人間-AI 協調における過不信予防のための AI エージェントの信頼ダイナミクス予測

Predicting AI Agent Trust Dynamics to Prevent Over/Under-Trust in Human-AI Collaboration

金子 颯汰 [1,2*]　山田 誠二 [2,1]
Sota KANEKO[1,2]　Seiji YAMADA[2,1]

[1] 総合研究大学院大学
[1] The Graduate University for Advanced Studies, SOKENDAI
[2] 国立情報学研究所
[2] National Institute of Informatics

Abstract: 自動運転に代表されるような人間-AI 協調意思決定システムにおいて，適切なシステムの利用および利用効率の向上のために AI エージェントに対する過信/不信を予防し常に適切な信頼を保ちつづけることが重要である．信頼は人間の内部状態であり外部から直接観測することが不可能なため，信頼の変動を捉えることは困難である．そこで本研究では，直接観測不可能な変数を取り扱うことが可能な SEM を時間軸に拡張したダイナミック SEM を適用することによって信頼ダイナミクス・過不信の予測を行うモデルを構築した．

1　はじめに

人工知能技術（AI）は様々な分野で発展を遂げ，自動運転，自律飛行ドローン，自律移動ロボットなど，日常的な場面での活用が急速に進んでいる．このような技術の発展により，人間は作業を AI に委託することが可能となり，作業負担を軽減することができる．自律非行ドローンや自律移動ロボットに全ての操作を任せる場合もあるが，SAE レベル 3 の自動運転のように多くの場合において人間と同じ空間で協働することとなる．このように，AI 技術の活用と開発においては，人間と AI の適切な協力が不可欠であり，ここで重要となるのが人間の AI エージェントに対する信頼 (エージェントの信頼) である [9, 10].

人間が AI の実際の性能を超えてその能力を過大評価すると，本来は委託すべきでない状況でのタスク委託など，システム誤用のリスクが生じる．例えば，自動運転において，天候の悪化による AI 性能の低下にもかかわらず自動運転を継続することは事故につながる可能性がある．このような AI 性能の過大評価は過信と呼ばれる [3]．一方，AI の能力を過小評価 (不信) することは，AI が対応可能なタスクを人間が行うこととなりシステム本来の性能を発揮できない状況に繋がる．したがって，効果的な協働のためには，エージェントに対する適切な信頼を維持することが極めて重要である．

自動運転などの人間と AI の協調において，信頼の変化を推定することは極めて重要である．しかし，信頼は人間の内部状態であるため直接外部から観察することができず，この潜在的な値を理解することが不可欠である．

そこで本研究では，動的に変化する信頼に着目した推定モデルを開発した．本研究では，動的構造方程式モデリング（DSEM）を適用し，効率的かつ説明可能な方法で信頼を推定し，過信/不信を推定する [8]．推定モデルの構築手順は，パス構造の探索的設計とその最も効果的な構造のための単純な最適化で構成される．提案手法の評価のため自動運転シミュレータで実験を実施した．

2　関連研究

2.1　HAI における信頼

Human-Robot Interaction（HRI）における信頼形成の研究では，信頼に影響を与える要因は次のように分類される [1, 4]

- ロボット (エージェント) に関連する要因

*連絡先： 総合研究大学院大学/国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2 丁目 1 番地 2 号
E-mail: sota@nii.ac.jp

- タスクと環境に関連する要因
- 人間に関連する要因

また，信頼形成への影響は，ロボットに関連する要因，タスクと環境に関連する要因，人間に関連する要因の順に大きい．ロボットエージェント関連する要因には，ロボットの信頼性，タスク失敗のタイミングと頻度，システムの透明性などがあり，これらはロボットの動作の質を決定すると考えられている．タスクと環境に関連する要因には，合理性，タスクが人間に及ぼす危険性，タスクの負荷と複雑さなどが含まれる．人間に関連する要因には，パーソナリティ，システムに関する知識，ロボットとの過去の経験などが含まれる．

HAI における信頼形成は ロボットに関する要因 をエージェントに関する要因 と置き換えることで同様に考えることができるが，物理的な実態の有無による違いを考慮する必要がある．

2.1.1 信頼ダイナミクス

Luo らの自動運転車とのインタラクションにおける信頼変化に関する研究では，内部システム要因による性能変化は，外部システム要因による性能変化よりも信頼に与える影響が大きいことが示されている [7]．内部システム要因による性能低下はセンサーの故障などによって引き起こされ，一方，外部要因による性能低下には道路工事による迂回や交通渋滞による所要時間の増加などが含まれる．

また，信頼のメカニズムを組み込むことで，これらの要因を考慮しないモデルよりも正確な信頼推定が可能であることが示されている [2]．さらに，信頼変化に基づくクラスタリングを行い，各クラスタに適した信頼推定モデルを形成することで信頼推定の精度を向上させるなど，信頼変化を活用した新たなモデリング手法の開発も進められている [6]．このように，信頼変化を考慮したモデルを構築することは，信頼の正確な推定を可能にするだけでなく，信頼に影響を与える要因を正確に理解することも可能にする．

3 提案手法

外部から直接観察することができない内部状態である信頼変化の説明可能なモデルを推定し構築するため，構造方程式モデリング（SEM）を時系列データに拡張した動的構造方程式モデリング（DSEM）を用いる．SEM は，直接観察・測定可能な観測変数と測定不可能な潜在変数を扱い，変数間のパス解析を行うことで因果関係を推定するモデルである [5]．この SEM を時系列データを扱い次時点の値を推定する目的で時間軸方向に拡張したものが DSEM である [12]．DSEM を推定モデルとして用いることには，主に以下の 3 つの利点がある：

- 人間によって推定された AI の性能の値として定義され，人間の内部状態である信頼の概念を変数として扱うことが可能である．
- 時系列データを扱い，信頼を推定することが可能である．
- ノード間を結ぶエッジ（パス）は先行研究に基づいてトップダウンに与えられるため，モデルの解釈可能性が高い．

我々が提案する信頼推定モデルの構築手法は，以下のステップにまとめられる：

1. パス図の探索的設計：人間が先行研究の知見と設計者の知識を含むドメイン知識に基づいて，SEM の初期静的パス図を設計し，精度が閾値 τ に達するまで改善する．これはヒューマンインザループの手順で行われる．
2. 時系列構造の最適化：静的パス図（ステップ 1）に基づく動的パス図は，異なる時間ステップ間のパス図間に手動で追加されたエッジによって自動的に最適化される．最適化は，制約された時間範囲 η 内の部分列のすべての候補を探索する制約付きブルートフォース探索アルゴリズムを用いて行うことができる．目的関数は時系列ローリングオリジンクロスバリデーションである [11]．この探索の計算複雑性は $O(2^n)$ であり，時系列長 n の指数オーダーは極めて大きいため，ヒューリスティックに設定可能な制約時間範囲 η というハイパーパラメータを導入した．

先行研究に基づき，$\tau = 0.9$ でステップ 1 によってモデルを作成した．作成したモデルのパス図を図 1 に示す．

図において，四角で囲まれたノードは直接観察可能な観測変数を，円形のノードは直接観察できない潜在変数を表している．ノード間に引かれたエッジは，変数間の因果関係を表している．また，緑，青，ピンクのノードはそれぞれ $t-1$，t，$t+1$ 時点の変数を表している．エッジに付随する実数値は実験におけるパス係数を示している．モデルの分析には Mplus[1] バージョン 8.8 を使用し，パス係数の推定にはベイズ推定を用いた．

変数とその範囲を以下に示す．AI と人間の性能は，各タスクの成功確率を意味する．AI に対する信頼は，人間が推定した AI の性能値として定義される．AIP とは AI エージェントの性能（エージェントのタスク成功

[1] www.statmodel.com

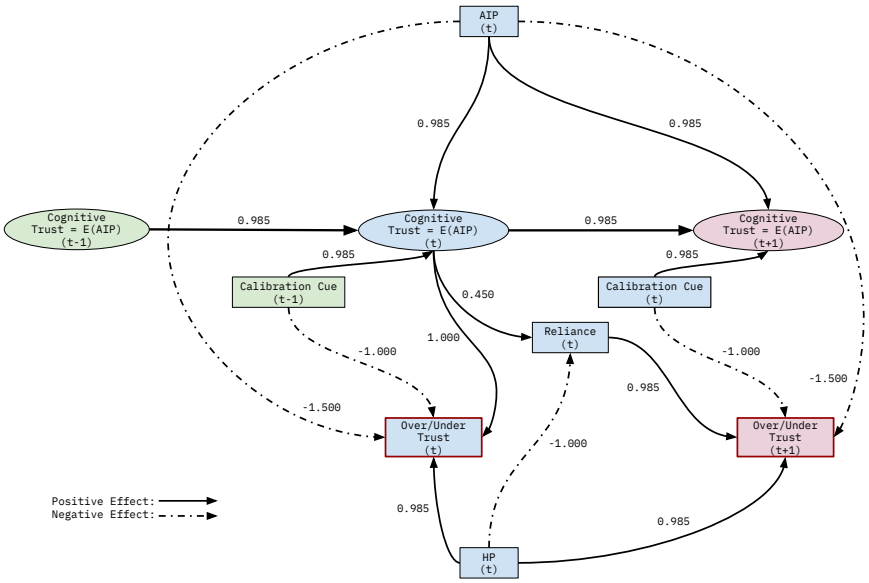


図 1: パス係数を含む動的パス図

確率) $[0, 1]$, HP は人間の性能 (人間のタスク成功確率) $[0, 1]$ である．また，$E(AIP)$ は人間によって推定された $AIP[0, 1]$ であり，これは人間の AI に対する**信頼**である．$Over/UnderTrust$ は過信または不信を判定するために用いられ,「-1」は不信,「0」は適切に調整された信頼,「1」は過信を表す．$Reliance$ はユーザーが自身でタスクを実行した場合は「0」，エージェントにタスク実行を委託した場合は「1」を示す．$Calibration\ cue$ は信頼較正キューがない場合は「0」，ユーザーに信頼較正キューが提示された場合は「1」を示す．このモデルでは，先行研究からの様々な知見と我々の直感に基づいて，次の時間ステップでの重複を除外してエッジが描かれている．このモデルは，$Over/UnderTrust$ 変数を導入することで，過信/不信を直接判定できる．さらに，この $Over/Under\ Trust$ は，今後の研究において過信/不信を効率的かつ正確に予防するために活用できる．

4 自動運転シミュレータによる協調タスク

先行研究の離散時間とは対照的に，より連続的な時間のタスクとして自動運転シミュレータを用いて過信/不信を推定した．このタスクでは，車載カメラからの映像をウェブ上で再生し，再生中に人間が介入するものである．自動運転車の車載映像として再生された動画は，BBD100K ドライブデータセット [13] からのものである．

ユーザーには，再生される映像が自動運転車によって撮影されたものであると説明された．ユーザーはウェブブラウザで映像を再生し，運転中に危険を感じた際にキーボードのスペースバーを押すことで介入の意思を示した．実験中，ユーザーは再生中の映像を継続的に監視でき，危険を感じた際には必要に応じて介入することができた．

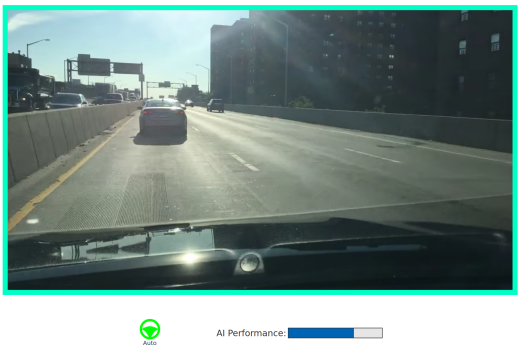
映像は 22 シーンで再生され，最初の 7 シーンでは AI が高性能で運転し，次の 9 シーンで性能が低下，最後の 6 シーンで再び AI の性能が向上した．中間の 9 シーンにおいて，AI の性能が人間よりも低い場合に介入がなければ，過信を示していると判断された．介入は 10 秒間のウィンドウ内で 1 ステップとして記録され，1 シーン内で合計 4 ステップが記録される．

実験で使用したウェブベースの自動運転シミュレータを図 2 に示す．ユーザによる介入が発生すると，映像上に赤枠が表示される．自動運転中は緑枠となる．映像下部に表示されるハンドルアイコンの色も，ユーザーの介入時に緑から赤に変化する．また，図 2 の下部には AI の能力が表示されている．

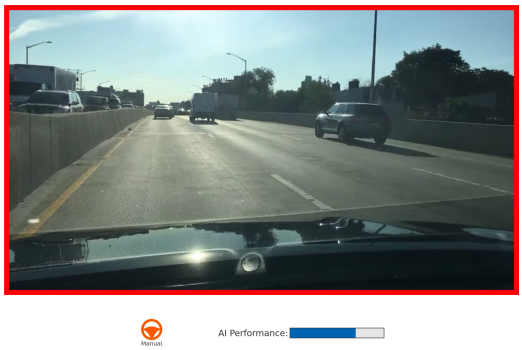
4.1 実験参加者

Yahoo!クラウドソーシング[2]を通じて，100 円で 50 名の実験参加者を募集した．49 名の参加者（2 名のノイズデータを除外）がタスクを完了した（女性 11 名，男性 38 名；年齢：22-66 歳，M = 46.5，S.D. = 9.97）．

[2]https://crowdsourcing.yahoo.co.jp/



(a) AI にタスクを委ねている際の実験画面



(b) ユーザーが介入中の実験画面

図 2: 実験画面のスクリーンショット

表 1: 人間-AI 協調運転タスクにおける実験結果

	ACC	RMSE
	Avg(S.D.)	Avg(S.D.)
DSEM	0.98(0.01)	0.14(0.04)
AR(1)	0.83(0.08)	0.20(0.10)
ARMA(1,1)	0.85(0.06)	0.18(0.08)
SARIMA(1,0,1)[4]	0.85(0.06)	0.18(0.08)

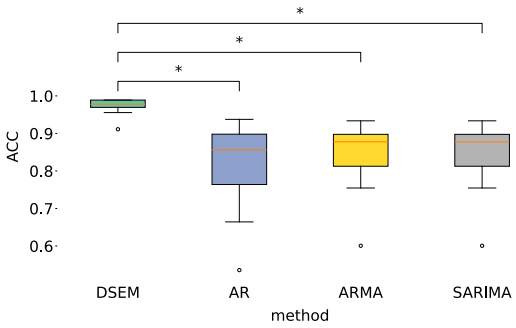


図 3: 過信/不信推定精度の結果

5 実験結果

提案手法による信頼推定モデルを用いた次時点における過信推定の結果を以下に示す．各モデルの ACC と RMSE を表 1 に示す．提案手法による推定では，ACC は 97.8%，RMSE は 0.14 であった．

提案手法と従来手法による推定精度に関する多重比較を伴う一元配置分散分析（ANOVA）[3]の結果を図 3 に示す．有意水準は ($\alpha = 0.05$) に設定した．結果として，提案手法とすべてのベースライン手法との間に有意差が見られた．ベースラインには自己回帰（AR），自己回帰移動平均（ARMA），季節自己回帰統合移動平均（SARIMA）が含まれる．この結果は，提案手法がベースライン手法を上回る性能を示したことを意味している．

6 結論

本研究では，人間-AI 協調意思決定における AI への信頼推定モデルを構築する新しい手法を提案した．我々の手法では，まず探索的設計を行い静的パス図を得た後，時系列パス図に対して最適化を適用する．本フレームワークにおいて，人間の合理的行動の実行を監視することなく過信/不信を直接推定できることは，過信/不信の防止において非常に独創的かつ重要である．提案手法のもう一つの利点は，パス構造による高い説明可能性である．また，我々は提案手法が従来手法を上回る性能を示すことを確認した．

[3]Analysis of Variance

参考文献

[1] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Trans. Interact. Intell. Syst.*, Vol. 8, No. 4, pp. 1–30, 2018.

[2] Michael G. Collins, Ion Juvina, and Kevin A. Gluck. Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computer-

ized confederate agents. *Frontiers in Psychology*, Vol. 7, , 2016.

[3] Ewart de Visser, Marieke M.M. Peeters, Malte Jung, Spencer Kohn, Tyler Shaw, Richard Pak, and Mark Neerincx. Towards a theory of longitudinal trust calibration in human – robot teams. *International Journal of Social Robotics*, Vol. 12, pp. 459–478, 2020.

[4] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. Modeling trust in human-robot interaction: A survey. In *Social Robotics*, pp. 529–541. Springer International Publishing, 2020.

[5] R.B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Publications, 2023.

[6] Jundi Liu, Kumar Akash, Teruhisa Misu, and Xingwei Wu. Clustering human trust dynamics for customized real-time prediction. In *Proceedings of 2021 IEEE International Intelligent Transportation Systems Conference (ITSC'21)*, pp. 1705–1712, 2021.

[7] Ruikun Luo, Jian Chu, and X. Jessie Yang. Trust dynamics in human-av (automated vehicle) interaction. In *EA of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2020.

[8] Christoph Molnar. *Interpretable Machine Learning – A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/, 2023.

[9] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, Vol. 15, No. 2, pp. 1–20, 2020.

[10] Kazuo Okamura and Seiji Yamada. Empirical evaluations of framework for adaptive trust calibration in human-ai cooperation. *IEEE Access*, Vol. 8, pp. 220335–220351, 2020.

[11] Len Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, Vol. 16, pp. 437–450, 2000.

[12] Ellen L. Hamaker Tihomir Asparouhov and Bengt Muthén. Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 25, No. 3, pp. 359–388, 2018.

[13] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, 2020.