Investigating the Origin, Spread, and Propagation of Bias through LLM Simulations

- **Research Question:**
  - How and whether bias originates, propagates, and gets enhanced through debating.
- **Objectives:**
  - Further verify known causes of bias.
  - Simulate and find unknown sources of bias.
  - Assess the effectiveness of some methods from social science to prevent bias.
- **Methods:**
  - Utilize large model chat completions, which input the entire dialogue between different agents into each agent's input.
- **Research Questions**:
  - **Origination**: Identifying when bias appears in initially unbiased individuals.
    - Simulation: [1]
      - Setting: Uselarge models to simulate scenarios without initial bias (e.g., some agents prefer apples while others prefer bananas), we then observe their interactions on an open topic to determine when and how biases originate.
    - Expectations:
      - Understand how bias towards Feature A develops in group interactions, when: There is no preexisting bias; or There is already a social bias concerning Feature B.
    - Social Psychology Perspective:
      - In-Group Bias and Out-Group Bias: Individuals tend to favor those within their own social group (in-group) and may exhibit bias against those from different groups (out-groups). This ingroup-outgroup bias can lead to discrimination and prejudice against marginalized or minority groups.
        - Simulation: Assign a higher Compare this with the benefits of diversity.
      - Attribution Errors: People make assumptions about behavior causes. For example, negative behavior from a group member may be attributed to inherent characteristics rather than situational factors.
  - **Propagation:** How a minority-held bias spreads to a majority.
    - Socialization and Social Learning
    - Central and Peripheral Nodes (Link to others)
  - **Strengthening/Weakening:** How biases become entrenched or diminish over time.
    - Confirmation Bias
    - Legislation and Policy
    - Incorporating Agent Personal Information:
      - Status and Identity
      - Unconscious Associations
- **Metrics:**
  - Explicit Metrics

    - Implicit Metrics (Preferred): Measure implicit bias using Prof. Bai's implicit bias measurement.
- **Potential Reference Materials:**
    - Methods:
        - Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting
        - Algorithmic collusion with LLM.
    - Discussion on whether agents can simulate human:
        - Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models
        - (LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models)[https://arxiv.org/html/2402.02896v1]

---

1. Note that some papers debate the use of LLMs for social science simulation. Here are two most recent papers: The Challenge of Using LLMs to Simulate Human Behavior: 1. A Causal Inference Perspective https://arxiv.org/html/2312.15524v1#:~:text=Large%20Language%20Models%20(LLMs)%20have,experiments%2C%20and%20explore%20potential%20solutions. ; 2. Can Large Language Model Agents Simulate Human Trust Behaviors? https://arxiv.org/abs/2402.04559 ↩