Can causal structure bias causal strength perceptions?

- **Normatively**, the number of a cause's effects should have **no** influence on its power to create each.
- **Dilution** (Stephan et al., 2023): A cause's perceived power over an effect **decreases** with more effects.
- **Boon-Bane Effect** (Sassman et al., 2020): Perceived power increases instead if the effects are **negative** (e.g., disease symptoms).
- Assuming variations among effects may explain the discordant findings (Park & Sloman, 2013).
- Findings could be **limited to Common Cause nets**.
- We manipulate network structure to adjudicate between the accounts.

Tested Structures

<table>
<thead>
<tr>
<th>Tested Structures</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain: A→C→B</td>
<td>$P(A, B, C) = \frac{P(B</td>
</tr>
<tr>
<td>Common Cause (CC): A→B</td>
<td>$P(A)P(B</td>
</tr>
<tr>
<td>Control: A→B</td>
<td>$P(A)P(B)$</td>
</tr>
</tbody>
</table>

Judgement Predictions of Related Works

- **Influence on its power to create each**.
- **Supported** by previous studies.

**LLMs:** Is language the main vehicle for deviations from normativity?

- Causal info is shared via **language** but may also be learned by interacting with the world.
- LLMs are trained on language. If they show a bias, language must be a vehicle for it.
- Prior studies show suboptimal LLM causal reasoning (Binz & Schulz, 2023; Willig et al., 2022).

**Human Data**

N= 320 US and UK residents (122 males) average age = 37.28 years (SD = 13.12; range: 18 to 76).

**LLM Data**

- Queried GPT3.5-Turbo (OpenAI, 2022), GPT4 (OpenAI, 2023), and Luminous Supreme Control (Aleph Alpha, 2023).
- Manipulated temperature to compare deterministic and non-deterministic responses with human data.
- Looked for sampling parameters that fit best:
  - 1. The human data
  - 2. The normative model.
  - Using Wasserstein Distance to compute distributions.

**Discussion**

- **The causal structure** (Chain vs Common Cause) changes causal intuitions.
- Both human participants and Large Language Models (LLMs) deviated from normativity by judging **intermediate causes** in causal chains as more potent than simple causation or Common Causes.
- Variations in LLM hyperparameters revealed that models with higher temperatures, which incorporate more randomness, showed **biases in the human judgments**.
- Possible explanations:
  - **“Mechanisms Hypothesis”**: middle nodes may be seen as mechanisms for the initial causes (Menzies, 2012). Mechanistic causes are preferred over correlational ones (Johnson & Ahn, 2017).
  - **“Causal Relay Hypothesis”**: the strength of the C→B link in a chain is supported by the A→C sequence, indicating that the perceived causal strength might be influenced by the support provided by preceding causes in the chain.

**Future Work**

- Probabilistic manipulation (A→C) in a chain to **differentiate between the Mechanisms and the Causal Relay Hypothesis**.
- Asking subjects whether they see the intermediate node in a chain as a **mechanism**.
- Examining the embedding space for clues to LLM representations that mimic human biases.
- Examining whether exposure to normative Bayesian reasoning could help improve the reliability of AI in domains requiring precise causal judgments.
- Future research should explore if different **architectures and training methods** result in more, or less biased causal reasoning.
- Studies should examine whether **increasing temperature** always induces human-like biases in LLM causal reasoning.