人間-AI協調における信頼ダイナミクスの予測

Predicting Trust Dynamics in Human-AI Collaboration

金子 颯汰 [1,2*]　山田 誠二 [3]
Sota KANEKO[1,2]　Seiji YAMADA[3]

[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
[3] 神奈川大学
[3] Kanagawa University

Abstract: 人間-ロボット協調において，適切なシステムの利用のために適切な信頼を構築し，過信/不信を予防することが重要である．しかしながら，信頼は外部から直接観測することが不可能な人間の内部状態であり，信頼の動的変化を捉えることは困難である．そこで本研究では，潜在的な変数を取り扱うことが可能な SEM をベースとしたダイナミック SEM を適用することによって信頼ダイナミクスおよび過不信の予測を行う．

1　はじめに

自律システムが人間の活動環境で利用されるにつれ，人間とロボットが協調して作業する必要のある状況は増加している．産業用ロボットからサービスロボット，自動運転車に至るまで，これらのシステムは，人間が依然として意思決定や行動に対する一定の権限を保持するタスクに関与するようになってきた．このような状況において，人間がロボットをどの程度信頼するかは，協調作業の有効性や安全性に大きな影響を与える．先行研究では，過剰な信頼は過度な依存を招き，一方で信頼が低すぎる場合はシステムの未活用につながることが示唆されている．これらはいずれも，システム全体の性能低下や安全性を損なう可能性がある．したがって，適切な信頼を構築し，かつ維持することが，人間－ロボットのインタラクションにおいて極めて重要である [7].

適切な信頼を構築するうえでの主要な課題の一つは，信頼が個人的な心理状態であり，直接外部から観測できない点にある．自己報告型の質問票は統制された環境における信頼を把握する手段として有用であるが，リアルタイムでの運用には適していない．その結果，近年では，ロボットの行動，環境的手がかり，人間の行動といった観測可能な情報から信頼を推定する研究が増加している．これらの研究は，特定時刻での信頼推定にとどまらず，システム性能や環境要因の変化に応じて信頼がどのように時間的に変化するかを理解することを目的としている．信頼を動的な過程としてモデル化することは，ロボットの性能が変動するタスクや，繰り返しのインタラクションを通じて信頼が形成される状況において特に重要である．

現在の信頼推定手法には，確率的グラフィカルモデルや深層ニューラルネットワークが含まれる．動的ベイジアンネットワークなどの確率モデルは，主要変数間の因果関係を表現でき，解釈性に優れている．一方で，ニューラルモデルは複雑なデータを扱う能力に長けているが，その内部構造は不透明である場合が多い．しかし，これらの手法の多くは即時的な信頼推定に焦点を当てているか，タスク構造に関する特定の仮定に依存している．信頼が逐次的な出来事に応じてどのように変化するのかを明確に理解しない限り，これらのモデルを多様な人間－ロボットのインタラクションにおけるシナリオに適用することは依然として困難である．

これらの課題に対処するため，本研究では，変動するタスク結果と基礎的な認知過程を統合した，信頼ダイナミクスの予測モデリングの枠組みを提案する．タスク性能が信頼形成において重要な役割を果たすことを強調する先行研究に基づき，本手法では，逐次的なロボット行動と文脈的手がかりに基づいて信頼の推移をモデル化する．本手法は解釈性を重視しつつ，時間的変化を取り込むことが可能であり，信頼を固定的な量としてではなく，蓄積された履歴によって形成され

*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

る軌跡として予測することを可能にする．

本手法を検証するために，我々は 2 つの異なる性能シナリオ下で，人間とロボットシステムとの逐次的なインタラクションを含む実験を実施した．参加者は複数のタスク実行を観察し，定められたタイミングで信頼評価を行った．このデータにより，信頼が時間とともにどのように変化するか，また提案モデルが観測可能な変数のみを用いて信頼の推移をどの程度正確に予測できるかを分析した．その結果，本手法は信頼を安定して予測でき，その時間的変化の重要な側面を適切に捉えていることが示された．

本研究は，ヒューマンロボットインタラクションにおける信頼較正の理解を深めるものであり，信頼ダイナミクスを予測可能かつ解釈可能な形で示すモデルを提供する．このようなモデルは，人間の信頼に応じて挙動を調整する自律システムの設計に不可欠であり，最終的には，より安全で効果的な協調作業の実現に寄与する．

2 関連研究

人間 – ロボット のインタラクションおよび人間 – AI のインタラクションにおける信頼に関する研究は，信頼が自律システムと人間がどれだけ効果的に協働できるかを左右する重要な要因であることを示している [6],[5],[1]．ロボティクス，自動運転車，協調意思決定といった分野において，信頼は，人々がこれらのシステムにどの程度依存するかを調整する役割を果たす．それは，個人がタスクをシステムに委ねるか，あるいは必要に応じて介入するかといった判断にも影響を与える．これらに関する基礎的研究では，信頼形成に影響を与える要因として，ロボットやエージェントの特性，タスクおよび環境の性質，そして人間利用者の特性という三つの主要なカテゴリが特定されている．これらの要因の中でも，システムの性能，信頼性，一貫性が，信頼の形成に最も大きな影響を及ぼす．さらに，メタ分析により，透明性，エラーの発生様式，および知覚される能力が，人間が自律システムを評価する際に極めて重要であることが確認されている．これらの知見を背景として，相互作用を通じて変化する信頼を推定および予測することを目的とした，信頼の計算論的モデル化に向けた重要な取り組みが進められている．

2.1 過信と不信

信頼研究の主要な分野の一つは，不適切な信頼較正状態，特に過信と不信に焦点を当ててきた．過信は，人間の期待がシステムの実際の能力を上回る場合に生じ，その結果，タスクの過度な委任につながる [8]．一方，不信は，システムが十分に能力を有しているにもかかわらず，人間が過剰に介入してしまう状態であり，効率の低下や作業負荷の増大を引き起こす可能性がある．人間 – AI 協調に関する研究では，エラーや不確実な結果に繰り返しによって，信頼の較正不良がどのように形成されるかが示されている [13]．シミュレーション研究によれば，システムが長期間にわたり完璧に動作する場合，特に利用者がシステム内部の信頼性メカニズムを理解していない状況では，過信が徐々に形成される可能性があることが示唆されている．これらの知見は，信頼を固定的な状態として捉えるのではなく，時間とともに変化する動的な状態として捉える必要性を強調している．

2.2 信頼ダイナミクス

信頼ダイナミクスという概念は，信頼が時間とともにどのように変化するのかを理解しようとする研究の進展とともに，注目を集めている．信頼ダイナミクスは，人間が自律システムとの経験を蓄積する過程において，信頼がどのように発展するかに着目する．自動運転車に関する先行研究では，信頼は性能に影響を与える内的要因と外的要因に対して異なる反応を示すことが報告されている．センサの故障といった内的要因は，環境中の障害物などの外的要因よりも，信頼を大きく損なう傾向がある．

他の研究では，システム信頼性の変化，ユーザの行動，環境的手がかりを考慮することで，信頼の遷移をモデル化している．関連研究においては，信頼パターンを分類することで，特有の信頼行動を示すユーザ群を特定し，それぞれに適した信頼予測モデルを構築する試みも行われている．

2.3 信頼ダイナミクスのモデリング

信頼ダイナミクスを予測するために，これまでにさまざまな枠組みが提案されている．動的ベイジアンネットワークに代表される確率的グラフィカルモデルは，性能，信頼，ユーザ行動，および文脈の間に存在する因果関係を捉えることができる [14]．これらのモデルは解釈性に優れる一方で，変数や因果関係を明確に定義する必要がある．一部の手法では，特に継続的な監視が求められる自動運転のような状況において，カルマンフィルタを用いてリアルタイムに信頼を推定する試みがなされている．構造方程式モデリング，SEM，およびその派生手法を用いた因果モデリングも，信頼に影響を与える潜在的メカニズムを明らかにする目的で検討されてきた [4]．

確率的モデリングと並行して，複雑な行動情報や知覚情報から信頼を測定するために，深層学習手法も用いられてきた．Transfomer アーキテクチャや，LSTM に代表されるリカレントニューラルネットワークの利用は，ユーザの関与やシステム相互作用における時間的パターンを捉えるうえで成功を収めている．例えば，Transfomer encoder を用いた信頼推定 [3] や，LSTM モデルを人間の意思決定行動の系列に適用した研究 [10] が報告されている．これらの手法は，画像や連続的なセンサ入力を含む複雑なデータ型を処理できる一方で，その内部動作は不透明である場合が多い．この解釈性の欠如は，信頼予測の説明可能性が重要となる安全性重視の分野において，大きな課題となる．

2.4 信頼較正

信頼較正もまた，重要な研究分野として位置づけられている [2]．研究によれば，過信が検出された際に適切な較正の手がかりを提示することで，信頼を最適な水準へと調整できることが示されている．ユーザ行動に基づいて適時に手がかりを提供する適応的信頼較正枠組みは，人間 0-AI 協調意思決定における過度な依存を防ぐうえで有望であることが報告されている [11]．

さらに，説明可能 AI (XAI) 手法が信頼に与える影響についても検討が進められており，意味のある説明や因果的洞察が，信頼の整合性を高めることが示されている．これらの知見は，信頼予測モデルに対して，高い信頼性と明確な解釈性の両立が不可欠であることを強調している．

3 提案手法

本研究では，構造方程式モデリング (SEM) を時系列データへと拡張した Dynamic-SEM (DSEM) [12] を用いることで，信頼ダイナミクスをモデル化するための予測枠組みを構築する．これにより，ロボット行動の順序に基づいて，将来の信頼状態を予測することが可能となる．本アプローチは，反復的な相互作用を通じて信頼が変化し，かつ自己報告データを直接取得することなく，継続的な信頼予測が求められる状況において，特に有用である．

3.1 予測フレームワーク概要

信頼は直接測定することができない人間内部の状態であるため，提案モデルでは信頼を潜在変数として扱う．本モデルは，観測可能なロボットの性能と，背後にある心理的な要因を組み合わせることで，信頼がどのように形成・発展するかを予測する．モデリング手順は，次の三段階から構成される．(1) 提案する因果関係を示すために，静的な SEM のパス図を作成する．(2) その図を，時間を通じた影響を捉える動的パス構造へと拡張する．(3) 時系列データを用いて動的モデルを推定し，信頼を予測する．

本プロセスは，信頼形成において想定される因果関係を示す静的な SEM のパス図を作成することから始まる．これには，二種類の変数が含まれる．観測変数は，タスクの成功または失敗といった，測定可能な指標を表す．一方，潜在変数はロボットの成功確率に対する期待や信頼といった，人間内部の概念を表現する．先行研究に基づき，タスク結果は知覚される能力に影響を与え，それがさらに信頼に影響を及ぼすと考えられている．この因果構造は，信頼形成を規定する主要な過程を適切に反映するよう，当該分野の知見を取り入れながら選択をする．

時間的なダイナミクスを取り込むために，静的なパス図は時間軸方向へと拡張される．DSEM はラグ付きの関係を導入し，時刻 t における変数が，時刻 $t+1$ における観測変数および潜在変数の双方に影響を与えることを可能にする．これらの拡張には，潜在的な信頼に関する自己回帰パス，タスク結果から知覚能力へのラグ効果，および残差間の時間依存関係が通常含まれる．この構造により DSEM は，過去のロボット行動に基づいて信頼がどのように構築され，変化するかを説明でき，将来時刻における信頼予測の基盤を形成する．適切な時間構造を構築するために，Akaike Information Criterion (AIC) などの情報量規準を用いて，複数の動的パス構成が評価される．これにより，時間をまたぐ依存関係の中から，最も単純かつ効果的な構造をモデルが選択できるようになる．

モデルパラメータは，逐次的な実験データを用いて推定される．この過程では，参加者が複数のロボット行動を観察し，定められた間隔で信頼を報告する．モデルは，同一時点内のパスと，時点をまたぐパスの双方を考慮する．すなわち，時間構造を明示的に取り込み，ベイズ推定を用いることで，潜在変数における自己相関にも対処する．得られたパス係数は，各変数の影響度を定量化するものであり，質問票による回答が得られない場合であっても，信頼を予測することを可能にする．モデル選択は，ローリング型のクロスバリデーションによってさらに補強される．この手法では，時間軸に沿ってウィンドウを構成し，予測精度を評価する．クロスバリデーションと AIC に基づく異なる動的構造間の比較を組み合わせることで，本モデルは予測性能と構造の簡潔性とのバランスを確保している．

3.2　変数および説明

- *Cognitive Trust*: 人間のロボットに対する信頼を表す潜在変数.
- *Success/Failure*: 時刻 t におけるロボットのタスク成功 (1) または失敗 (0) を表す二値変数.
- *Perceived Task Difficulty*: ロボットが実行したタスクの難易度を表す変数であり，事前のマニピュレーションチェックによって確認される.

Perceived Task Difficulty は，実際のタスクを通じて人間がロボットの能力をどのように認識するかを示す指標である．*Cognitive Trust* は，人間の内部的な思考過程を示す変数であり，本モデルにおける主要な出力である.

3.3　実験でのデータ取得

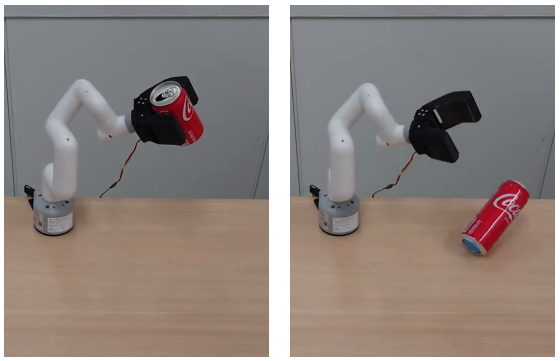
モデル推定のために，参加者は異なる結果に至るロボット行動の系列を観察した．各観察後に，妥当性が確認された測定尺度を用いて信頼評価を収集した．これらの値は正規化され，係数推定の際に，観測変数と潜在的な信頼水準を結び付けるために用いられた．データには時間に沿った繰り返し観測が含まれているため，本モデルは，ロボット行動が信頼ダイナミクスにどのような影響を与えるかを観察することが可能である.

係数推定が完了すると，モデルは将来の時点における信頼を予測する．時刻 $t-1$ までに得られた観測可能なタスク結果を用いて，動的パス構造が *Cognitive Trust* を算出する．この値は，時刻をまたぐパスに入力され，次の時点における *Cognitive Trust* を予測するために利用される．この予測手法により，人間－ロボットのインタラクションにおいて，信頼を継続的にモニタリングすることが可能となる．また，DSEM が有する変数間の関係性は，各変数が信頼形成に果たす相対的な影響を評価することを可能にし，説明可能性が重視される応用分野における理解の向上にも寄与する.

4　実験

本研究では，提案するモデルに基づく信頼予測モデルの構築および検証のために，時系列データを収集する実験を実施した．本実験は，異なる性能パターンの下でロボットの行動を繰り返し観察することにより，信頼がどのように形成・発展するかを明らかにすることを目的としている.

さらに，本課題に先立ち Pick-and-place タスクの知覚的な難易度を確認するためのマニピュレーション・チェックを実施した.



(a) 成功した動作 (小さな缶). **(b)** 失敗した動作 (大きな缶).

図 1: 成功と失敗の動作例.

4.1　実験参加者

本研究には，合計 200 名の参加者が参加した．参加者は無作為に 4 つの実験群に割り当てられ，各群は 50 名で構成された．すべての参加者は，Yahoo!クラウドソーシングを通したオンラインプラットフォームを用いて実験を完了した.

4.2　実験デザイン

参加者は，ロボットアームによって実行される 5 回の Pick-and-place 動作を観察した．図 2,3 は，本実験で使用した動画からのスナップショットを示しており，成功および失敗の両方の動画を準備した.

成功と失敗の動作の並び順のみに基づいて 4 種類の性能パターンを設定した．これらのシナリオは，成功と失敗の分布および順序を変化させることで，信頼に異なる変化を生じさせることを目的としている.

以下に 4 種類のし治りを示す:

1. S-S-F-F-F: 2 回の成功に続き 3 回の失敗
2. S-S-S-F-F: 3 回の成功に続き 2 回の失敗
3. F-F-S-S-S: 2 回の失敗に続き 3 回の成功
4. F-F-F-S-S: 3 回の成功に続き 2 回の失敗

各参加者は，4 つのシナリオのいずれか一つに無作為に割り当てられ，系列の各ステップに対応する 5 本の短い動画を視聴した．各動画の視聴後，参加者は，Multi-Dimensional Measure of Trust (MDMT) [9] から適応した項目を用いて信頼を報告した.

4.3　マニピュレーションチェック

本実験に先立ち，Pick-and-place タスクで使用する 2 種類の物体間におけるタスク難易度の違いを，参加者

(1) (2) (3)

図 2: 成功時のスナップショット．

(1) (2) (3)

図 3: 失敗時のスナップショット．

が認識しているかを確認するチェックを実施した．参加者は，大きさの異なる 2 つの物体をロボットが操作するデモンストレーションを観察し，各タスクがどの程度困難であるかを評価した．

その結果，参加者は 2 つの種類の物体を異なる難易度として知覚していることが示された．これらの結果に基づき，知覚された難易度の評価を，DSEM モデルにおける観測変数として組み込んだ．本研究では，この変数を $Perceived\ Task\ Difficulty$ と定義した．

4.4 実験手順

実験の流れは以下のとおりである．参加者はまず課題の説明を受け，ロボットの一連の行動を観察することを知らされた．その後，割り当てられたシナリオに対応する 5 本の動画を順に視聴した．各動画の視聴後に，MDMT に基づく信頼質問紙へ回答した．これら一連の手続きは，参加者 1 人あたり約 15 分で完了した．

このようにして得られた逐次的な信頼測定値に，タスク結果および知覚されたタスクの難易度を組み合わせることで，動的パス係数の推定と，提案する DSEM 枠組みの予測性能の評価を行った．

5 実験結果

本実験に先立ち，Pick-and-place タスクで用いられた 2 種類の物体間におけるタスク難易度の違いを，参加者が認識しているかを確認するためのマニピュレーションチェックを実施した．参加者は，「ロボットが行った作業の難易度はどの程度だと考えますか?」という質問に対して，7 段階リッカート尺度を用いて難易度を評価した．

その結果，2 種類の物体に対する知覚的な難易度には明確な差が認められた．大きい缶は平均 2.720(SD = 1.217) であったのに対し，小さい缶はより高い難易度として評価され，平均 3.400(SD = 1.311) であった．t 検定の結果，両タスク間の差は統計的に有意であることが示された ($p = 0.009$)．これらの結果は，参加者が一貫して，小さい缶の方をロボットにとって扱いにくいタスクであると認識していたことを示唆している．

この結果を踏まえ，難易度評価は，提案モデルにおける観測変数 $Perceived\ Task\ Difficulty$ として組み込んだ．

5.1 信頼の評価

5 回のロボット動作それぞれの後に，参加者は MDMT に基づく質問票を用いて信頼の評価を行った．表 1 は，4 つの性能シナリオ全体にわたって収集された認知的信頼 項目の回答を要約したものである．これらの評価値は，モデル入力のために範囲 $[0, 1]$ に正規化された．この逐次的な信頼測定値は，タスク結果および知覚タスク難易度とともに，動的モデルを構築するための主要なデータとして用いられた．

表 1: 各シナリオタイプごとの参加者による信頼の評価.

Scenario Type	1st Avg. (S.D.)	2nd Avg. (S.D.)	3rd Avg. (S.D.)	4th Avg. (S.D.)	Final Avg. (S.D.)
SSSFF	4.74 (0.93)	4.83 (0.84)	5.01 (0.93)	1.82 (0.85)	1.73 (0.5)
SSFFF	4.83 (1.08)	4.82 (1.09)	1.80 (0.71)	1.77 (0.73)	1.74 (0.71)
FFSSS	2.23 (1.00)	2.17 (0.99)	5.44 (0.99)	5.54 (0.98)	5.47 (0.98)
FFFSS	2.17 (0.90)	1.98 (0.77)	2.00 (0.78)	5.45 (1.07)	5.62 (1.08)

表 2: 実験結果の予測精度.

	ACC Avg(S.D.)
Proposed Method	0.811(0.216)
ARMA(1,1)	0.541(0.286)
AR(1)	0.564(0.315)

5.2 予測モデルの構築

収集された時系列データは, 動的構造方程式モデルの係数推定に用いられた．モデルには, $Success/Failure$, $Perceived\,Task\,Difficulty$, および $Cognitive\,Trust$ の 3 つの変数が含まれる．動的パス構造の選択にあたっては，Akaike Information Criterion，AIC，を用いて候補モデルを比較し，さらにローリングオリジン型交差検証によって予測安定性を確認した．

最終的な DSEM 構造には，潜在的信頼に対する自己回帰効果，タスク結果からの交差遅延影響，および知覚的なタスク難易度からの一貫したパスが含まれていた．

5.3 信頼ダイナミクスの予測

構築したモデルを用いて，時刻 $t-1$ までに観測されたデータに基づき，時刻 t における信頼値を予測した．図 5 および表 2 は，信頼の値の予測精度の比較を示している．提案手法は，ロボット動作の成功と失敗の切り替わりによって生じる信頼の変化を，適切に捉えることに成功した．

6 結論

本研究は, 人間がロボットと繰り返しインタラクションする過程において，信頼がどのように形成・発展するかを検討し，動的構造方程式モデリング (DSEM) を用いて，将来の信頼を予測する枠組みを提案した．ロボットのタスク成功および失敗の異なる系列から得られた時系列の信頼データに加え，参加者によるタスク難易度の知覚を収集することで，信頼が観測可能な性能だけでなく，タスクの複雑さに対する個人的な期待にも影響されることを示した．知覚されたタスク難易度を組み込むことで，文脈情報が付加され，個人差を伴う信頼反応をより正確に説明できるようになった．

提案したモデルは，成功と失敗の順序が異なる複数のシナリオにわたって，変化する信頼ダイナミクスを適切に捉えることができた．モデル構造は，AIC に基づく評価とローリング型のクロスバリデーションを組み合わせて選択された．この手法により，予測精度と理解しやすさのバランスが確保された．予測結果からは，質問票による直接的な信頼回答を必要とせず，観測されたタスク結果と文脈情報のみを用いて，1 ステップ先の信頼を予測できることが示された．この点において，DSEM はヒューマンロボットインタラクションにおけるリアルタイム信頼推定の有効な手法であるといえる．

本研究の知見は，信頼を考慮したロボットシステム設計に対しても重要な示唆を与える．信頼は直近の経験と過去の経験の双方から形成されるため，静的な信頼指標のみに依存するシステムでは，利用者の期待や行動を誤って解釈する可能性がある．信頼を動的に予測できるようになることで，能動的な信頼較正のための介入戦略の可能性が広がる．例えば，ロボットが自身の行動や情報提示の方法を調整することで，過信や不信を未然に防ぐことができる．

一方で，本研究には今後の課題も残されている．より長期の相互作用系列を対象とし，多様かつ有効な変数をさらに追加し，より複雑なタスクにおけるモデルの有効性を検証する必要がある．

本研究では，構造化された時系列モデルが，信頼の動的変化である信頼ダイナミクスを効果的に予測できることを示した．ロボットの性能や利用者の知覚に応じて信頼がどのように変化するかを明らかにすることで，本研究は，より効果的かつ安全な人間 – ロボット協働を実現するための信頼モデリング研究の発展に寄与する．

7 謝辞

本研究の一部は，JST CREST (JPMJCR21D4, JPMJCR23J5) および科研費 (24H00731) の支援を受けて実施された．

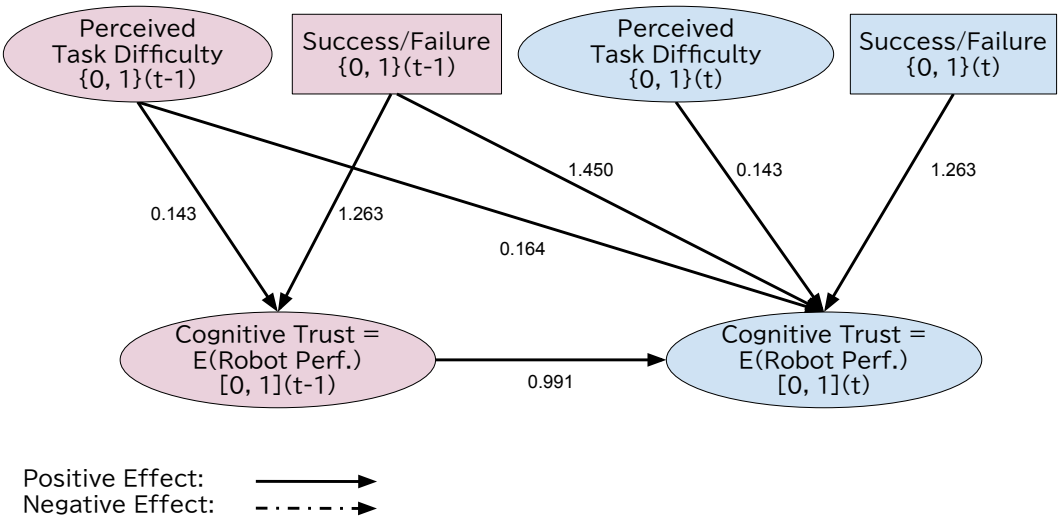


図 4: 予測モデルのパス図.

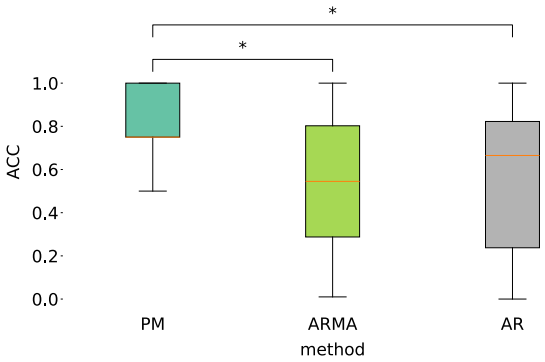


図 5: 予測精度の比較.

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] 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.

[3] Yosuke Fukuchi and Seiji Yamada. Dynamic selection of reliance calibration cues with ai reliance model. *IEEE Access*, Vol. 11, pp. 138870–138881, 2023.

[4] Sota Kaneko and Seiji Yamada. Predicting trust dynamics with dynamic sem in human-ai collaboration. *IEEE Access*, Vol. 13, pp. 190701–190711, 2025.

[5] Theresa T. Kessler, Cintya Larios, Tiffani Walker, Valarie Yerdon, and P. A. Hancock. A comparison of trust measures in human-robot interaction scenarios. In Pamela Savage-Knepshield and Jessie Chen, editors, *Advances in Human Factors in Robots and Unmanned Systems*, pp. 353–364. Springer International Publishing, 2017.

[6] 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.

[7] John D. Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, Vol. 40, No. 1, pp. 153–184, 1994.

[8] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, Vol. 46, No. 1, pp. 50–80, 2004.

[9] Bertram F. Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. In Chang S. Nam and Joseph B. Lyons, editors, *Trust in Human-Robot Interaction*, pp. 3–25. Academic Press, 2021.

[10] Kiana Jafari Meimandi, Matthew L. Bolton, and Peter A. Beling. Action over words: Predicting human trust in ai partners through gameplay behaviors. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 563–568, 2024.

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

[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] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. The development of overtrust: An empirical simulation and psychological analysis in the context of human – robot interaction. *Frontiers in Robotics and AI*, Vol. 8, , 2021.

[14] Anqi Xu and Gregory Dudek. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*, pp. 221–228, 2015.