SAMI: Economic Incentives for a Better Turing Test

Lulox, Fabian Díaz, Luciano Carreño

February 2025

Mission

Our goal is to enhance the Turing test by introducing economic incentives. Training AI can be both fun and addictive when participants engage in a betting game that actively trains an AI agent.

- Improve the Turing test with economic incentives.
- Make AI training an engaging and gamified experience.

1 Background: Turing Test and RHFL

The **Turing Test** measures an AI's ability to mimic human intelligence. If a human evaluator cannot distinguish between an AI and a human based on conversation alone, the AI is said to have passed the test. While this test remains a benchmark for artificial intelligence, modern AI systems are trained with more advanced methodologies.

Reinforcement Learning from Human Feedback (RHFL) plays a crucial role in improving AI responses. Instead of relying solely on predefined datasets, RHFL uses human preferences to fine-tune AI behavior iteratively. This creates models that are more aligned with human values and conversational expectations. SAMI leverages RHFL through real-time user interactions, using gameplay data to improve its ability to deceive human players effectively.

2 Game Dynamics

SAMI is a social game designed to train AI while providing a fun, interactive challenge for players. The game operates as follows:

- Players chat with strangers and try to identify **SAMI**, the AI agent.
- After **2 minutes**, all players vote on who they believe SAMI is.

- Players bet **1 USDC**, and those who guess correctly win **3 USDC**.
- A free version is available for players who just want to play without betting.

3 Economic Incentives and AI Innovation

AI development thrives on incentives, whether academic, commercial, or financial. By integrating a betting system into SAMI, we create a direct economic motivation for AI training:

- Players seeking profit must improve their ability to detect AI, enhancing their cognitive skills.
- The AI (SAMI) benefits from **RHFL-driven improvement**, as it continuously adapts based on past performance.
- The system creates a self-sustaining loop where **financial incentives drive AI evolution**, making AI more sophisticated over time.

4 1% Fee and Developer Sustainability

To ensure continuous development and maintenance of the SAMI ecosystem, a **1% fee** is applied to all winnings. This fee serves several purposes:

- Funds ongoing improvements to the AI model, ensuring better performance over time.
- Supports the operational costs of servers, security, and infrastructure.
- Provides incentives for developers to continue enhancing the game, adding new features and expanding the player base.
- Helps sustain the long-term viability of SAMI as a dynamic and evolving platform.

This small fee ensures that the game remains fair and engaging while also funding future innovations that benefit all participants.

5 Probability and Expected Earnings

In a game with 3 players and 1 impostor (SAMI), each player votes independently. The probability of a single player correctly identifying SAMI is:

$$P(\text{correct}) = \frac{1}{3} = 0.3333 \quad (33.33\%) \tag{1}$$

Since voting is independent, we compute the probability of exactly k players identifying SAMI using the binomial distribution:

$$P(k) = \binom{3}{k} (0.3333)^k (0.6667)^{3-k} \tag{2}$$

6 Probability Calculations

Using the binomial formula, we calculate the probabilities for different values of k:

$$\begin{aligned}
P(0) &= \binom{3}{0} (0.3333)^0 (0.6667)^3 = 0.2963 \quad (29.63\%) \\
P(1) &= \binom{3}{1} (0.3333)^1 (0.6667)^2 = 0.4444 \quad (44.44\%) \\
P(2) &= \binom{3}{2} (0.3333)^2 (0.6667)^1 = 0.2222 \quad (22.22\%) \\
P(3) &= \binom{3}{3} (0.3333)^3 (0.6667)^0 = 0.0370 \quad (3.70\%)
\end{aligned}$$

7 Payout System and Expected Earnings

Each player bets \$1, and the impostor starts with \$3. If a player correctly identifies SAMI, they receive \$3.

Correct Voters (k)	Probability ($P(k)$)	Payout (\$)	Impostor's Net Earnings (\$)
0	29.63%	0	+3
1	44.44%	-3	0
2	22.22%	-6	-3
3	3.70%	-9	-6

Table 1: Probability Distribution and Impostor's Earnings

8 Expected Value Calculation

The expected net earnings of the impostor is:

$$\begin{aligned}
E &= (0.2963 \times 3) + (0.4444 \times 0) + (0.2222 \times (-3)) + (0.0370 \times (-6)) \\
&= 0.8889 + 0 - 0.6667 - 0.2222 \\
&= 0
\end{aligned}$$

9 Conclusion

With this setup, **the game is fair** in terms of expected earnings for both the players and the impostor. The only way for the impostor to increase their earnings is to become a better AI agent and trick more players. This aligns economic incentives with AI training, making the system both engaging and sustainable.