第 30 回日本バーチャルリアリティ学会大会論文集（2025年9 月）

特定の人物模倣 AI の設計と違和感の評価

Design and Evaluation of Person-Specific Generative Agents:
Assessing Naturalness in Human Imitation Systems

西尾拓真 [1)]，大野凪咲 [1)]，宮下敬宏 [2)3)]，安藤英由樹 [1)3)]
Takuma NISHIO, Nagisa ONO, Takahiro MIYASHITA and Hideyuki ANDO

1）大阪芸術大学大学院 芸術研究科 芸術学専攻 （〒585-0001 大阪府南河内郡河南町東山 469）
2）国際電気通信基礎技術研究所 （〒619-0237 京都府相楽郡精華町光台 2 丁目 2-2）
3）大阪芸術大学 芸術学部 アートサイエンス学科 （〒585-8555 大阪府南河内郡河南町東山 469）

概要：物語キャラクターとの自然な対話体験の実現を目指し，その前段階として実在人物の「その人物らしさ」を再現する人物模倣 AI を構築し評価した．既存の人物模倣手法で筆者自身の模倣 AI を制作し，心理尺度により定量評価を実施した．履歴書的な人物らしさは高精度で再現可能だが，状況依存的な人物らしさの再現には課題があることが判明した．本報告は認知科学的知見に基づく設計指針を提案し，現在の技術の到達点と方向性を示す．
キーワード：人物模倣，音声対話型 AI，人物らしさ

1. はじめに

物語に登場する好きなキャラクターと実際に接してみたいという願望があるユーザは少なくない．一方で昨今の生成 AI 技術や合成音声モデルを用いることで，セリフや行動の履歴，発話や行動の描写，設定資料などから，そのキャラクターがどんな状況で何を言うか，何をするかをある程度推測でき，ユーザと擬似的に話すような体験を作ることが可能になった．しかし，その体験の質は「その人物らしさ」にどれだけ忠実かに依存する．そのキャラクターが好きで，自身の中の「その人物らしさ」が具体的であるほど，その違和感に敏感になってしまう．声色やキャラクターの描画などにおいては技術の改善により向上すると考えられるが，発話内容や行動に対する違和感はどのように生じているのだろうか？また，この違和感を低減させるため方法を検討するためにはどのような方策と取ればよいのか？これを明らかとするために我々はまず，実際の人物で「本人の応答」と「模倣 AI の応答」とを比較することで「その人物らしさ」についての知見を深める必要があると考えた．

本報告ではその予備検討として，既存の手法による個人の性格・信念・話法を模倣するエージェントアーキテクチャを用いて著者本人の発言内容を再現する対話システムを制作した．そしてそれを実際に使用してみることで著者本人との比較対象として妥当なレベルで「本人らしさ」を備えているかを検討し，現状の手法に不足している要素について考察する．

2. その人物らしさはどこにあるのか？

人物の模倣において「その人物らしさ」を考える場合，2 つの観点が考えられる．1 つめの観点は，声の物理的特徴，2 つめの観点は，話し方や口調といった発言内容である．

声の物理的特徴について，現在の音声合成技術はどこまで本人の声を模倣でき，それを他者が聞き分けられるかという点が重要である．近年の音声合成技術では，VITS2[1]をベースに文脈に応じたイントネーションを生成可能にした BERT-VITS2[2]が登場している．その性能評価では，話者内一貫性が 95%以上の水準を達成し，MOS (Mean Opinion Score) 評価において合成音声と本物の音声に統計的有意差がなかったことが報告されている[3]．このことから，特定の人物とその人物の声を学習した生成音声が同じ文章を同じようなイントネーションで読み上げた場合，聞き分けは困難であることがわかる．

次に発言内容について考えると，我々が普段話している内容における「その人物らしさ」は 2 つの階層に分かれていると考えられる．

第一の階層は，プロフィール・履歴書などの情報に基づく人物らしさである．これは経歴などから推測される統計学的・分類学的な性質，基本的な価値観・人生哲学など，比較的客観的で静的な情報から構成される「その人らしさ」である．

第二の階層は，特定の状況や人物などの対象と直面しているときに変化し，対象との関係性によって深化・蓄積され形成される人物らしさである．これは日常的な関係性の中で逐次形成される，動的な「その人らしさ」である．

第一の階層（履歴書的な人物らしさ）については，既存研究[4]により高精度での模倣が実現されている．LLM（大規模言語モデル）に入力するプロンプトを工夫することで，特定の個人を模倣し，自然な音声・内容で会話できる音声対話型 AI の開発が現実的なものとなりつつある．Park et al. (2024) は，実際の個人の性格・信念・話法を高い精度で再現するエージェントの生成的アーキテクチャ[4]を示した．この人物模倣アーキテクチャは，多岐にわたる質問項目を含む General Social Survey (GSS) [5]において，本人と比較して予測精度 85%という高い数値を達成している．このことから，履歴書的な情報を回答する面接のようなやり取りについては，既存の手法によってかなりの程度模倣可能であることがわかる．

しかし，第二の階層（状況依存的・関係性に基づく人

物らしさ）については課題が残る．我々が他者に対して持つ「その人物ならこの状況でこうするだろう」や「あの人に対してはこんなことを言いそう」といった予測は，単なる性格特性の把握を超えて，相互作用の履歴や文脈に依存した複雑な認知処理を必要とする．前述の人物模倣アーキテクチャでは，この部分を十分にカバーできていないと考えられる．したがって，まず既存の理論研究と照らし合わせることで，我々が日常的に行っている対人予測のメカニズムを明らかにし，それを踏まえて人物模倣アーキテクチャが実現できている範囲を考察する必要がある．

3. 人物を予測するメカニズム

そもそも，我々が他者とコミュニケーションを行う上で感じる「本人らしさ」とは何なのだろうか．特に前節で述べた第二の階層（状況依存的・関係性に基づく人物らしさ）を我々はどのように認識・予測しているのだろうか．

Cantor & Mischel (1979) は，人物認知における特性プロトタイプ（trait prototypes）を提唱し，人々が「外向的な人」「誠実な人」といった抽象的な人格カテゴリーの典型例を認知的基準として使用することを実証した[6]．

実際の対人関係における個別化された認知処理については，Hastie et al.（1980）が人物記憶（person memory）として詳しく論じている[7]．特に興味深いのは，個人の印象と矛盾する行動情報は一貫性のある行動よりもよく記憶されるという「非一貫性効果」の発見である．これは，人々が他者についての認知表象を動的に更新していることを示している．

Bartlett のスキーマ理論[8]に基づけば，個人に関する様々な情報を統合した人物スキーマ（person schema）が形成され，新たな対人相互作用における予測的な認知枠組みとして機能する．この認知メカニズムこそが，2 章の第二の階層（状況依存的・関係性に基づく人物らしさ）で述べた「その人物ならこの状況でこうするだろう」という予測を可能にする基盤となっている．

社会認知神経科学の発展により，人物認知はベイズ推論プロセスとして理解され，脳が他者に関する事前信念（prior beliefs）を感覚的証拠で継続的に更新する予測システムとして捉えられている[9]．「本人らしさ」の知覚は予測誤差処理の産物であり，人物スキーマに基づく予測と実際の観察との乖離度が予測誤差信号として処理される[10]．予測と実際の行動が一致する場合は「その人らしい」という一貫性を感じ，乖離が生じると「らしくない」という違和感として現れる．この予測誤差の大きさが，対人コミュニケーションにおける「らしさ」の知覚を決定的に左右すると考えられる．

4. 人物模倣 AI の設計指針

3 章で述べた人物を予測するメカニズムから，人物模倣 AI が「本人らしさ」を再現するために必要な設計指針を導出し，Park et al. (2024) の生成的アーキテクチャ[4]によってどこまで可能なのかと，不足している事項について検討を進めた．

Park et al. (2024) の生成的アーキテクチャ[4]は以下の 3 つの構造を持つ：

(1) American Voices Project[11]のプロトコルに基づく 2 時間のインタビュー文字起こし

(2) 心理学・社会学等の専門家ペルソナ AI によるメタ視点の分析情報

(3) Chain-of-Thought 法による段階的推論テンプレート

これらの構造が，対人認知メカニズムから導かれる以下の設計指針をどの程度実現しているかを検証する．

指針 1：人物スキーマに基づく予測可能性の確保

Bartlett のスキーマ理論[8]に基づけば，受け手は過去の相互作用から形成された人物スキーマを通じて行動や発言を予測する．同手法は，包括的なインタビューの文字起こしを入力することで，個人の価値観や性格特性から一貫性のある応答を生成できる．この点において，受け手の人物スキーマに適合する予測可能な応答パターンの実現が可能である．

指針 2：多層的人物表象の構築

Cantor & Mischel の特性プロトタイプ理論[6]によれば，人物認知は抽象的な人格カテゴリーから個人特異的な認知構造への階層的処理として機能する．同手法は，入力に専門家ペルソナによるメタ認知的分析を含めることで，深層的な特性抽出を実現している．これにより，受け手が抱く「その人らしさ」の期待との整合性を一定程度確保できる．

指針 3：個人特異的認知パターンの再現

Hastie et al.の人物記憶研究[7]に基づき，その人物特有の思考プロセスを再現する必要がある．同手法は，Chain-of-Thought 法による 4 段階の推論（選択肢理解→潜在的動機の推論→総合的判断→最終応答）を採用し，体系的な思考プロセスを模倣する．これにより，文脈に適応した自然な対話の流れを維持できる．

指針 4：認知表象の動的更新機能

Hastie et al.の非一貫性効果の知見[7]によれば，人々は他者についての認知表象を対話を通じて継続的に更新する．しかし，同手法はインタビューデータから構築された静的モデルに依存するため，対話中に得られる新情報による認知表象の更新ができない．長期的な対話において「本人らしさ」を保ちながら新たな側面を統合する能力が欠如している．

指針 5：予測誤差に基づく適応的調整

社会認知神経科学のベイズ推論モデル[9]によれば，「本人らしさ」の知覚は予測誤差処理の産物である．同手法は各応答を独立して生成するため，受け手の反応から予測誤差を検出し，応答パターンを調整する機能を持たない．相手の期待との乖離を検出して修正する能力が欠如しており，対話の自然な流れの中での微妙な調整ができない．

以上の分析から，同手法の人物模倣アーキテクチャ[4]は指針 1〜3 については一定程度の対応が可能であるが，指針 4・5 に関しては根本的な機能拡張が必要であることが明らかになった．特に，動的な認知更新と予測誤差に基づく適応的調整は，「本人らしい」対話を実現するための重要な要素でありながら，現在の静的なプロンプトベースのアプローチでは実現困難である．

本報告では，筆者自身をモデルにこのアーキテクチャ用いた人物模倣 AI を作成し，評価を行うことで，理論的分析で明らかになった限界が実際の使用場面でどのように現れるかを検証した．特に，第一の階層（履歴書的な人物らしさ）と第二の階層（状況依存的・関係性に基づく人物らしさ）の再現度の差異に着目し，現在の技術で実現可能な「本人らしさ」の範囲と限界を具体的に明らかにすることを試みた．

5. 実際の制作手法と評価手法

筆者自身の人物模倣 AI を制作するにあたって，文献[4]の手法に従い，American Voices Project[11]のインタビュープロトコルの質問内容を Claude Sonet 4 モデル[12]を用いて日本の大学生向けに翻訳・修正し，2 時間程

度のセルフインタビューを実施した．また，同著内記載のプロンプトを使用することで，心理学者，社会行動学者，政治学者，人口統計学者のそれぞれの専門家ペルソナを付与したClaude Opus 4モデル[12]にこのインタビュー全体の文字起こしを入力し，専門家による考察を生成した．これらインタビューの全体の文字起こしと専門家による考察を同著内記載の Chain-of-Thought プロンプトに挿入し，人物模倣アーキテクチャを再現した．

生成応答の妥当性を検討するため，複数のモデルを用いた LLM による出力内容の比較を行う．モデルは GPT-4o（OpenAI）[13]，Claude 4 Opus（Anthropic）[12]とし，それぞれに同一のプロンプトを入力することで出力の差異を分析する．なお，Anthropic のモデルに対しては 66000 字程度のプロンプトを直接入力し，OpenAI のモデルに対しては文字数制限のため，インタビュー全体の文字起こしと専門家による考察をそれぞれ別のファイルとしてベクトルデータベースに登録して比較を行った．

5.1 評価手法

制作した人物模倣 AI の「本人らしさ」を多面的に評価するため，以下の 3 つの心理尺度による定量的評価を採用した：

1. Big Five 性格検査（Big Five 尺度短縮版）[14]
2. 認知的完結欲求尺度（Need for Closure Scale 日本語版）[15]
3. 一般システム正当化尺度（General System Justification Scale）[16]

これら 3 つの尺度を選定した理由は，（1）時間的安定性が高く第一の階層（履歴書的な人物らしさ）の特徴と合致すること，（2）Park et al.が評価に用いたGSS[5]の質問項目と概念的に重複する領域をカバーしていること，（3）標準化された尺度であり信頼性・妥当性が確立されていることである．筆者本人と模倣 AI の両方に同一の質問項目を提示し，その回答の一致度を各因子・下位尺度ごとに比較することで，安定的な個人特性の再現精度を定量的に評価した．

6. 結果と考察

本報告で用いた評価指標の算出方法を以下に示す．完全一致率は本人と同じ回答をした項目数を全項目数で除した値，類似率は本人の回答±1 の範囲内の回答を類似とみなし，類似項目数を全項目数で除した値として算出した．

類似率を導入した理由は，心理尺度における隣接選択肢間の心理的距離は必ずしも大きくなく，人間の回答にも測定誤差や状況による揺らぎが存在するためである．この指標により，厳密な完全一致よりも実用的な「その人物らしさ」の再現度を評価できると判断した．

相関係数は Park et al.（2024）[4] に準拠し，個人レベル分析を実施した．Big Five 性格検査では 5 つの因子スコア間（N=5），認知的完結欲求尺度では 3 つの因子スコア間（N=3）でピアソン相関係数を計算した．具体的には，各心理尺度について筆者本人のスコアベクトルと人物模倣 AI のスコアベクトル間の相関を算出した．

筆者自身を対象とした人物模倣 AI の評価結果から，既存の人物模倣アーキテクチャが持つ可能性と限界が明らかになった（表 1）．全体的な傾向として，両モデルとも類似率では 80%以上の高い値を示したが，完全一致率では大きな差が見られた．この結果は，現在の人物模倣技術が個人の大まかな傾向や特性は捉えられているものの，より繊細な個人差の再現には課題があることを示唆している．

特筆すべきは，GPT-4o と Claude Opus 4 が異なる強みを示したことである．GPT-4o は完全一致率で優れた結果を示し（全体 51.7%），特に誠実性（71.4%）と開放性（66.7%）で高い精度を達成した．一方，Claude Opus 4 は相関係数で圧倒的に優れた結果を示し（Big Five: r=0.874，認知的完結欲求: r=0.923），全体的な主観評価の再現において本人との高い一貫性を示した．これは個人の認知的傾向の大局的なパターンを捉えることができていることを示す．

Park et al.（2024）の研究[4]では，1,052 名の参加者に対して Big Five での正規化相関 0.80 を報告している．本研究の Claude Opus 4 の相関係数（Big Five: r=0.874）は，Park et al.の大規模サンプルでの結果（生の相関 r=0.78）を上回っており，単一事例においても高い再現性が可能であることを示している．

筆者本人との比較において大きな差異が出た箇所として，協調性では両モデルとも完全一致率 16.7%と極めて低く，対人関係に関わる微妙な特性の再現が困難であることが示された．認知的完結欲求尺度の「予測可能性に対する選好」では両モデルとも完全一致率 0%であり，状況依存的な判断の模倣に失敗していることがわかる．

これらの結果から，第一の階層（履歴書的な人物らしさ）については，類似率 80%以上という形で一定程度再現可能であることが示された．GPT-4o の表層的精度と Claude Opus 4 のパターン一貫性という異なる強みの存在は，将来的なモデル統合の可能性を示唆している．

これは，物語に登場するキャラクターを違和感なく AI で模倣するという本研究の目的において，キャラクターの基本設定（性格，口調等）に相当する第一の階層は現在の技術である程度の再現精度が実現できることの示唆となる．特に，原作者や物語が直接的に言及しないようなキャラクターの資質・特性を既存の心理学的指標に当てはめて定量化することが可能になり，キャラクター性を分類学的に体系化できる可能性がある．これにより，定量化された特性カテゴリーとユーザの期待との乖離を予測誤差として検出し，リアルタイムで応答パターンを調整する適応的機能への応用可能性が示唆される．

しかしながら，4 章で指針 4，5 と対応する機能の欠如について言及したように，第二の階層（状況依存的・関係性に基づく人物らしさ）の再現は，現在の静的アーキテクチャでは困難であることが明確化された．認知的完結欲求尺度の低い精度は，その人物特有の思考プロセスや判断基準の動的な側面が捉えられていないことを示す．特定の尺度（協調性，予測可能性選好）での極端に低い精度は，個人特異的な認知パターンへの対応が不十分であることを示唆している．

表 1：本人と人物模倣 AI の心理尺度における一致度比較

尺度/因子	完全一致率		類似率(±1の範囲)		相関係数	
	vs GPT	vs Claude	vs GPT	vs Claude	vs GPT	vs Claude
BigFive(全体)	51.7	27.6	86.8	82.8	0.687	0.874
-情緒不安定性	40	40	100	100		
-外向性	60	20	80	80		
-開放性	66.7	33.3	100	100		
-調和性	16.7	16.7	66.7	83.3		
-誠実性	71.4	28.6	85.7	57.1		
認知的完結欲求尺度(全体)	20	10	45	50	0.406	0.923
-秩序に対する選好	28.6	0	28.6	28.6		
-予測可能性に対する選好	0	0	60	40		
-決断性	25	25	50	75		
システム正当化尺度(全体)	25	37.5	62.5	75		

7. おわりに

本報告では，物語のキャラクターと違和感なく対話できる体験の実現を目指し，その前段階として実在人物の「本人らしさ」を再現する対話システムの制作と評価を行った．Park et al.（2024）の人物模倣アーキテクチャ[4]を用いた評価の結果，第一の階層（履歴書的な人物らしさ）については一定の再現が可能である一方，第二の階層（状況依存的・関係性に基づく人物らしさ）の再現には根本的な機能拡張が必要であることが明らかになった．

本報告で明らかになった課題を 3 章で述べた人物を予測するメカニズムと照らし合わせることで，重要な設計要素が浮かび上がる．第一に，認知表象の動的更新機能の必要性である．人々は対話を通じて他者についての認知表象を継続的に更新するが，同手法は静的な個人モデルに基づくため，対話中に得られる新情報による認知更新ができない．これに対し，長期記憶を RAG（Retrieval-Augmented Generation）＋ベクトルデータベースに動的に追加し，短期記憶を会話ログで補完することで，認知更新のような振る舞いを実装できる可能性がある．

第二に，予測誤差に基づく適応的調整機能の必要性である．同手法では，各応答を独立して Chain-of-Thought で生成するため，相手が困惑している様子や会話が停滞している状況でも同じパターンで推論を続けてしまう．これに対し，会話履歴に基づく重み付けフィードバックを推論プロンプトに動的に反映させることで，状況に応じた適応的な応答生成が可能となる．

本報告の知見から，「本人らしい」対話体験の実現には，以下の 3 段階のアプローチが必要であることが示唆される：

第 0 段階：物理的同一性の検証

まず前提として，物理的に同じ声質・同じ発話内容であっても本人と模倣 AI の判別が可能かどうかの検証が必要である．現在の音声合成技術は高品質であるが，微細な韻律や間の取り方などに違いが残る可能性がある．この段階では，適切な制約条件（例：読み上げタスク，定型的な応答）を設定することで，物理的な識別可能性を最小化できる可能性がある．

第 1 段階：第一の階層の実現

本研究で検証した段階である．履歴書的・プロフィール的な情報に基づく応答の再現は，現在の技術でも類似率 80%以上で実現可能である．Park et al.の研究が示すように，インタビューのような構造化された状況では高い予測精度を達成できる．これは質問者と回答者の役割が明確で，文脈が限定されているためである．

第 2 段階：第二の階層への拡張

状況依存的・関係性ベースの「その人らしさ」の実現には，前述の動的更新機能と適応的調整機能の実装が不可欠である．この段階では，対話の文脈や相手との関係性に応じて柔軟に応答を変化させる能力が求められる．

本研究は第 0 段階を前提として第 1 段階の検証を行ったが，実用的な観点からは，これらの段階を統合的に考慮する必要がある．物語のキャラクターとの対話においては，まず制約された状況（特定のシナリオ内での会話，役割が明確な対話場面）から始め，段階的に制約を緩和していくアプローチが現実的である．各段階で必要となる技術要素と制約条件を体系的に整理し，段階的な実装と評価を進めることが，違和感のない対話体験の実現への道筋となるだろう．

謝辞　この成果の一部は JST CREST(JPMJCR22P4)による．

参考文献

[1] Kong, J., Park, J., Kim, B., Kim, J., Kong, D., & Kim, S. (2023). VITS2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. In Proceedings of Interspeech 2023 (pp. 4374-4378).

[2] fishaudio. (2023). Bert-VITS2: VITS2 backbone with multilingual-bert [Computer software]. GitHub. https://github.com/fishaudio/Bert-VITS2

[3] Zackary Rackauckas, Julia Hirschberg. (2025). Benchmarking expressive Japanese character text-to-speech with VITS and Style-BERT-VITS2. *arXiv preprint arXiv:2505.17320*.

[4] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative Agent Simulations of 1,000 People. arXiv preprint arXiv:2411.10109.

[5] NORC at the University of Chicago. (1972-present). General Social Survey. https://gss.norc.org/

[6] Cantor, N., & Mischel, W. (1979). Prototypes in person perception. Advances in Experimental Social Psychology, 12, 3-52.

[7] Hastie, R., Ostrom, T. M., Ebbesen, E. B., Wyer, R. S., Jr., Hamilton, D. L., & Carlston, D. E. (Eds.). (1980). Person memory: The cognitive basis of social perception. Hillsdale, NJ: Lawrence Erlbaum Associates.

[8] Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. Cambridge University Press.

[9] Clifford CW, Mareschal I, Otsuka Y, Watson TL. (2015). A Bayesian approach to person perception. Conscious Cogn. 36:406-13.

[10] Van Overwalle F. (2009). Social cognition and the brain: a meta-analysis. Hum Brain Mapp. (3):829-58.

[11] Stanford Center on Poverty and Inequality & Princeton University. (2018-present). American Voices Project. https://inequality.stanford.edu/american-voices-project

[12] Anthropic. (2025). Introducing Claude 4. https://www.anthropic.com/news/claude-4

[13] OpenAI. (2024). GPT-4o. https://openai.com/index/hello-gpt-4o/

[14] 並川努，谷伊織，脇田貴文，熊谷龍一，中根愛，野口裕之．(2012)．Big Five 尺度短縮版の開発と信頼性と妥当性の検討．心理学研究，83(2)，91-99.

[15] 鈴木公基，桜井茂男．(2003)．認知的閉鎖欲求尺度の作成と信頼性・妥当性の検討-Need for Closure Scale の邦訳-．心理学研究，74(3)，270-275.

[16] Kay, A. C., Jost, J. T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. Journal of Personality and Social Psychology, 85(5).