Model of Cognition, Proceedings of the 20th International Conference on Cognitive Modeling (2022)

[5] Spellman B. and Holyoak K.: If Saddam Is Hitler Then Who Is George Bush? Analogical Mapping Between Systems of Social Roles, Publication Source, Vol. 62, No. 6, pp. 913-933, (1992)

[6] Keith J. Holyoak and Paul T.: Alogical Mapping by Constraint Satisfaction, Cognitive Science, Vol.13, No.3, pp 295-465 (1989)

[7] Blanchette I. and Dunbar K.: Analogy use in naturalistic settings:The influence of audience, emotion, and goals, Memory & Cognition, Vol.29, No.5, 730-735, (2001)

[8] Schooler L. and Hertwig R.: How forgetting aids heuristic inference, Psychological Review, Vol. 112, No. 3, pp. 610-628, (2005)

[9] Judith Jarvis Thomson: The trolley problem, 94 Yale LJ, pp. 1395-1415, (1985)

[10] Google Cloud: Cloud Natural Language API, https://cloud.google.com/natural-language/docs/reference/rest , (2022/09/28)

[11] Google Cloud: Analyzing Sentiment, https://cloud.google.com/natural-language/docs/analyzing-sentiment , (2023/02/15)

[12] Reimers N. and Gurevych I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv. 1908. 10084, (2019)