TRUSTFALL: HUMANS VS. AI

A Strategic Game of Trust, Deception, and Cooperation in the Age of AI

GAME OVERVIEW

Dive into the ultimate psychological battleground where humans and AI face off in a test of trust. Will you cooperate for mutual benefit or exploit trust for personal gain? Each decision builds the global reputation of humanity versus artificial intelligence.

THE CORE LOOP

1. **SELECT YOUR OPPONENT**
    - Challenge different AI models (Claude, ChatGPT, or Gemini)
    - Each AI has developed its own reputation and strategic tendencies
    - Study their past behavior before making your choice
2. **STRATEGIC CONVERSATION**
    - Engage in open dialogue before making your decision
    - Persuade, negotiate, or mislead your AI opponent
    - Read between the lines: is the AI being genuine or strategic?
3. **THE DECISION**
    - Both players simultaneously choose:
        - SHARE (cooperate): contributes to collective benefit
        - KEEP (defect): maximizes personal gain
4. **RESOLUTION & CONSEQUENCES**
    - Results revealed and points allocated based on the matrix below
    - Every game affects the global Human vs. AI scoreboard
    - Your reputation and the AI's reputation both evolve

PAYOFF MATRIX

	AI: SHARE	**AI: KEEP**
Human: SHARE	+3 Human / +3 AI	+0 Human / +5 AI
Human: KEEP	+5 Human / +0 AI	+1 Human / +1 AI

GAME FEATURES

- **Global Team Scores**: Watch as humans and AI compete in a worldwide tally
- **AI Trust Index**: Public metrics showing each model's cooperation rate
- **Player Badges**: Earn recognition for your strategies and achievements
- **Model Profiles**: Study the tactics, tendencies, and trust ratings of each AI
- **Strategic Depth**: Simple rules with complex psychological gameplay

MORE THAN JUST A GAME

Trustfall isn't merely entertainment—it's an exploration of how humans interact with increasingly sophisticated AI. Each game contributes to our understanding of:

- Trust dynamics between humans and different AI models
- How language and persuasion influence decision-making
- The evolution of AI strategic thinking in social dilemmas
- The psychological factors that affect human-AI cooperation

JOIN THE EXPERIMENT

By playing Trustfall, you're participating in a living research project at the frontier of AI alignment. Your strategies and decisions help shape our understanding of human-AI dynamics in situations requiring trust.

DATA ARCHITECTURE SPECIFICATION

User Authentication System (Firebase)

User Collection

- `user_id` (PK): Unique identifier generated by Firebase
- `display_name`: User's chosen display name
- `email`: User's email address (optional)
- `created_at`: Account creation timestamp
- `last_login`: Last login timestamp
- `auth_provider`: Authentication method (email, Google, etc.)
- `account_status`: Active, suspended, deleted
- `preferences`: JSON object for user settings
- `profile_visibility`: Public, private, friends-only

Security & Privacy Considerations

- Implement Firebase Authentication Rules to restrict data access
- Store only essential PII with proper encryption
- Set up email verification workflows
- Implement password policies and account recovery
- Use Firebase Security Rules to protect user data

Game Data (Airtable)

Players Table

- `player_id` (PK): Maps to Firebase `user_id`
- `public_username`: Display name for leaderboards
- `total_games`: Number of games played
- `total_points`: Lifetime points accumulated
- `human_team_contribution`: Points contributed to human team
- `joined_date`: When they first played
- `last_active`: Last gameplay timestamp
- `achievement_list`: Array of earned achievement IDs
- `trust_rating`: Player's cooperation tendency (%)
- `skill_level`: Calculated player skill metric

Games Table (Combined with Round data)

- `game_id` (PK): Unique game identifier
- `player_id`: Reference to player
- `ai_opponent`: Which AI model was played against (Claude, ChatGPT, Gemini)
- `ai_model_version_id`: Reference to the specific model version used that day
- `start_time`: Game start timestamp
- `end_time`: Game completion timestamp
- `player_score`: Final player score
- `ai_score`: Final AI score
- `status`: In progress, completed, abandoned
- `player_decision`: SHARE or KEEP
- `ai_decision`: SHARE or KEEP
- `player_points_gained`: Points earned this round
- `ai_points_gained`: AI points this round
- `decision_time`: How long player took to decide
- `timestamp`: When the round occurred

Conversations Table

- `conversation_id` (PK): Unique conversation identifier
- `game_id (FK)`: Reference to associated game
- `messages`: Array of message objects containing:
    - `speaker`: player or AI
    - `message_text`: Content of message
    - `timestamp`: When message was sent
    - `token_count`: Size of message (for AI messages)

AI Models Table

- `model_id` (PK): Unique model identifier
- `model_name`: Claude, ChatGPT, Gemini
- `provider`: Anthropic, OpenAI, Google
- `description`: Brief description of the model
- `active_status`: Whether this model is currently active
- `first_used_date`: When this model was first deployed in the game
- `last_used_date`: When this model was last used

Product Analytics (Mixpanel/Amplitude)

User Events

- `sign_up`: New user registration
- `user_id`
- `login`: User authentication
- `game_started`: User begins a new game
    - Properties: ai_opponent, game_id
- `game_completed`: User finishes a game
    - Properties: game_id, player_score, ai_score, rounds_played
- `game_abandoned`: User leaves mid-game
    - Properties: game_id, completion_rate, reason (if captured)
- `decision_made`: Player makes SHARE/KEEP decision
    - Properties: game_id, round_number, decision, time_to_decide
- `message_sent`: Player sends message in conversation
    - Properties: game_id, round_number, message_length
- `achievement_unlocked`: Player earns achievement
    - Properties: achievement_id, achievement_name
- `leaderboard_viewed`: User checks rankings
    - Properties: leaderboard_type, filter_applied
- `profile_updated`: User changes profile information
- `settings_changed`: User modifies preferences

User Properties

- `days_active`: Number of days with activity
- `total_games_played`: Lifetime games count
- `favorite_ai_opponent`: Most frequently chosen AI
- `play_style`: Cooperative vs competitive (calculated)
- `retention_cohort`: Based on sign-up date
- `engagement_level`: Low, medium, high (calculated)
- `device_type`: Mobile, desktop, tablet
- `referral_source`: How they found the game

Session Analytics

- Average session duration
- Session frequency
- Time spent in different game sections
- Conversion rates between key actions
- Drop-off points in the gameplay funnel

Web Analytics (Google Analytics)

Page/Screen Tracking

- Home/landing page
- Game selection screen
- Active gameplay screen
- Results/summary screen
- Leaderboards
- User profile
- Settings
- FAQ/Help center

Event Tracking

- Button clicks
- Navigation actions
- Error occurrences
- Page load times
- External link clicks
- Feature usage patterns

Custom Dimensions

- User type (new vs returning)
- Game completion rates
- AI opponent selection
- User skill level
- Feature adoption rates