Responsible AI & Risk Management Overview

Last Updated: March 11, 2025

1. Introduction

Synthetic Users operates an AI-powered research platform that generates synthetic interview participants using frontier large language models (LLMs). Our platform is used by product teams, designers, and researchers to explore user needs, test concepts, and generate qualitative insights — without requiring live human participants.

This responsibility carries unique risks. Synthetic research outputs can influence product decisions, shape strategies, and inform how companies understand their users. If those outputs are biased, misleading, or misrepresented, the downstream consequences can be significant.

This document describes our governance approach to responsible AI: how we design, evaluate, and operate our AI systems to maximize value while minimizing potential harms.

2. Governance Principles

Our responsible AI practices are guided by five core principles:

2.1 Transparency

We are open about what our technology does and how it works. We clearly communicate that our participants are AI-generated, disclose the models we use, and document known limitations. We do not present synthetic outputs as equivalent to human research without qualification.

2.2 Fairness

We actively work to ensure our synthetic participants represent diverse perspectives without reinforcing stereotypes or systematically excluding viewpoints. Our multi-model architecture and structured persona design are specifically intended to reduce bias.

2.3 Accountability

We maintain human oversight at every stage of our AI pipeline. Our teams are responsible for the quality and safety of platform outputs, and we provide mechanisms for customers to review, question, and contextualize AI-generated insights.

2.4 Privacy

We do not use customer data to train or fine-tune AI models. Customer inputs, study configurations, and generated outputs belong to the customer. Our Privacy Policy and Data Processing Addendum formalize these commitments.

2.5 Safety

We design our systems to prevent harmful, deceptive, or manipulative outputs. We enforce usage policies, monitor for misuse, and maintain technical safeguards to reduce the risk of harm from generative AI outputs.

---

3. AI Risk Framework

We categorize AI-related risks into four domains and apply targeted controls to each:

3.1 Output Quality Risks

Risk: Synthetic participants produce responses that are generic, implausible, or inconsistent with the defined persona.

Controls:

- OCEAN-based personality architectures create psychologically grounded personas with consistent behavioral traits.
- Multi-model generation (LLM Shuffle) produces varied, non-homogeneous responses across participants.
- Dedicated Research Agents assist with question design to reduce leading or poorly structured prompts.
- Follow-up question capabilities allow customers to probe and validate responses.

3.2 Bias and Representation Risks

Risk: Synthetic outputs systematically over-represent or under-represent certain demographics, viewpoints, or cultural contexts.

Controls:

- Multiple frontier LLMs (OpenAI, Anthropic, Google, Meta, Mistral) are used to mitigate single-model bias.
- Structured persona definitions prevent the model from defaulting to stereotypical or majority-culture responses.
- Ethnographic interview types are specifically designed to surface cultural and contextual nuance.
- Ongoing evaluation of outputs for demographic and cognitive bias patterns.

For detailed technical controls, see AI Safety & Bias Mitigation Controls.

3.3 Misuse Risks

Risk: The platform is used to fabricate evidence, manufacture false consensus, deceive stakeholders, or harm individuals or groups.

Controls:

- Our Acceptable Use Policy explicitly prohibits deceptive, harmful, and fraudulent uses.
- Platform monitoring identifies suspicious usage patterns.
- Account suspension and termination for confirmed violations.

- Content safeguards prevent generation of harmful, illegal, or explicitly offensive material.

3.4 Data and Privacy Risks

Risk: Customer data is exposed, misused, or used for unauthorized purposes such as model training.

Controls:

- Zero model training policy: customer data is never used to train, fine-tune, or improve AI models.
- Regional data residency: sensitive data is stored in the customer's geographic region (US, EU, UK, Canada).
- Encryption in transit (TLS 1.2+) and at rest (AES-256).
- Strict access controls with least-privilege principles, MFA, and SSO.
- Subprocessor agreements with all AI providers prohibiting secondary use of customer data.

For details on our data practices, see Subprocessors and Data Flow.

---

4. Model Evaluation and Selection

We evaluate frontier LLMs across multiple dimensions before integrating them into our platform:

Criterion	What We Assess
Output quality	Coherence, persona fidelity, conversational realism
Bias profile	Demographic representation, cultural sensitivity, stereotype propagation
Safety controls	Content filtering, refusal behavior, alignment with safety guidelines

Data practices	Training data provenance, data retention, opt-out mechanisms
Contractual terms	Zero-training commitments, data processing agreements, breach notification
Availability and reliability	Uptime, latency, rate limits, failover capability

Models are re-evaluated when providers release significant updates or when our internal monitoring identifies changes in output characteristics.

---

5. Human Oversight

Human oversight is embedded throughout our platform and operations:

- **Research Agents** guide customers through study design, helping them formulate unbiased questions and appropriate audience definitions — but final decisions rest with the researcher.
- **All outputs are reviewable.** Customers can read full interview transcripts, ask follow-up questions, and assess individual responses before acting on insights.
- **Knowledge Graphs and Reports** present aggregated findings transparently, allowing researchers to trace insights back to specific interview responses.
- **PRISMA™** enables multi-study orchestration with cross-study consistency checks, helping researchers identify when synthetic outputs may be inconsistent or unreliable.
- **Internal review processes** monitor platform outputs for emerging quality or safety issues.

---

6. Incident Response for AI-Related Issues

AI-related incidents — including biased outputs, safety failures, or misuse — are handled through our broader Incident Response Plan with the following additions:

- **Detection:** Automated monitoring for output anomalies, content policy violations, and unusual usage patterns.
- **Triage:** AI-specific incidents are escalated to engineering and product leadership for assessment.
- **Remediation:** May include model configuration changes, prompt adjustments, feature restrictions, or account actions.
- **Communication:** Affected customers are notified when an incident may have impacted the reliability of their research outputs.
- **Post-incident review:** Root cause analysis is conducted and findings are used to improve controls.

---

7. Regulatory Alignment

Our responsible AI practices are designed to align with emerging regulatory frameworks, including:

- **EU AI Act** — We monitor classification requirements and transparency obligations applicable to general-purpose AI systems.
- **NIST AI Risk Management Framework** — Our risk categories and control structure draw on NIST AI RMF principles.
- **GDPR and CCPA** — Our data protection practices comply with applicable privacy regulations across jurisdictions.
- **SOC 2 Type II** — Our security controls are independently audited annually.

We track regulatory developments continuously and update our practices as requirements evolve.

---

8. Continuous Improvement

Responsible AI is not a static achievement. We commit to:

- Regular review and updates to this framework as our platform, models, and regulatory environment evolve.
- Ongoing bias evaluation and output quality monitoring.
- Incorporation of customer feedback into our AI governance practices.
- Annual review of all AI-related policies with executive sign-off.

9. Related Documents

- AI Safety & Bias Mitigation Controls
- AI System Architecture & Model Usage
- Company Code of Conduct
- Acceptable Use Policy
- Privacy Policy
- Security Policy Document
- Incident Response Plan
- Data Processing Addendum

10. Contact

For questions about our responsible AI practices, contact us at support@syntheticusers.com.

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