AI Safety & Bias Mitigation Controls

Last Updated: March 11, 2025

1. Introduction

Synthetic Users generates AI-powered interview participants that organizations rely on to inform product and research decisions. The quality and fairness of those outputs directly impacts the quality of the decisions made from them.

This document details the technical and procedural controls we use to identify, measure, and mitigate bias in our AI systems, and to ensure the safety and reliability of platform outputs.

2. Sources of Bias in Synthetic Research

Bias in AI-generated research can originate from multiple sources. Understanding these sources is essential to mitigating them effectively:

Source	Description	Risk
Model training data	LLMs inherit biases present in their training corpora, including demographic stereotypes, cultural blind spots, and majority-culture defaults.	Synthetic participants may over-represent dominant perspectives and under-represent minority viewpoints.
Single-model dependence	Relying on one LLM amplifies that model's specific biases and behavioral patterns.	Homogeneous, predictable outputs that lack the variability of real human responses.
Prompt design	Poorly structured or leading prompts can steer outputs toward predetermined conclusions.	Research findings that confirm existing assumptions rather

		than surfacing genuine insights.
Persona specification	Vague or stereotypical persona definitions can produce shallow, caricatured responses.	Outputs that reinforce stereotypes rather than representing authentic perspectives.
Researcher interpretation	Even unbiased outputs can be selectively interpreted or over-generalized.	Misapplication of synthetic research findings to real-world decisions.

---

3. Multi-Model Evaluation (LLM Shuffle)

Our primary architectural control against bias is the use of multiple frontier LLMs to generate synthetic participants.

How It Works

- Each study distributes interview generation across multiple models, including offerings from OpenAI, Anthropic, Google, Meta, and Mistral.
- Model selection is diversified across participants within a study so that no single model dominates the output.
- Different models bring different linguistic patterns, reasoning approaches, and implicit perspectives, producing a richer and more varied set of responses.

Why It Matters

- **Reduces systematic bias:** Individual model biases are diluted when outputs are drawn from multiple sources.
- **Increases output diversity:** Responses exhibit greater variability in tone, reasoning style, and perspective.
- **Prevents model lock-in:** No single provider's limitations or failures can compromise the entire research output.

- **Enables comparative analysis:** When models produce divergent responses to the same prompt, it signals areas where human judgment and further investigation are warranted.

Monitoring

- We track output distribution across models to ensure balanced allocation.
- We evaluate for convergence patterns that may indicate shared biases across models on specific topics.
- Model allocations are adjusted when evaluation reveals disproportionate bias from a specific provider.

---

4. Structured Persona Architectures

Synthetic Users employs psychologically grounded persona definitions to produce realistic, consistent, and non-stereotypical interview participants.

OCEAN-Based Personality Framework

Each synthetic participant is defined using the OCEAN (Big Five) personality model:

- **Openness** — Receptivity to new ideas, creativity, intellectual curiosity.
- **Conscientiousness** — Organization, dependability, discipline.
- **Extraversion** — Sociability, assertiveness, energy level.
- **Agreeableness** — Cooperation, empathy, trust.
- **Neuroticism** — Emotional sensitivity, stress response, anxiety.

How Personas Reduce Bias

- **Grounded in psychology:** OCEAN traits are empirically validated and culturally studied, providing a more rigorous foundation than ad hoc persona descriptions.
- **Behavioral consistency:** Trait-based definitions produce participants whose responses are internally consistent across an interview, rather than shifting based on model defaults.

- **Diversity by design:** Varying OCEAN parameters across participants within a study ensures a range of perspectives, communication styles, and decision-making patterns.
- **Stereotype resistance:** Structured traits prevent the model from filling in unspecified attributes with stereotypical defaults (e.g., assuming demographic characteristics from occupation or interest descriptions).

Audience Definition Controls

Customers define their target audience through structured inputs:

- Demographic parameters (age range, location, profession, etc.)
- Behavioral characteristics and context
- Problems, goals, and pain points relevant to the research

These structured inputs constrain the model's generation space, reducing the likelihood of irrelevant or stereotypical outputs.

---

5. Controlled Prompts Through Research Agents

Prompt quality is one of the most significant determinants of output quality. Our Dedicated Research Agents act as a control layer between the researcher and the LLMs.

Research Planning Agent

- Recommends appropriate interview methodologies based on the research objective.
- Suggests sampling strategies to ensure adequate representation across the target audience.
- Flags potential gaps or biases in the proposed research design.

Question Design Agent

- Assists in formulating interview questions that are open-ended, neutral, and non-leading.

- Identifies questions that may prime the synthetic participant toward a specific response.
- Structures question sequences to build context progressively, reducing order-effect bias.

Analysis Agent

- Summarizes and clusters interview findings across participants.
- Highlights areas of consensus and divergence.
- Flags outputs that appear anomalous, generic, or inconsistent with the defined persona.

Why Agent-Mediated Prompts Matter

- Researchers with varying levels of experience receive consistent methodological guidance.
- Common prompt design errors (leading questions, double-barreled questions, loaded framing) are caught before they reach the model.
- The research process is more reproducible and auditable.

---

6. Human-Reviewable Outputs

Every output generated by the platform is designed to be transparent and reviewable:

Full Transcript Access

- Customers can read complete interview transcripts, not just summaries or extracted insights.
- Individual responses can be examined in the context of the full conversation flow.

Follow-Up Questioning

- Researchers can ask clarifying or probing follow-up questions to any synthetic participant.

- This allows testing of response consistency and depth — a key mechanism for identifying shallow or generic outputs.

Knowledge Graphs

- Research data is visualized through Knowledge Graphs that map themes, relationships, and patterns.
- Researchers can trace aggregated insights back to specific participant responses.
- Anomalous clusters or unexpected patterns are surfaced visually for human review.

Reports with Interactive Follow-Up

- Generated reports can be questioned and refined through follow-up queries.
- Researchers can request alternative perspectives, challenge findings, or drill into specific themes.

PRISMA™ Multi-Study Orchestration

- Cross-study consistency checks identify when findings from different studies conflict or converge.
- Global reports aggregate findings while preserving traceability to individual studies and participants.

---

7. Content Safety Controls

We maintain multiple layers of content safety to prevent harmful outputs:

Input Safeguards

- Study configurations and audience definitions are validated against our Acceptable Use Policy.
- Inputs designed to generate harmful, illegal, discriminatory, or explicitly offensive content are rejected.

- Structured input fields constrain the scope of generation, reducing the attack surface for prompt manipulation.

Output Safeguards

- All LLM providers used by Synthetic Users maintain their own content safety filters, which apply before outputs reach our platform.
- Platform-level output monitoring identifies responses that may violate content policies.
- Outputs that contain personally identifiable information, harmful instructions, or discriminatory content are flagged for review.

Usage Monitoring

- Automated systems monitor for patterns indicative of misuse, including attempts to generate deceptive content, fabricate evidence, or manufacture false consensus.
- Accounts exhibiting suspicious behavior are flagged for review and may be suspended pending investigation.

---

8. Evaluation and Testing

We conduct ongoing evaluation of our AI systems across multiple dimensions:

Output Quality Evaluation

- Periodic assessment of persona fidelity: do synthetic participants behave consistently with their defined traits?
- Conversational realism: do responses feel natural, contextually appropriate, and sufficiently varied?
- Depth and specificity: do participants provide substantive responses or default to generic, surface-level answers?

Bias Evaluation

- Comparative analysis of outputs across models to identify systematic differences in how demographics, cultures, or viewpoints are represented.
- Review of outputs for common bias patterns: stereotype reinforcement, majority-culture defaults, gender or racial assumptions.
- Assessment of whether persona diversity parameters produce genuinely different perspectives or superficially varied versions of the same viewpoint.

Safety Evaluation

- Testing of content safety filters against adversarial inputs.
- Review of edge cases where safety controls may be overly permissive or overly restrictive.
- Monitoring of new model releases for changes in safety behavior.

---

9. Limitations and Honest Disclosure

We are transparent about the inherent limitations of synthetic research:

- **Synthetic participants are not real people.** They simulate human perspectives based on model training data and persona definitions, but they cannot replicate lived experience, genuine emotion, or true cultural immersion.
- **Outputs reflect model capabilities.** The quality and diversity of synthetic responses are bounded by the capabilities and biases of the underlying LLMs, even with multi-model generation.
- **Not a replacement for all human research.** Synthetic research is most effective as a complement to traditional methods — for rapid iteration, hypothesis generation, and early-stage exploration. High-stakes decisions that affect people's lives should incorporate research with real participants.
- **Bias cannot be fully eliminated.** Our controls significantly reduce bias but cannot eliminate it entirely. We encourage customers to critically evaluate all outputs.

---

10. Related Documents

- Responsible AI & Risk Management Overview
- AI System Architecture & Model Usage
- Acceptable Use Policy
- Product Features
- Company Code of Conduct

11. Contact

For questions about our safety and bias mitigation controls, contact us at support@syntheticusers.com.

Synthetic Users, Inc. 4223 Glencoe Ave, Suite C215-523, Marina del Rey CA 90292