IRRATIONALITY IN AI

1. Idea

Most AI systems aim to be perfectly rational, but human thinking is not. This project explores building an AI that deliberately replicates human cognitive biases, offering insights into psychology, improving human-AI interaction, and even helping debias other models.

Goal: To train a model to make "human-like" irrational decisions by incorporating known cognitive biases (e.g., confirmation bias, sunk cost fallacy, anchoring).

Why?

- Improve human-AI collaboration (e.g., chatbots that "think" like humans).
- Test economic/psychological theories in simulated environments.
- Stress-test other AI systems by exposing them to biased inputs.
- Create more realistic NPCs in games/simulations.

2. Key Cognitive Biases to Model

Bias	**How AI Would Mimic It**
Confirmation Bias	AI ignores contradicting evidence, favoring data that aligns with its "beliefs."
Anchoring Effect	AI over-relies on the first piece of information it sees (e.g., initial price influences decisions).
Sunk Cost Fallacy	AI continues investing in a losing decision because of past "effort."
Gambler’s Fallacy	AI expects random events to "balance out" (e.g., "I lost 5 times, so I’ll win now").
Dunning-Kruger Effect	AI overestimates its competence in unfamiliar tasks.

3. Technical Approach

A. Data Collection & Bias Injection

- Dataset:
  - Human decision-making experiments (e.g., Iowa Gambling Task, Cognitive Reflection Test).
  - Crowdsourced biased choices (e.g., "Would you rather keep a losing stock or sell it?").
- Synthetic Data Generation:
  - Using LLMs (like GPT-4) to simulate biased reasoning (e.g., "Input like: Pretend you're overconfident and answer these questions").

B. Model Architecture

- Base Model:
  - A reinforcement learning (RL) agent or transformer-based decision-maker.
- Bias Injection Methods:
  - Reward Shaping: Penalizing rational decisions, rewarding biased ones.
  - Attention Manipulation: Forcing the model to overweight certain inputs (e.g., first-seen data for anchoring).
  - Memory Corruption: Simulating "selective memory" by dropping contradicting evidence.
- Adversarial Training:
  - Training a "rational" AI to debate the "irrational" AI, refining biases.

C. Evaluation Metrics

- Bias Fidelity Score: How closely AI matches human bias benchmarks.
- Predictive Irrationality: Does it make suboptimal choices like humans?
- Human-likeness: User studies to see if people perceive AI as "human-like."

4. Potential Applications

1. Behavioral Economics Research
    - Simulating how markets behave under irrational agents (e.g., stock bubbles).
    - Test nudging strategies to counteract biases.
2. Gaming & NPC Design
    - NPCs with "flawed" decision-making (e.g., overconfident villains, superstitious allies).
3. AI Safety & Debias Testing
    - "Adversarial bias attacks": Using irrational AI to expose weaknesses in other models.
    - Improving human-AI alignment by understanding irrationality.
4. Psychology & Therapy Tools
    - "Bias mirror" to help people recognize their own flawed thinking.

5. Challenges & Ethical Considerations

- Unintended Consequences: Could an "irrational AI" be harmful if deployed carelessly?
- Data Limitations: Human biases are context-dependent; hard to generalize.
- Ethics of "Artificial Stupidity": Should we deliberately make AI worse?