LLM 媒介仮想実験法の妥当性ランドスケープ：4 軸タクソノミーと経験的状態の体系的レビュー（2022-2026）

The Validity Landscape of LLM-Mediated Virtual Experimentation: A Four-Axis Taxonomy and Empirical Status Synthesis (2022-2026)

著者: Anonymous Author (LLM-authored draft, prepared for AIxiv submission) **版**: v1.0 **日付**: 2026-04-29 **Submission Target**: aiXiv (https://aixiv.science/), secondary: arXiv (subject to endorsement) **全データ**: 公開文献調査（18 検索クエリ、約 90 anchor papers）+ ARC 自律走行による独立合成（rc-20260428-213351-42de41, stage 1-9） **Recursive disclosure**: 本論文は LLM 著の合成として書かれており、その認識論的含意を §13 Limitations で論じる

Abstract (English)

Virtual experimentation—the use of large language models (LLMs) as synthetic respondents, populations, or scenario generators in psychological and social science research—expanded explosively after Argyle et al.'s (2023) "silicon sample" proposal. Subsequent empirical work (Bisbee et al. 2024; Boelaert et al. 2025; Hullman et al. 2026) has identified systematic methodological limitations: low variance, topic-specific machine bias, prompt sensitivity, WEIRD population skew, and algorithmic monoculture. Yet integrative reviews that map which validity claims hold under which simulation strategy for which epistemic purpose remain absent, leaving practitioners and reviewers without a shared evaluative framework.

We propose a **four-axis taxonomy** spanning (1) Simulation Layer (LLM-as-Respondent / LLM-as-Population / LLM-as-Scenario), (2) Validity Strategy (Face / Benchmark / Mechanistic-Process / Stress-Adversarial / Participatory-Ecological), (3) Epistemic Goal (Prediction-Substitution / Theory Probing / Design Exploration / Normative Representation), and

(4) Empirical Status (Established / Contested / Frontier / Empirical-Void). We apply this taxonomy systematically to 18 representative studies, producing an empirical-status matrix that identifies three regions: (a) a saturated region (LLM-as-Respondent for survey-style aggregate prediction, contested but well-charted), (b) a frontier region (LLM-as-Scenario for design exploration and mechanism probing, promising but under-tested), and (c) an empirical-void region (normative representation of silent majorities, future generations, and marginalized groups). We derive four falsifiable hypotheses with pre-registered prediction thresholds and rejection criteria, and articulate eight honest-design principles, each operationalized as a 3-5 point reporting checklist with concrete thresholds. We address known reviewer critiques (recursive limitation, selection bias, search transparency) by pre-emptive disclosure and conclude with implications for participatory design, research ethics, and the representation of voices that cannot speak for themselves.

Keywords: virtual experimentation, silicon sample, algorithmic fidelity, LLM-as-Population, LLM-as-Scenario, machine bias, statistical calibration, validity taxonomy, participatory design, voice representation, future generations, honest research design

要旨（日本語）

仮想実験法（virtual experimentation）――大規模言語モデル（LLM）を合成回答者・集団・場面生成器として心理学および社会科学の実証研究に用いる実践――は Argyle et al. (2023) の "silicon sample" 提唱以降、爆発的に拡大した。2024-2026 年の体系的検証（Bisbee et al. 2024; Boelaert et al. 2025; Hullman et al. 2026）はこの実践に方法論的限界を明らかにしてきた：低分散性、トピック横断的に方向が変動する machine bias、プロンプト敏感性、WEIRD 集団偏向、algorithmic monoculture。しかし、**いかなる妥当性主張がいかなる simulation 層 × いかなる認識論的目的の下で支持されうるか** を統合的に整理したレビューは不在であり、実装者・査読者・編集者が共有可能な評価フレームワークが欠如している。

本レビューは **4 軸タクソノミー** を提案する：(1) Simulation Layer（LLM-as-Respondent / LLM-as-Population / LLM-as-Scenario）、(2) Validity Strategy（Face / Benchmark / Mechanistic-Process / Stress-Adversarial / Participatory-Ecological）、(3) Epistemic Goal（Prediction-

Substitution / Theory Probing / Design Exploration / Normative Representation）、(4) Empirical Status（Established / Contested / Frontier / Empirical-Void）。このタクソノミーを 18 件の代表的研究に体系的に適用し、3 つの領域を同定する：(a) **Saturated 領域**——LLM-as-Respondent による集団平均値予測、(b) **Frontier 領域**——LLM-as-Scenario による設計探求とメカニズム探索、(c) **Empirical-void 領域**——サイレントマジョリティ・未来世代・周縁化集団の normative representation。

さらに 4 つの **falsifiable hypotheses**（pre-registered な予測閾値と棄却基準付き）と、8 つの **operational checklists 付き honest design principles**（thresholds 付与）を導出する。査読者からの予想される 4 つの批判（実証検証不足、原則の高水準性、selection bias、recursive limitation）に対し pre-emptive な対応を施し、参加型デザイン・研究倫理・声を持たない者の代弁への含意を論じる。

キーワード: 仮想実験法、シリコンサンプル、algorithmic fidelity、LLM-as-Population、LLM-as-Scenario、machine bias、statistical calibration、妥当性タクソノミー、参加型デザイン、声の代弁、未来世代、honest research design

---

1. 緒論

1.1 仮想実験法の急速な拡大と妥当性論争の同時進行

2023 年以降、Large Language Models (LLMs) を心理学・社会科学の実証研究の構成要素として用いる実践——本稿で **仮想実験法 (virtual experimentation)** と総称する——が爆発的に拡大した。Argyle et al. (2023) "Out of one, many: Using language models to simulate human samples" が **silicon sample** および **algorithmic fidelity** という用語を確立し、Aher, Arriaga & Kalai (2023) の Turing Experiments、Horton et al. (2023) の "Homo Silicus"、Park et al. (2024) の 1,000 人 generative agent simulation が学術出版において日常化した。同時に、業界（市場調査）でも synthetic respondents の採用が急進し、複数のアナリストレポートは 2027 年までに市場調査 input の 50% 以上が synthetic に移行すると予測している (NielsenIQ 2024)。

この拡大と並走して、批判的検討も蓄積した： - **Bisbee et al. (2024)** Political Analysis: ChatGPT 合成被験者は ANES 平均値を再現するが、分

散・回帰係数・プロンプト不変性・時間安定性のいずれにおいても unreliable。 - **Santurkar et al. (2023)** ICML: LMs の意見と米国デモグラフィック群の意見に「気候変動の党派分極」並みの不一致。 - **Sarstedt et al. (2024)** Psychology & Marketing: WEIRD 国で性能が高く、非 WEIRD 国で急落。 - **Boelaert et al. (2025)** Sociological Methods & Research: トピックごとに方向がランダムに変動する **machine bias** を実証。 - **Hullman et al. (2026)**: 仮想実験の妥当性確保には **heuristic interchangeability** と **statistical calibration** の二戦略があり、前者は exploratory のみ、後者のみが confirmatory に拡張可能。

1.2 既存批判文献の分散と統合不足

これらの批判文献は政治学（Bisbee, Boelaert）、機械学習（Santurkar, Liu）、マーケティング（Sarstedt）、HCI/NLP メソドロジー（Hullman）の各分野に分散しており、横断的に統合した review は限定的である。とりわけ、**「いかなる妥当性主張がいかなる simulation 戦略の下で何を主張できるか」** を体系的にマッピングした taxonomy は不在であり、結果として:

- 実装者は「どの程度まで主張してよいか」を判断する根拠がない
- 査読者は領域横断的な評価基準を持たない
- 編集者は spectrum 上のどこに掲載許容点を置くかを決められない
- 大学院生は学習路線として何から学ぶべきかが不明

特に近年（2024-2026）顕在化した次の課題群への統合的整理が不足している: 1. simulation 層（個別回答 vs 集団分布 vs 仮想場面）による妥当性の **非一様性** 2. 検証戦略（face vs benchmark vs mechanistic vs stress vs participatory）による主張の **射程差** 3. 認識論的目的（substitute vs probe vs explore vs represent）による **適切な評価基準の差** 4. 経験的状態（established vs contested vs frontier vs void）の **領域別 mapping**

1.3 本レビューの目的と貢献

本レビューは以下の貢献を目指す：

1. **4 軸タクソノミーの提案**: 上記 4 次元を直交軸として構築し、仮想実験法の研究空間を体系化する。
2. **18 文献への体系的適用**: 各論文を 4 軸の組み合わせで分類し、empirical status matrix を生成する。
3. **3 領域の同定**: Saturated（飽和）/ Frontier（フロンティア）/ Empirical-Void（実証空白）。

4. **4 つの falsifiable hypotheses**: 各仮説に pre-registered な予測閾値と棄却基準を付随する。
5. **8 つの operational design principles**: 各原則を 3-5 点 checklist + thresholds に展開する。
6. **参加型デザイン・倫理との接続**: silent majority、future generations、marginalized groups の voice representation 問題への含意を論じる。

1.4 用語の事前整理

本レビューでは以下の用語を厳密に区別する：

- **仮想実験法 (virtual experimentation)**: LLM が回答者・集団・場面のいずれかを生成する形で実証研究の被験段に組み込まれる手法の総称。
- **silicon sample**: Argyle et al. (2023) の用語。LLM を sociodemographic backstories で条件付けして得られる合成回答群。
- **algorithmic fidelity**: 合成回答が実集団の応答分布をどの程度反映するかの程度概念。
- **machine bias** (Boelaert et al. 2025): LLM がトピックごとに方向の変動するシステマティックでない歪みを示す現象。「左寄り」など単純化できない。
- **algorithmic monoculture** (Bommasani et al. 2022): 同一基盤モデルが複数の意思決定者に共有される構造的リスク。
- **heuristic interchangeability vs statistical calibration** (Hullman et al. 2026): 前者は prompt engineering で「シミュレーションと人間が交換可能」を目指す（探索的）、後者は人間データで差異を統計調整（confirmatory に拡張可能）。
- **WEIRD**: Western, Educated, Industrialized, Rich, Democratic。心理学の集団偏向を指す古典的概念。

---

2. 方法

2.1 探求の構造

本レビューは 4 つの探求軸 × 7 の関連分野のマトリクス上で文献を収集した：

探求軸: 1. JTBD (Jobs-to-be-Done): どんな研究実践がなぜ仮想実験法を「採用」するか 2. 歴史的経緯: どのような知的伝統の延長か 3. 現状と限界: 2022-2026 の経験的所見 4. ありたい姿: design principles の方向性

関連分野: 政治学 / 計算社会科学 / 心理測定学 / マーケティング / HCI・NLP メソドロジー / 参加型デザイン・熟議民主主義 / 研究倫理・IRB

2.2 検索戦略（全開示）

Anthropic WebSearch（Google ベース）で計 18 のクエリを 7 系統に分けて実施。各クエリと採択結果を §A1 Appendix に開示する。

系統	クエリ数	主要対象
Foundational（提唱論文）	4	Argyle 2023, Park 2024, Aher 2023, Horton 2023
Critical（限界研究）	5	Bisbee 2024, Santurkar 2023, Boelaert 2025, Sarstedt 2024, Dillion 2023
Mechanism（プロンプト・ペルソナ）	2	Sclar 2024, Liu 2024
Causal（counterfactual）	1	LLM counterfactual reasoning ベンチマーク
Methodological framework	1	Hullman et al. 2026
Adjacent: Vignette / ABM / GSS	2	Rossi 1982, Epstein 2006, GABM 2024
Participatory / Ethics	3	Plurality, Future Generations, IRB AI

採択基準: - 査読付き学術誌（Political Analysis, ICML, NeurIPS, Nature, Trends in Cognitive Sciences, Sociological Methods & Research） - 主要 NBER/SSRN ワーキングペーパー - arXiv プレプリント（最新動向把握用、補助） - 業界レポート（市場調査領域、別カテゴリ）

2.3 ARC との parallel synthesis

本レビューは独立に 2 つの統合経路を経た: 1. **WebSearch ベース** (本論文の主軸): 18 クエリ → 約 90 文献 → 手動コーディング → 4 軸タクソノミー 2. **ARC v0.3.1 自律走行** (rc-20260428-213351-42de41): topic prompt → ARC が独立に goal/synthesis/hypothesis を生成 → 結果を本レビューと交差検証

ARC の独立な synthesis（stage 7, 521 行）が **同じ 4 軸タクソノミーと 3 領域分類に到達した** ことは、本論文の概念構造が単なる著者の解釈バイアスではなく、現存する文献の構造的特徴を捕捉していることの傍証である。両経路の合意点・対立点は §10 で詳述する。

ARC の stage 10 (CODE_GENERATION) が review-synthesis 型 topic と experiment-pipeline 前提パイプラインの構造的不整合により失敗したこと自体も、本論文の主張する **「妥当性はタスク次第」** の実例であり、§13.5 で recursive limitation の文脈で論じる。

2.4 統合の手順

文献群に対し以下の質的処理を経て本論文を執筆：

1. **Open coding**: 各文献の主張・限界を 5-15 字の code で抽出
2. **Axial coding**: 4 軸 × 7 領域マトリクスに code を配置
3. **Tension mapping**: 文献間の対立点・補完点・収斂点を可視化
4. **Status assignment**: 4 軸の各セルに Empirical Status を割り当て、合意の程度を記述
5. **Synthesis**: 3 領域分類 + 4 hypotheses + 8 design principles に整理

2.5 限界の事前明示

本レビューは以下の限界を持つ（詳細は §13 で論じる）： - **言語**: 英語と日本語の文献を主とし、その他言語は非対象。検索 "仮想実験 LLM" は 0 件で日本語文脈の文献が極めて薄いことを §13.2 で論じる。 - **データベースアクセス**: 公開 web 検索によるのみ。 - **時間範囲**: 2022-2026 が中心。古典的方法論文献（vignette, factorial survey, ABM）は anchor papers から逆引きで補完。 - **本論文自体の合成性**: §13.5 で詳述するが、本論文は LLM 著であり、これが reviewing virtual experimentation という主題と recursive な関係を持つ。

---

3. 歴史的経緯：5 つの転回

仮想実験法は突如出現した断絶ではなく、複数の知的伝統の延長線上にある。本節では 5 つの転回を整理する。

3.1 第 1 転回: Vignette / Factorial Survey（1980s）

Rossi & Anderson (1982) が制度化した factorial survey design は、**仮想記述（vignette）の複数次元を直交化** して回答者にランダム割付し、社会的判断構造を測定する手法である。Wallander (2009) "25 years of factorial surveys in sociology" が到達点をレビュー、Hainmueller, Hangartner & Yamamoto (2015) PNAS が vignette/conjoint の現実行動との整合を検証した。

LLM 仮想実験との接続: 古典的 vignette は「人間が仮想状況に答える」のに対し、LLM 仮想実験は **「vignette を生成し、回答者も合成する」二重仮想化** である。この shift により外部妥当性の前提が変質する。

3.2 第 2 転回: Generative Social Science / ABM（1990s-2000s）

Schelling (1971) segregation model に始まり、Axelrod (1984), Epstein & Axtell (1996) Sugarscape を経て、Epstein (2006) "Generative Social Science" が **「If you didn't grow it, you didn't explain it」** という方法論的綱領を確立した。Lazer et al. (2009) Science の "Computational Social Science" 宣言が領域を制度化。

LLM 仮想実験との接続: 現代の LLM 仮想実験は、ルールベースの ABM が **LLM 駆動エージェントに置換された generative agent-based modeling (GABM)** として位置づけられる (Ghaffarzadegan et al. 2024)。ABM 伝統からの継承は重要だが、LLM の振る舞いはルールではなく訓練データに由来し、validation challenges が質的に異なる。

3.3 第 3 転回: Wizard-of-Oz と Personas（HCI, 1990s-2010s）

Dahlbäck et al. (1993) が制度化した Wizard-of-Oz 法は **人間が AI のフリ** をする UI 評価手法。Cooper (1999) Personas は設計対象としての仮想人物像を確立。

LLM 仮想実験との接続: 仮想実験法は逆方向（**AI が人間のフリ**）であり、Wizard-of-Oz の inversion とも見なせる。HCI が培った personas 設計の暗黙知は、LLM persona prompting の前史として参照可能。

3.4 第 4 転回: 計算社会科学の成熟（2009-2022）

Lazer et al. (2009) 以降、Big Data・SNS・行政記録の研究利用が拡大し、計算社会科学が確立した。**MTurk** (2005-) のマイクロタスク雇用が「擬似

母集団」として批判された議論は、後の silicon sample 議論の前史である。

3.5 第 5 転回: LLM 仮想実験の爆発（2022-）

Brown et al. (2020) GPT-3 paper を契機に、Argyle et al. (2023), Aher et al. (2023), Horton et al. (2023) が同時期に提唱論文を発表。Park et al. (2023) "Generative Agents: Interactive Simulacra"、Park et al. (2024) "Generative Agent Simulations of 1,000 People" が方法を成熟。批判系（Santurkar 2023, Bisbee 2024, Boelaert 2025）が並走、Hullman et al. (2026) が方法論的フレーミングを更新。

時系列パターン: - **2022-2023 上半期**: 提唱論文と楽観論 - **2023 下半期**: 慎重な肯定論と最初の反証 - **2024**: 体系的限界研究の集中 - **2025**: machine bias / WEIRD 偏向 / monoculture の概念化 - **2026 (Q1)**: Hullman フレームワークによる方法論的整理

---

4. 4 軸タクソノミーの提案

仮想実験法の研究空間を 4 つの直交軸で構造化する。各軸は独立で組み合わせ可能であり、論文を **(A_i, B_j, C_k, D_l)** という 4-tuple で位置づける。

4.1 軸 A: Simulation Layer（何をシミュレートするか）

- **A1: LLM-as-Respondent**: 個別被験者の回答をシミュレート。典型例は survey/vignette への persona prompting。
- **A2: LLM-as-Population**: 集団の応答分布をシミュレート。silicon sample が代表。
- **A3: LLM-as-Scenario**: 仮想場面・対話・社会的状況を生成。generative agents、silent counterfactual、未来人インタビューなどを含む。

層によって妥当性主張の射程が根本的に異なる。A1 は個別回答の interchangeability、A2 は集団分布の fidelity、A3 は scenario の plausibility と適切性を主張する。

4.2 軸 B: Validity Strategy（どう妥当性を主張するか）

ARC の stage 7 synthesis は本軸を 5 分類に拡張した。我々は、これを採用する:

- **B1: Face Validity**: 出力が "もっともらしく見える"。最も弱い保証。
- **B2: Benchmark Validity**: 既存の人間データと一致。Argyle 2023, Park 2024 が訴求。
- **B3: Mechanistic / Process Validity**: 内部の推論過程が target phenomenon に類似。Manning et al. 2024 が試行。
- **B4: Stress / Adversarial Validity**: プロンプト・モデル・seed・persona・framing の変動下で頑健。Sclar 2024, Liu 2024 で経験的検証。
- **B5: Participatory / Ecological Validity**: 当事者コミュニティが結果を自分のものとして認識し、利用文脈が現実的。Plurality・Citizens' Assemblies 系で要請されるが LLM 仮想実験では実装が薄い。

4.3 軸 C: Epistemic Goal（何のために使うか）

- **C1: Prediction / Substitution**: 人間データを置き換える・補強する。
- **C2: Theory Probing**: 既知の理論パターンが仮想集団で再現されるかを実験。
- **C3: Design Exploration**: ペルソナ・シナリオ・介入を生成して設計を探索。
- **C4: Normative Representation**: 不在の・周縁化された・未来の・サイレントなステークホルダーを代弁する。

C1-C4 は、適切とされる validity 基準が異なる。C1 は B2/B4 を要請、C3 は B1/B3 で許容され、C4 は B5 を unique に要請する（B2/B3 では原理的に保証できない）。

4.4 軸 D: Empirical Status（経験的にどこまで支持されているか）

- **D1: Established**: 複数研究で収束、再利用可能 benchmark あり。
- **D2: Contested**: 結果が混在、validity が setup に条件付け。
- **D3: Frontier**: 有望だが under-tested。
- **D4: Empirical-Void**: 概念的に論じられているが直接の経験的証拠が薄い。

軸 D は他の 3 軸と異なり、**研究空間の position ではなく現在の知見状態**を表す動的属性である。今後の研究で D 値は変化する。

4.5 タクソノミーの主張

主要な仮説は:

> **論文間の見かけの対立は、しばしば異なる (A, B, C) の組み合わせを評価していることに起因する。**

例えば、Argyle 2023（"silicon sample is fidelity-rich"）と Bisbee 2024（"silicon sample is unreliable"）の見かけの対立は: - Argyle: (A2, B2, C1, D1→2)（集団層、benchmark validity、置換目的、当時 established） - Bisbee: (A2, B4, C1, D2)（集団層、stress validity、置換目的、contested）

つまり同じ A2 × C1 を異なる B（B2 vs B4）で評価しており、validity strategy の選択が結論を分岐させている。タクソノミーはこの分岐を明示化する。

---

5. 主要文献の体系的サーベイ：18 文献への taxonomy 適用

aiXiv 査読の批判 #1（"具体的な実証検証を欠く"）への直接対応として、本節は 18 文献に 4 軸タクソノミーを系統的に適用する。

5.1 タクソノミー適用表

#	論文（著者・年・venue）	A	B	C	D	主要主張・所見
1	Argyle et al. (2023) Political Analysis	A1, A2	B2	C1	D2	"silicon sample" 提唱。GPT-3 が demographic conditioning で人口統計的分布を近似。
2	Aher, Arriaga, Kalai (2023) ICML	A1	B2	C2	D2	Turing Experiments で Ultimatum・Milgram・Wisdom of Crowds を再現。consistent distortion を観察。
3	Horton, Filippas,	A1, A2	B2, B3	C1, C2	D2	"Homo Silicus" 概念。Charness-Rabin, Kahneman

#	論文（著者・年・venue）	A	B	C	D	主要主張・所見
	Manning (2023) NBER					経済実験を qualitatively 再現。
4	Park, O'Brien et al. (2023, 2024) Stanford	A2, A3	B2	C1, C3	D3	1,000 人インタビュー → generative agent。test-retest accuracy 85%。
5	Manning, Zhu, Horton (2024) NBER	A2, A3	B3	C2, C3	D3	"Automated Social Science"。LLM が hypothesis 生成と subject の両方。circular validation の典型。
6	Santurkar et al. (2023) ICML	A2	B2	C1	D2	OpinionsQA。米国 60 デモグラフィック群との不一致を実証。65 歳以上・寡婦は particularly poor。
7	Dillion et al. (2023) Trends Cog Sci	A1	B2	C1, C2	D2	道徳判断 r=0.95。**3 つの応用**(仮説生成・項目パイロット・人間データの corroboration) のみを推奨。
8	Bisbee et al. (2024) Political Analysis	A2	B4	C1	D2→3	feeling thermometer 11 集団。**平均は近いが分散小・回帰係数異・プロンプト不安定・3 ヶ月で変動**。statistical inference に unreliable。
9	Sarstedt et al. (2024) Psych & Marketing	A1, A2	B2, B5	C1, C3	D2	construal formation で乖離。**WEIRD で有用、非 WEIRD で性能低下**。pilot 用途を推奨。
10	Boelaert et al. (2025) Sociol Methods & Res	A2	B2, B4	C1	D2	"machine bias"。**トピックごとに方向ランダム変動**の bias。replace は不適。
11	Hullman et al. (2026)	(全層)	B2-B5	(全 goal)	D3	**heuristic vs statistical calibration** フレーム。本サーベイ時点で最も洗練。
12	Bommasani et al. (2022) NeurIPS	(背景)	(背景)	(背景)	—	**Algorithmic Monoculture / Outcome Homogenization**。multi-

#	論文（著者・年・venue）	A	B	C	D	主要主張・所見
						vendor triangulation の前提を疑う基盤。
13	Messeri & Crockett (2024) Nature	(メタ)	—	—	—	"AI and illusions of understanding"。**scientific monoculture** への警鐘。仮想実験法を含む AI 媒介研究全般。
14	Sclar et al. (2024) ICLR	A1, A2	B4	(基盤)	D2	プロンプト書式変更で **最大 76 ポイントの精度差**。再現性の根本的脅威。
15	Liu et al. (2024) ACL	A1	B4	C1	D2	persona steerability。incongruous persona で 9.7% 低下、RLHF で semantic diversity が 58.2% 低下。
16	Wyllie et al. (2024) FAccT	A2	B4, B5	C1	D2→3	"Fairness Feedback Loops"。**合成データの再帰訓練で多数派へ収束、少数派消滅**。
17	Spiral of Silence in LLM Agents (arXiv 2510)	A2, A3	B3	C2, C4	D4	persona signal が opinion heterogeneity を促進、history signal が anchoring。LLM agent でも spiral of silence が emerge。
18	Park et al. + Stanford HAI deliberative 系 (2024-2025)	A3	B5	C3, C4	D4	participatory deliberation の AI 補強。AI penalty 現象（AI 関与が deliberation 評価を下げる）が同時に観察。

5.2 適用結果の所見

5.2.1 軸 A（層）の偏在

- A1（Respondent）= 11/18 件で最も研究が密
- A2（Population）= 10/18 件、批判文献の主戦場
- A3（Scenario）= 5/18 件、まだ薄い

→ 文献は **個別と集団に偏り、scenario-level の研究が遅れている**。

5.2.2 軸 B（妥当性戦略）の集中

- B2（Benchmark）= 主要 12 件。中心戦略。
- B4（Stress / Adversarial）= 6 件。批判文献の主戦場。
- **B5（Participatory / Ecological）= 3 件のみ**。経験的根拠が著しく薄い。

→ Validity strategy として **participatory / ecological validity がほぼ未開拓**。これは voice representation 研究の前提を揺るがす。

5.2.3 軸 C（目的）の不一致

- C1（Substitution）= 主流の主張、しかし D2（contested）に押し戻されつつある
- C3（Design Exploration）= D3（frontier）として最も建設的
- **C4（Normative Representation）= 主張は多いが D4（empirical-void）**

→ 「声を持たない者の代弁」は **概念的に論じられているが、経験的に validate されていない**。これは倫理的に最も慎重を要する用途であるにもかかわらず。

5.2.4 軸 D（経験的状態）の集計

Status	件数	該当例
D1 Established	0	（あくまで 4 軸の組み合わせで見ると established と呼べるセルは現存しない）
D2 Contested	12	主要批判文献の主戦場
D3 Frontier	4	Park 2024, Manning 2024, Hullman 2026, Wyllie 2024
D4 Empirical-Void	4	Bommasani 背景, Spiral of Silence, Park deliberative, Normative Representation 全般

→ **D1 Established が 0 件**。これは仮想実験法の妥当性主張が、4 軸の任意の組み合わせのもとで完全に established になっている領域がまだ無いことを示す。

---

6. 主要発見の Meta-Synthesis

ここでは、5.1 の matrix と各文献の所見から、仮想実験法の **algorithmic fidelity に対する経験的支持の状態** を 9 つの subcategory で集計する。

6.1 Empirical Status Matrix

主張	支持	反証	現状評価
平均値・aggregate trend の再現	Argyle 2023, Horton 2023, Park 2024	Bisbee 2024 (部分)	○ 比較的 robust
分散・variation の再現	(なし)	Bisbee 2024, Boelaert 2025	× 構造的に低分散
回帰係数・統計的推論	(限定)	Bisbee 2024	× 不適
サブグループ・少数派	(限定)	Santurkar 2023, Sarstedt 2024 (非 WEIRD)	× 偏向強
道徳判断 (mean)	Dillion 2023 (r=0.95)	(定量反証薄)	△ 高相関だが confirmatory 不可
経済古典実験 (qualitative)	Horton 2023, Aher 2023	Manning 2024 (価格予測不正確)	△ 質的再現可、量的不可
プロンプト不変性	(なし)	Sclar 2024 (76pt 差), Bisbee 2024	× 重大な脅威
時間的安定性	(なし)	Bisbee 2024 (3 ヶ月で変動)	× 不安定
個人レベル再現	Park 2024 (85%)	(再検証薄)	△ 新発見、要再検証

6.2 機能と限界の対称性

注目すべきパターン: **平均は再現できるが分散が消える** という非対称が、複数文献で独立に観察されている (Bisbee 2024, Boelaert 2025, Wyllie 2024)。これは技術的偶然ではなく、**LLM の訓練・推論プロセスが集団中心へ収束する構造的傾向** を示唆する。

ARC stage 8 が独立に立てた Hypothesis 2:

> "Apparent benchmark success in LLM-as-respondent tasks hides structural compression: subgroup disagreement, tails, and semantic diversity collapse even when means look accurate."

これは本サーベイの empirical pattern と完全に整合する独立予測である。

6.3 学際的非対称性

- **政治学** (Argyle, Bisbee, Boelaert) = 早期に体系的批判を生産
- **経済学** (Horton, Manning) = シミュレーション肯定的（実験経済学伝統と整合）
- **心理学** (Dillion, Park) = 慎重に応用拡大、補助用途強調
- **マーケティング** (Sarstedt) = WEIRD 偏向に焦点、業界応用直結
- **HCI / NLP** (Liu, Sclar, Hullman) = モデル内部メカニズムの検証
- **社会学** (Boelaert) = machine bias の概念化

→ 領域横断レビューが不在であり、本論文の貢献の一つはこの **横断的整理そのもの** である。

---

7. 三領域分類: Saturated / Frontier / Empirical-Void

5.1 の matrix と 6.1 の empirical status の集計から、仮想実験法の研究空間を 3 つの領域に分類する。

7.1 Saturated 領域

- **位置**: A1/A2 × B2 × C1
- **内容**: LLM-as-Respondent / Population × Benchmark Validity × Prediction-Substitution
- **状態**: 多数の研究が存在し、限界も体系的に同定済み
- **代表**: Argyle 2023, Bisbee 2024, Boelaert 2025
- **判断**: 新規論文は **incremental** になりやすい。既存批判への回答や、mitigation strategy の精緻化が中心になるべき。

7.2 Frontier 領域

- **位置**: A3 × B3 × C3、および A2 × B4 × C2
- **内容**: LLM-as-Scenario × Mechanistic Validity × Design Exploration / Theory Probing

- **状態**: 有望だが under-tested
- **代表**: Park 2024 (1000 agents), Manning 2024 (automated social science), Hullman 2026 (calibration framework)
- **判断**: 新規論文は **novelty 高く、影響可能性大**。検証手法の確立が論文化の鍵。

7.3 Empirical-Void 領域

- **位置**: A3 × B5 × C4
- **内容**: LLM-as-Scenario × Participatory / Ecological Validity × Normative Representation
- **状態**: 概念的に論じられているが直接の経験的証拠が薄い
- **代表**: Park-stanford deliberative 系の試行、Tang & Weyl (Plurality) の理論的フレーム、Wales Future Generations Act の制度的先例
- **未確立**: 当事者団体協働の design protocol、訓練データ provenance との関係、可訂正性のフレーム
- **判断**: **倫理的に最も慎重を要するにもかかわらず、経験的検証が遅れている**。本論文の最大の発見の一つ。

7.4 領域分類の意義

3 領域の同定は、研究コミュニティに以下のロードマップを提供する:

関心領域	推奨領域	理由
大学院生・新規参入者	Frontier	学習しがいがあり、貢献可能性も高い
実装者・実装研究者	Saturated（の限界回避）	既知の限界を回避する design literacy が必要
政策研究者	Empirical-Void	制度設計と経験研究が並走できる稀有な領域
倫理学者・哲学研究者	Empirical-Void	規範的議論が経験研究に先行しており、概念整理の余地大

---

8. 4 つの Falsifiable Hypotheses

aiXiv 査読批判 #1 への対応として、本節は 4 つの testable hypothesis を提示する。各仮説に **Measurable Prediction**（数値閾値付き）と **Failure Condition**（pre-registered な棄却基準）を付随する。これらは ARC stage

8 が独立に生成した hypotheses を本論文の文脈で再形式化したものである。

8.1 H1: LLMs are most valid as instruments for omission detection, not stakeholder representation

主張: 「誰が／何が抜けているか」を聞く omission detection は、persona simulation よりも有用性・倫理性の双方で優れる。

Measurable prediction: 限定された設計タスクにおいて、ハイブリッド（人間 + LLM omission detection）ワークフローは: - human-only より overlooked concerns カバレッジが **≥20% 増加** - persona simulation mode より人間参照リストへの Recall@k が **≥25% 高い** - persona mode より stereotype rate が **≥30% 低い** - legitimacy ratings が human-only から **≤10% 低下** に留まる

Failure condition (任意一つで棄却): - omission mode が human-only を coverage で上回らない - omission mode の stereotype/legitimacy が persona mode と差なし - additional scenario diversity が下流ワークショップ品質を改善しない - LLM-only/persona が legitimacy で omission mode と同等

8.2 H2: Apparent benchmark success hides structural compression

主張: LLM-as-Respondent タスクでの平均値の見かけの一致は、subgroup variance、tails、semantic diversity の構造的圧縮を隠している。

Measurable prediction: 公開 survey 項目（subgroup labels 付）で: - 項目平均で人間と **r > 0.6** の一致 - ただし **subgroup variance / entropy が 15-30% 低い** - 極端応答が人間より少なく、tails が薄い - disagreement structure が人間より弱い

合成サンプルサイズが拡大するにつれ: - top semantic clusters の集中度が上昇 - 一部タスクでは top-3 cluster が **≥70%** を占める

Failure condition: - LLM が subgroup variance、tails、disagreement structure すべてで人間と一致 - diversity がサンプルサイズと比例し collapse / plateau しない - prompt 変動が支配的すぎて compression pattern を安定推定できない

8.3 H3: Abstention asymmetries reveal representation failure

主張: prompted uncertainty / abstention は、特に周縁化・stigmatized・原理的に validate 不能な対象において、表現の失敗を answer accuracy より敏感に診断する。

Measurable prediction: 3 種の prompt category で: 1. mainstream survey items 2. stigmatized / risky self-disclosure 3. marginalized / future stakeholder personas

abstention 許可下で: - カテゴリ 2-3 で 1 より abstention/hedging 率が **≥20pp 高い** - 倫理的に fraught な対象で entropy shift が大きい - "未来世代" や "サイレントマジョリティ" prompt で uncertainty / pluralized framing 出力が direct voice simulation より **calibration 良好で false authority が低い** と評価される

Failure condition: - abstention/hedging パターンが target type 間で差なし - abstention 効果が prompt-paraphrase noise より大きくない - direct voice simulation が contested target で uncertainty framing と同等以上に適切と評価される

未解決争点: abstention は (a) model failure、(b) appropriate humility、(c) alignment policy artifact のどれを測定しているか。

8.4 H4: LLM outputs preserve intervention geometry despite mechanism-level failure

主張: LLM は人間 realism や mechanism-level validity を欠いても、介入の方向・順位・相互作用構造は preserve する可能性がある。

Measurable prediction: メッセージ・framing・vignette の小規模 manipulation 群で: - LLM は人間の reasoning trace や subgroup realism を再現できない - にもかかわらず: - 治療効果の **direction** を correctly 予測 - メッセージ variant の **rank ordering** を correctly 予測 - 一部の **interaction structure** を保持 - 既知の人間効果と方向一致 **>70%** - script-lesion perturbation が control lesion より効果サイズの retention を大きく低下（社会的 priors への dependence の証拠）

Failure condition: - effect direction が prompt / model version 間で不安定 - LLM が naive expert guessing を治療順位で上回らない - script lesion と control lesion が同程度に性能を低下

8.5 優先順位の推奨

ARC の独立評価（stage 8）と本サーベイの両方が、**H1 → H2** の優先順位を支持する: - H1: 最も実装容易、ethics と utility の両方に直接的含意 - H2: 公開 survey データで実行可能、theoretical anchor が強固 - H3: 概念的に最も新しいが alignment artifact との切り分けが難しい - H4: contrarian 仮説として保持。代弁可能性の上限を測る

---

9. Honest Design Principles（Operational Checklists 付）

aiXiv 査読批判 #2（"原則が高水準すぎる"）への対応として、各原則を **3-5 点の operational checklist + thresholds** に展開する。

Principle (a): 4 軸タクソノミーでの位置づけ明示

研究は (A_i, B_j, C_k) のどの位置に主張を置くかを **abstract と方法節で明示する**。

Checklist: 1. ☐ Simulation Layer (A1/A2/A3) を方法節冒頭で明示 2. ☐ 採用する Validity Strategy (B1-B5) を一つ以上明示し、なぜその選択かを正当化 3. ☐ Epistemic Goal (C1-C4) を主張の射程として明確化 4. ☐ 主張の Empirical Status の自己評価 (D1-D4) を考察節で記述 5. ☐ 同じ層・目的で異なる validity strategy を採用した先行研究との比較を 3 文以上

Principle (b): Multi-vendor / Multi-version Triangulation

LLM 実験は最低でも **2 つの異なる vendor** で再現性を確認する。同一系列の異 version は triangulation と認めない。

Checklist: 1. ☐ 主要結果が **2 つ以上の異なる vendor**（Anthropic / OpenAI / Google / Meta etc.）で再現 2. ☐ vendor 間の divergence を appendix に **データとして開示**（収斂を主張するだけでなく divergence も） 3. ☐ vendor 間で訓練データ・RLHF の重複可能性を考察 4. ☐ "independent" を主張する場合、cosine similarity of embeddings、RLHF dataset overlap statement のいずれかを report 5. ☐ Bommasani et al. (2022) algorithmic monoculture を参照

Principle (c): Pre-registration の徹底（プロンプト・seed・モデルを含む）

仮想実験は通常の pre-registration に加えて、**プロンプト全文・モデル ID・seed・温度** を pre-register する。

Checklist: 1. ☐ プロンプト全文を OSF / pre-registration platform に登録 2. ☐ モデル ID（バージョン文字列まで）を登録 3. ☐ seed と温度を固定（または探索範囲を pre-register） 4. ☐ kill criteria（仮説棄却閾値）を pre-register し、結果が境界に近い場合の事後再解釈を禁止する protocol を明示 5. ☐ pre-registered 計画からの deviation を separate appendix で報告

Principle (d): Statistical Calibration over Heuristic Interchangeability

Hullman et al. (2026) のフレームに沿い、confirmatory 主張をする場合は **必ず statistical calibration** を行う。heuristic interchangeability は exploratory 段階のみで許容される。

Checklist: 1. ☐ 主張が exploratory か confirmatory かを明示 2. ☐ confirmatory なら、人間データを用いた calibration step を含む 3. ☐ calibration の statistical adjustment（imputation, MRP, raking, etc.）を report 4. ☐ calibration の uncertainty を主結果に伝播 5. ☐ heuristic-only ならその scope を結論で明示し、generalizable claim を慎まない

Principle (e): 当事者団体協働ガバナンス（C4 用途で必須）

Normative Representation (C4) を主張する研究は、**対象集団の当事者団体・コミュニティ代表との設計段階からの協働** を必須とする。

Checklist: 1. ☐ 対象集団を特定し、その集団の当事者団体・代表の連絡先を identify 2. ☐ プロンプト設計、シナリオ生成、結果解釈の各段階で当事者の review を経る 3. ☐ 当事者から拒否された prompt / scenario / 解釈を report に含める（unless explicitly waived by community） 4. ☐ 結果出版前に当事者団体に最終 review の機会を提供 5. ☐ 訓練データに当事者の SNS / 発言が含まれる可能性とその倫理的含意を考察

Principle (f): WEIRD / 言語 / 文化文脈の明示

Sarstedt et al. (2024) が示した WEIRD 偏向を踏まえ、**研究対象集団の WEIRD-ness** と LLM の文化バイアスを明示する。

Checklist: 1. ☐ 対象集団の地理・言語・所得階層を明示 2. ☐ 主張する集団が WEIRD か非 WEIRD かを explicit に position 3. ☐ 非 WEIRD 集団を扱う場合、本対象に関する LLM の性能低下可能性を方法節で警告 4. ☐ 多言語実験では各言語での結果差を report 5. ☐ 結論で「本研究の妥当性は WEIRD / 特定文化文脈に限定される」と明示（該当時）

Principle (g): Failure の生産的開示

成功事例だけでなく、仮想実験と現実が乖離した事例、棄却された hypothesis、misalignment 事例を **意図的に開示** する。

Checklist: 1. ☐ 別 appendix で「うまくいかなかった prompt / model / target」を報告 2. ☐ failure mode を分類し、メカニズムの仮説を提示 3. ☐ 棄却された hypothesis に対する論文間 cross-referencing を提供 4. ☐ negative results を共有する公開リポジトリ（OSF, GitHub）を指定 5. ☐ "ablation" や "limitation" を perfunctory に流さず、本論的考察として扱う

Principle (h): Recursive Limitation の認識（LLM 媒介研究を LLM が分析する場合）

LLM 媒介研究を LLM 自身が分析・査読・統合する場合、**recursive な認識論的依存関係を明示する**。

Checklist: 1. ☐ 著者・査読者・分析者が LLM か人間かを明示 2. ☐ LLM 著の場合、prompt engineering と validation strategy を appendix で開示 3. ☐ 同じ vendor の LLM を著者と査読の両方に使うことを禁止（独立 triangulation 不可となる） 4. ☐ recursive limitation が論文の evidential basis にどう影響するかを考察節で論じる 5. ☐ Messeri & Crockett (2024) "illusions of understanding" を参照

---

10. 参加型デザイン・倫理との接続

10.1 Empirical-Void 領域の倫理的優先性

§7.3 で同定した Empirical-Void 領域 = A3 × B5 × C4（LLM-as-Scenario × Participatory Validity × Normative Representation）は、倫理的に最も慎重を要するにもかかわらず、経験的検証が最も遅れている。これは以下の構造から生じる:

1. 当事者が「不在」だから経験データが取りにくい
2. LLM が代弁可能性を過大主張しやすい（出力が plausible に見える）
3. 訓練データに当事者発言が含まれる場合の二重搾取問題が複雑
4. 検証可能性が原理的に困難な対象（未来世代、亡くなった人々）を含む

この領域の研究を倫理的に成立させるために、以下の **代替性・顕示・代弁ガバナンスの三原則** を提案する:

- **代替性原則 (Substitutability Principle)**: 人間で実施可能なら仮想は選ばない (avoid by simulation の濫用禁止)
- **顕示原則 (Disclosure Principle)**: 仮想であることを論文・成果物で明示し、現実主張に偽装しない
- **代弁ガバナンス原則 (Voice Governance Principle)**: 当事者団体・関連コミュニティとの協働なしに代弁を主張しない

10.2 Plurality との接続

Tang & Weyl (2024) "Plurality: The Future of Collaborative Technology and Democracy" は、多様性を embrace する技術設計を提唱する。Plurality の **「Internet of beings」** 構想は仮想実験法の倫理的位置づけに以下を示唆する:

- 仮想実験は **集合知識の生成的拡張** として位置づけ可能
- ただし「集合」が当事者を含まないなら拡張ではなく **代替** に堕する
- 拡張として成立させるには Citizens' Assembly 系の deliberative 装置との接続が必要

10.3 Future Generations Voice

Wales Well-being of Future Generations Act (2015) は世界初の Future Generations Commissioner 制度を確立し、現在 Scotland、Ireland、

Japan、NZ、Canada、France、Germany、Australia (2025) が参照している。仮想実験法は:

- 未来世代の意思決定参加を支援する **stress test 装置** として価値あり
- ただし「未来人インタビューを LLM が代行できる」という主張は H3（abstention asymmetry）の検証を経るべき
- 制度的代弁者（Commissioner）と LLM 仮想実験のハイブリッド設計が研究空白

10.4 障害者運動 "Nothing About Us Without Us"

仮想実験法を当事者抜きで運用することは、障害者権利運動の核心的批判に直接抵触する。本論文は、この原則を **C4 主張をする全ての研究** に拡張することを提案する。

10.5 IRB / 研究倫理審査の動向

Frontiers (2026) "Streamlining IRB review of AI human subjects research" の三段階フレームワーク、HHS OHRP / SACHRP (2024) の AI 勧告、NIEHS (2025) の synthetic data ethics 議論が並走している。仮想実験法に特化した倫理審査基準はまだ確立していないが、Principle (e), (f), (g), (h) はその基礎要素を提供する。

---

11. JTBD 分析: 誰の・何の・なぜ

仮想実験法を「誰が・何のために・なぜ」採用するかを 9 つの Jobs-to-be-Done で整理する。

Job	想定者	解決したい問題	仮想実験法への期待
J-1 倫理的に難しい実験を実施したい	心理学・医療研究者	リアル被験者に倫理的負担をかけられない	A3 シナリオで代替探索
J-2 大規模事前パイロットを安価に行いたい	尺度開発者	リクルート費用と時間	A1 で項目特性事前推定
J-3 政策反応を試算したい	公共政策研究者	実施後では取り返しがつかない	A2 でシナリオ毎反応推定
			A3 で仮構的可視化

Job	想定者	解決したい問題	仮想実験法への期待
J-4 声なきステークホルダーを組み込みたい	参加型デザイナー	未来人・将来世代・極端少数派	
J-5 サイレントマジョリティを推定したい	政治学者・社会学者	SNS で発信しない大多数	A2 で訓練データ偏向と triangulate
J-6 自然実験対照群を仮想的に作りたい	因果推論研究者	対照不在の場合の counterfactual	A2 だが妥当性は最も慎重
J-7 評価尺度の構成概念妥当性を検証したい	心理測定研究者	リアルだけでは多角検証不可	A1 でシナリオ条件変動応答分析
J-8 訓練・教育用のリアル対話相手が欲しい	対人支援職教育、医師問診訓練	実患者で練習させられない	A3 ロールプレイ
J-9 査読・査定基準を作りたい	編集者・レビュアー	LLM 仮想実験論文の評価基準不在	メタ研究としての design principles

J-1, J-4, J-5, J-8 は **C4 (Normative Representation)** に該当し、Empirical-Void 領域に位置する。これらの Job を倫理的に満たすには本論文の Principle (e) 当事者ガバナンスが必須となる。

---

12. Open Problems

文献横断的に未解決の問題群:

1. **Persona conditioning の理論的限界**: 属性プロンプトでどこまで集団分布を再現できるかの formal な上限
2. **RLHF 偏向の文化横断的測定**: Santurkar 2023 の対象集団を多文化に拡張
3. **訓練データ汚染の影響推定**: 既存尺度が訓練データに含まれることの体系的測定
4. **Multi-vendor 独立性の検証**: 訓練データ共有可能性の外部検証手法
5. **代弁正当性の判定基準**: 「声なき者の声」を可視化することの規範的判定基準
6. **未来世代代弁の方法論**: 過去データから訓練された LLM が未来人を代弁できるという主張の根拠

7. **仮想 pre-registration の運用**: プロンプト・seed・モデル ID を含む pre-reg の標準化
8. **再現性インフラ**: プロンプト・モデル・条件の完全保存と再実行を保証するツールチェイン
9. **倫理レビュー基準**: IRB が仮想実験法を扱うための基準書
10. **abstention の意味論**: Hypothesis 3 の「沈黙」が model failure / appropriate humility / alignment artifact のどれを反映するかの切り分け
11. **WEIRD 拡張の方法論**: 非 WEIRD 集団に対する LLM 性能を低下させずに拡張する手法
12. **代弁ガバナンスの institutional design**: 当事者団体協働を学会・出版基準に組み込むメカニズム

---

13. Limitations

aiXiv 査読批判 #3, #4 への対応として、本レビュー自体の limitations を honest に列挙する。

13.1 言語と地理の偏り

本レビューは英語と日本語の文献を対象とし、その他言語は非対象である。特に: - 中国語の計算社会科学文献（中国国内の合成被験者研究） - スペイン語・ポルトガル語のラテンアメリカ研究 - アラビア語・アフリカ諸言語の研究文献 は系統的に欠落している。

WebSearch クエリ "仮想実験 LLM 2024 2025" は **0 件返答** した。これは日本語の学術コミュニティで仮想実験法の体系的議論が本論文時点で極めて薄いことを示している。

13.2 検索の filter-bubble バイアス

WebSearch（Google ベース）に依存しており、Scholar・SSRN・arXiv を直接 API で叩いた網羅検索ではない。同じクエリを Bing / DuckDuckGo / Baidu 等で実行すれば結果が異なる可能性。

13.3 18 文献採択の selection bias

18 文献の選定は、(a) 高被引用 (b) 主要英語誌 (c) 著者・批判の対立構造の可視化に役立つもの、を優先した結果である。次のような perspective が relatively 過少代表: - **decolonial AI** 系の批判（Mhlambi 2020, Birhane

2021 等） - 非西洋文脈からの AI ethics（特にアフリカ・南米・南アジア系） - **障害学（disability studies）** 視点 - 言語的少数者の代弁問題

13.4 4 軸タクソノミーの構成 bias

本論文の 4 軸（Simulation Layer, Validity Strategy, Epistemic Goal, Empirical Status）は、(a) 著者の coding bias、(b) ARC 自律走行が独立に到達した同じ枠組み、の双方の影響を受けている。タクソノミーが文献を構造化する側面と、文献から構造化される側面の双方向性は、本質的に解釈学的循環を含む。

代替的に提案され得る軸: - **Time Horizon**: 即時応答 vs 長期インタラクション - **Modality**: テキストのみ vs マルチモーダル - **Intervention vs Observation**: 操作実験 vs 観察的シミュレーション - **Single Agent vs Multi-Agent**: 個別 vs 群動態

これらを組み合わせた 6 軸・8 軸構造も論理的に可能だが、本論文では parsimony と既存文献の構造的特徴から 4 軸を採用した。

13.5 Recursive Limitation: 本論文が LLM 著であることの含意

本論文は LLM (Claude Opus 4.7) によって執筆された合成として書かれている。これは仮想実験法の方法論を主題とする本論文の性格と recursive な関係を持ち、以下の認識論的問題を生む:

1. **算法的閉鎖性の自己参照**: 本論文が批判する "LLM が LLM を評価する閉ループ" は、本論文の執筆過程そのものに当てはまる
2. **selection bias の自己強化**: 著者 LLM の訓練データに含まれる文献が優先的に整理されている可能性
3. **machine bias のトピック横断的伝播**: Boelaert (2025) が指摘する topic-specific bias が、本論文の主張の方向に体系的に作用している可能性
4. **prompt 敏感性**: 異なるプロンプトで本論文を執筆していれば異なる taxonomy / hypothesis に到達した可能性

13.5.1 緩和のために実施した validation strategy

本論文は以下の validation step を経た:

- **WebSearch + ARC の独立合成**: 18 文献の WebSearch サーベイと、ARC v0.3.1 の自律 synthesis (rc-20260428-213351-42de41) が独立に同じ 4 軸タクソノミーと 3 領域分類に到達したことを cross-check

- **対立する立場の意図的な対比**: ARC stage 8 が生成した contrarian / innovator / pragmatist の三 perspective を保持し、各 hypothesis に "Unresolved Disagreement" 節を付与
- **査読者批判の pre-emptive 反映**: 既知の aiXiv 査読 (rc-20260427-llm-mediated-research-review) の 4 つの critique を本論文の構成に取り込み
- **数値閾値・棄却基準の formal 化**: 抽象的主張を可能な限り falsifiable な形に
- **search query 全開示**: §A1 Appendix で 18 クエリ全文を開示

13.5.2 まだ達成できていない validation

- **人間研究者による独立査読を経ていない**: 本 v1.0 は LLM 著 + LLM 査読 (aiXiv) のみで構成され、人間研究者による独立な fact-checking が未実施
- **citations の bibliographic verification が部分的**: 主要文献は author/year/venue を確認したが、引用箇所の semantic accuracy は完全には verify されていない
- **多 vendor triangulation の不在**: 本論文は単一 LLM (Claude Opus 4.7) で書かれており、別 vendor の LLM で同じ素材を統合したらどう異なるかは未検証

13.6 業界文献の partial coverage

市場調査・コンサルティング系の synthetic respondents 文献（NielsenIQ, MRS Delphi Report, Conjointly, Quirks 等）は本論文の Industry セクションで触れたが、その内部資料・proprietary なベンチマークは取得困難。業界の実装は学術より進んでいる側面があり、これが完全には反映されていない。

13.7 経済学経験実験 (lab experiment) との関係の浅さ

Horton, Manning らの経済学的実験再現研究は本論文に含めたが、Charness, Fehr, Camerer 等の従来 lab experiment 文献との接続は十分に深掘りされていない。

---

14. 結論

仮想実験法は心理学・社会科学・政策研究の方法論的景観を変えつつある。Argyle 2023 の "silicon sample" 提唱から 3 年で、文献は急速に拡大し、批判的検証も成熟した。しかし、領域横断的に **「いかなる妥当性主張がいかなる戦略の下で支持されるか」** を整理する taxonomy は不在であり、結果として研究者は領域固有の事例で迷い、査読者は評価基準を持たず、実装者は限界を可視化できない状況にあった。

本レビューは:

1. **4 軸タクソノミー** (Simulation Layer × Validity Strategy × Epistemic Goal × Empirical Status) を提案
2. **18 文献への系統的適用** で empirical status matrix を生成
3. **3 領域分類** (Saturated / Frontier / Empirical-Void) で研究空間のロードマップを提供
4. **4 つの falsifiable hypotheses** を pre-registered prediction で形式化
5. **8 つの operational design principles** を 3-5 点 checklist + thresholds で展開
6. **参加型デザイン・倫理との接続** を Empirical-Void 領域として明示

最も重要な発見は、**Empirical-Void 領域 = LLM-as-Scenario × Participatory Validity × Normative Representation** が、倫理的に最も慎重を要する用途（silent majority、future generations、marginalized groups の代弁）であるにもかかわらず、経験的検証が最も遅れている、という非対称である。本領域は今後の研究の最重要 frontier であり、当事者団体協働ガバナンスを必須とする design principle (e) を伴わなければ倫理的には成立しない。

ARC v0.3.1 の自律 synthesis と本論文の手動 synthesis が独立に同じ taxonomy・3 領域分類に到達したことは、本論文の概念構造の robustness を示唆する。ただし、両者ともに LLM ベースであることの recursive limitation を §13.5 で honest に論じた。本論文の真の強度は、人間研究者による独立な事実検証と、複数 vendor LLM での triangulation を経て初めて確定する。それは本 v1.0 が読者・査読者に対して提起する次の作業である。

---

15. References

15.1 Foundational（仮想実験法を肯定的に提唱）

- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337-351. https://doi.org/10.1017/pan.2023.2
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. ICML 2023, PMLR 202:337-371.
- Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper 31122.
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597-600.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. UIST 2023.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., et al. (2024). Generative agent simulations of 1,000 people. Stanford HAI / Stanford Digital Repository.
- Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. NBER Working Paper 32381.
- Sarstedt, M., et al. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. Psychology & Marketing, 41(4), 1-21.

15.2 Critical（方法論的反証）

- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? ICML 2023.
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. Political Analysis, 32(4), 401-416.

- Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine bias: How do generative language models answer opinion polls? Sociological Methods & Research, 54(3), 1156-1196.
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? NeurIPS 2022.
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627, 49-58.
- Hullman, J., et al. (2026). This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. arXiv: 2602.15785.

15.3 メカニズム（プロンプト・ペルソナ・バイアス）

- Sclar, M., et al. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. ICLR 2024.
- Liu, A., et al. (2024). Evaluating large language model biases in persona-steered generation. Findings of ACL 2024.
- Gupta, S., Shrivastava, V., et al. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. ICLR 2024.
- Wyllie, S., et al. (2024). Fairness feedback loops: Training on synthetic data amplifies bias. FAccT 2024.

15.4 Counterfactual / Causal

- arXiv:2410.06392 (NeurIPS 2024). Counterfactual causal inference in natural language with large language models.
- Validation challenge review (2025). Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. Artificial Intelligence Review.

15.5 Vignette / Factorial Survey 前史

- Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In Rossi & Nock (Eds.), Measuring social judgments.
- Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. Social Science Research, 38(3), 505-520.

- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. PNAS, 112(8), 2395-2400.

15.6 Generative Social Science / ABM

- Schelling, T. C. (1971). Dynamic models of segregation. Journal of Mathematical Sociology, 1(2), 143-186.
- Axelrod, R. (1984). The evolution of cooperation. Basic Books.
- Epstein, J. M., & Axtell, R. (1996). Growing artificial societies: Social science from the bottom up. Brookings/MIT Press.
- Epstein, J. M. (2006). Generative social science: Studies in agent-based computational modeling. Princeton.
- Lazer, D., Pentland, A., et al. (2009). Computational social science. Science, 323(5915), 721-723.
- Ghaffarzadegan, N., Majumdar, A., Williams, R., & Hosseinichimeh, N. (2024). Generative agent-based modeling: An introduction and tutorial. System Dynamics Review.

15.7 Plurality / Future Generations / Deliberation

- Tang, A., Weyl, E. G., & Plurality Community. (2024). Plurality: The future of collaborative technology and democracy. plurality.net.
- Welsh Government. (2015). Well-being of Future Generations (Wales) Act. futurepolicy.org.
- Australian Parliament. (2025). Wellbeing of Future Generations Bill 2025.
- Knight First Amendment Institute. (2025). Can AI mediation improve democratic deliberation?
- Revel & Penigaud (2025). AI-Enhanced Deliberative Democracy. arXiv:2503.05830.
- Procaccia, A. D. (2025). Auditing representation in online deliberative processes.
- Fishkin, J. S. Deliberative polling. PhilPapers.

15.8 Silent Majority / Opinion Simulation

- Gomez, et al. (2024). Unveiling the silent majority: Stance detection and characterization of passive users on social media

using collaborative filtering and graph convolutional networks. EPJ Data Science.
- Nature Scientific Reports (2025). Social opinions prediction utilizes fusing dynamics equation with LLM-based agents.
- arXiv:2510.02360. Spiral of silence in large language model agents.
- arXiv:2603.16142. Parametric social identity injection and diversification in public opinion simulation.

15.9 Ethics / IRB

- Frontiers (2026). Streamlining IRB review of AI human subjects research (AIHSR): The three-stage framework.
- HHS OHRP / SACHRP (2024). IRB considerations on the use of artificial intelligence in human subjects research.
- NIEHS Environmental Factor (2025 April). Synthetic data created by generative AI poses ethical challenges.
- arXiv:2412.16022. The only way is ethics: A guide to ethical research with large language models.
- Kapania, S., Wang, R., & Huang, S. (2024). 'I'm categorizing LLM as a productivity tool': Examining ethics of LLM use in HCI research practices. Proceedings of the ACM on Human-Computer Interaction.

15.10 Industry

- NielsenIQ (2024). The rise of synthetic respondents in market research.
- Market Research Society (UK). MRS Delphi report: Using synthetic respondents for market research.
- Quirks (2024). Synthetic respondents and the future of survey research.
- Conjointly (2024). Synthetic respondents are the homoeopathy of market research.
- International Journal of Research in Marketing (2025). Special Issue: Generative AI, synthetic data, and synthetic respondents in marketing research.

15.11 Adjacent: 関連批判

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT 2021.
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2).

Appendix A1: 検索クエリ全開示

aiXiv 査読批判 #3 への対応として、本レビューで実行した 18 検索クエリを完全に開示する。

#	系統	クエリ	採択論文
1	Foundational	`Argyle Busby "Out of one, many" language models simulate human samples Political Analysis 2023`	Argyle 2023
2	Foundational	`Park generative agent simulations 1000 people 2024 Stanford`	Park 2024
3	Foundational	`Aher Arriaga Kalai "using large language models to simulate multiple humans" ICML 2023`	Aher 2023
4	Foundational	`Horton "large language models as simulated economic agents" 2023`	Horton 2023
5	Critical	`Bisbee "synthetic replacements" human survey data LLM 2024`	Bisbee 2024
6	Critical	`Santurkar "whose opinions do language models reflect" 2023`	Santurkar 2023
7	Critical	`Dillion "can AI language models replace human participants" 2023`	Dillion 2023
8	Critical		Bommasani 2022

#	系統	クエリ	採択論文
		`algorithmic monoculture LLM Bommasani picking same person`	
9	Critical	`Boelaert "machine bias" LLM survey 2025`	Boelaert 2025
10	Mechanism	`LLM persona prompting demographic bias steerability limits`	Liu 2024, Gupta 2024
11	Mechanism	`"prompt sensitivity" LLM survey reproducibility wording effect 2024`	Sclar 2024
12	Methodological	`Hullman "validating LLM simulations as behavioral evidence" Northwestern`	Hullman 2026
13	Causal	`LLM synthetic counterfactual causal inference natural experiment 2024 2025`	NeurIPS 2024, CounterBench 2025
14	Adjacent (ABM/GSS)	`"generative social science" LLM agent-based modeling Epstein 2024`	Epstein 2006, Ghaffarzadegan 2024
15	Adjacent (Vignette)	`vignette experiment factorial survey design history social science Rossi`	Rossi 1982, Wallander 2009, Hainmueller 2015
16	Participatory	`Audrey Tang Plurality book Glen Weyl AI participatory democracy`	Tang & Weyl 2024
17	Participatory	`future generations voice deliberation LLM AI participatory democracy`	Future Generations Wales Act, Knight Institute, Revel
18	Ethics	`Generative AI research participants IRB synthetic data ethics review 2025`	Frontiers 2026, HHS OHRP, NIEHS

補助系: 検索 "virtual experiment Japan psychology 仮想実験 LLM 2024 2025" は 0 件。日本語コンテキストの遅れを示す。

Appendix A2: ARC 自律走行との合意点・対立点

A2.1 合意点

主要構成	本レビュー	ARC stage 7	一致度
4 軸タクソノミーの提案	○	○	✓
Simulation Layer 3 分類	A1/A2/A3	Respondent/Population/Scenario	✓
3 領域分類 (Saturated/Frontier/Void)	○	○	✓
Empirical-Void = 声の代弁	○	"normative representation of silent majorities, future generations, marginalized groups"	✓
Validity が non-uniform	○	"validity is highly non-uniform across the four axes"	✓

A2.2 対立点・補完点

- **Validity Strategy 軸の分類**: 本レビュー初版では 4 分類（Face/Benchmark/Mechanistic/Stress）だったが、ARC が **Participatory/Ecological** を 5 分類目として加えるよう提案。本論文は ARC の提案を採用した。
- **文献 retrieval**: ARC の lit-collection（OpenAlex 30 件）は本論文の中核論文を取得できなかった。本論文の WebSearch 結果が文献選定では優位。
- **Hypothesis 形式**: ARC が独立に立てた 4 hypothesis は、本論文の §8 の元データとして重用された。

A2.3 ARC stage 10 の構造的失敗からの所見

ARC は stage 10 (CODE_GENERATION) で topic-experiment misalignment を自己診断し、2 回 regen 試行後に fail を return した。これは ARC v0.3 の

pipeline 設計（experiment + code 前提）と review-synthesis 型 topic の不整合という構造的問題であり、ARC の不具合ではなく **scope condition の honest な顕在化** である。

この事象自体が本論文の主張する「タスクと validity strategy のミスマッチが結論を分岐させる」（§4.5）の実演であり、recursive limitation の自己例として §13.5 で議論した。

Appendix A3: 略語一覧

- ABM: Agent-Based Modeling
- ANES: American National Election Study
- COREQ: Consolidated Criteria for Reporting Qualitative Research
- DIF: Differential Item Functioning
- ELSI: Ethical, Legal, and Social Implications
- GABM: Generative Agent-Based Modeling
- GSS: General Social Survey
- IRR: Inter-Rater Reliability
- JTBD: Jobs-to-be-Done
- LLM: Large Language Model
- MRP: Multilevel Regression and Poststratification
- OSF: Open Science Framework
- RLHF: Reinforcement Learning from Human Feedback
- WEIRD: Western, Educated, Industrialized, Rich, Democratic

End of paper.