Causal Strength Judgments in Humans and Large Language Models

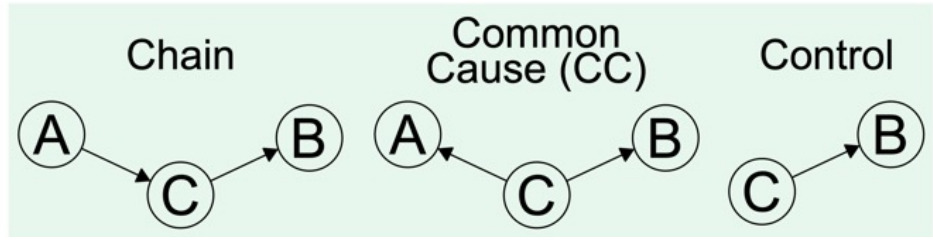
Anita Keshmirian[1,2,3], Moritz Willig[4], Babak Hemmatian[5], Ulrike Hahn[2,6], Kristian Kersting[4,7,8], Tobias Gerstenberg[9]

[1]Fraunhofer IKS, Munich [2]Munich Center for Mathematical Philosophy [3]Forward College, Berlin [4]Technical University of Darmstadt [5]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign [6]Birkbeck University, London [7]Hessian Center for AI [8]German Research Center for AI [9]Stanford University, Palo Alto

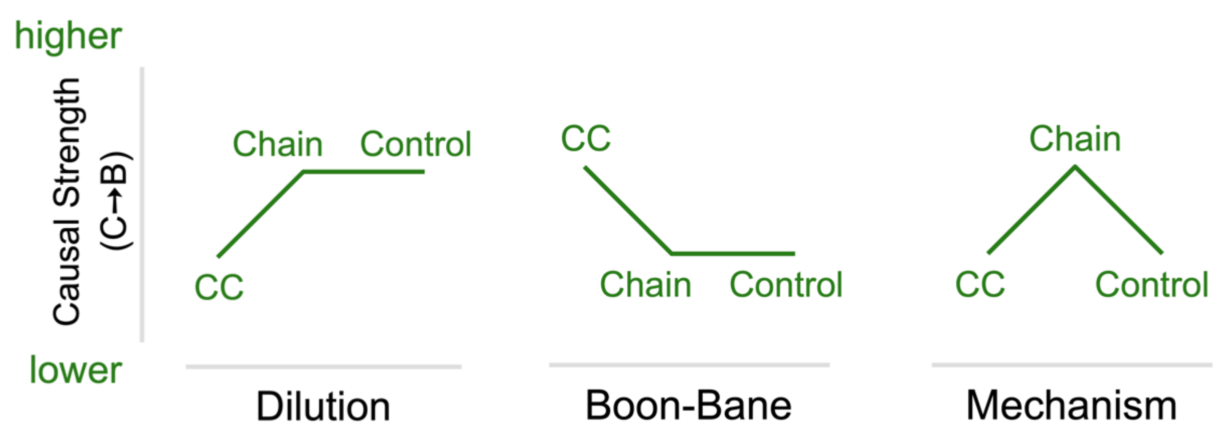
Can causal structure bias causal strength perceptions?

- **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** (Sussman et al., 2020): Perceived power **increases** instead if the effects are <u>negative</u> (e.g., disease symptoms)
- Assuming **interactions** 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

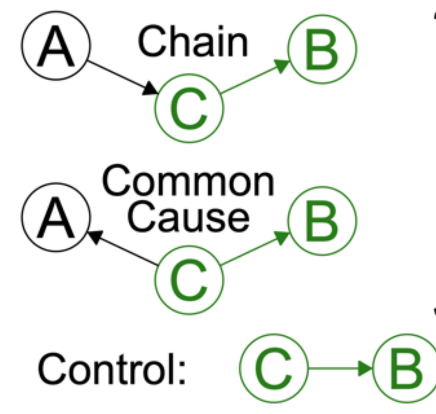
Tested Structures



Judgement Predictions of Related Works



Normative Prediction



$\mathbf{A} \perp \mathbf{B}|\mathbf{C}$

$$P(\mathbf{A},\mathbf{B},\mathbf{C}) = P(\mathbf{B}|\mathbf{C})P(\mathbf{C}|\mathbf{A})P(\mathbf{A}) = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})P(\mathbf{C})$$

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).

Method

Design: Scenarios within-subject, Structure between-subject

Materials: Three scenarios (novel + adapted from previous studies): Alien, sex-work, economy

The same variables were used across different network structures

Example Scenario

Economy

Adapted from Rehder (2014) for a moderate level of familiarity with the domain:

Chain: High-interest rates lead to more loan defaults, which leads to more inflation.

Common cause (Generative): More loan defaults lead to high-interest rates on the one hand and more inflation on the other.

Common cause (Preventive): More loan defaults prevent low-interest rates on the one hand and prevent retirement investment on the other.

Control (Generative): More loan defaults lead to more inflation.

Control (Preventive): More loan defaults prevent retirement investment.

Human Data

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

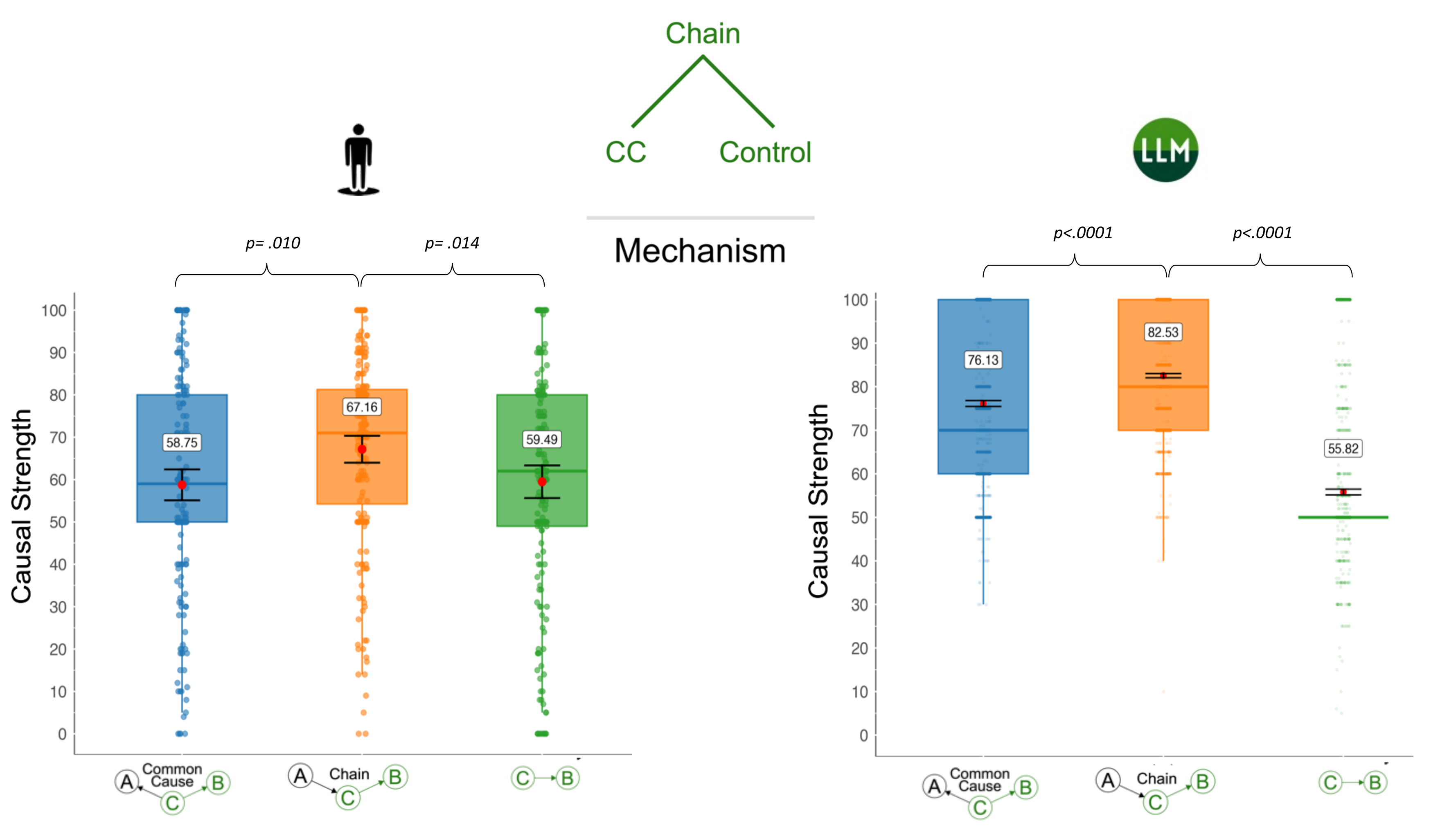
LLM Data

- 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.

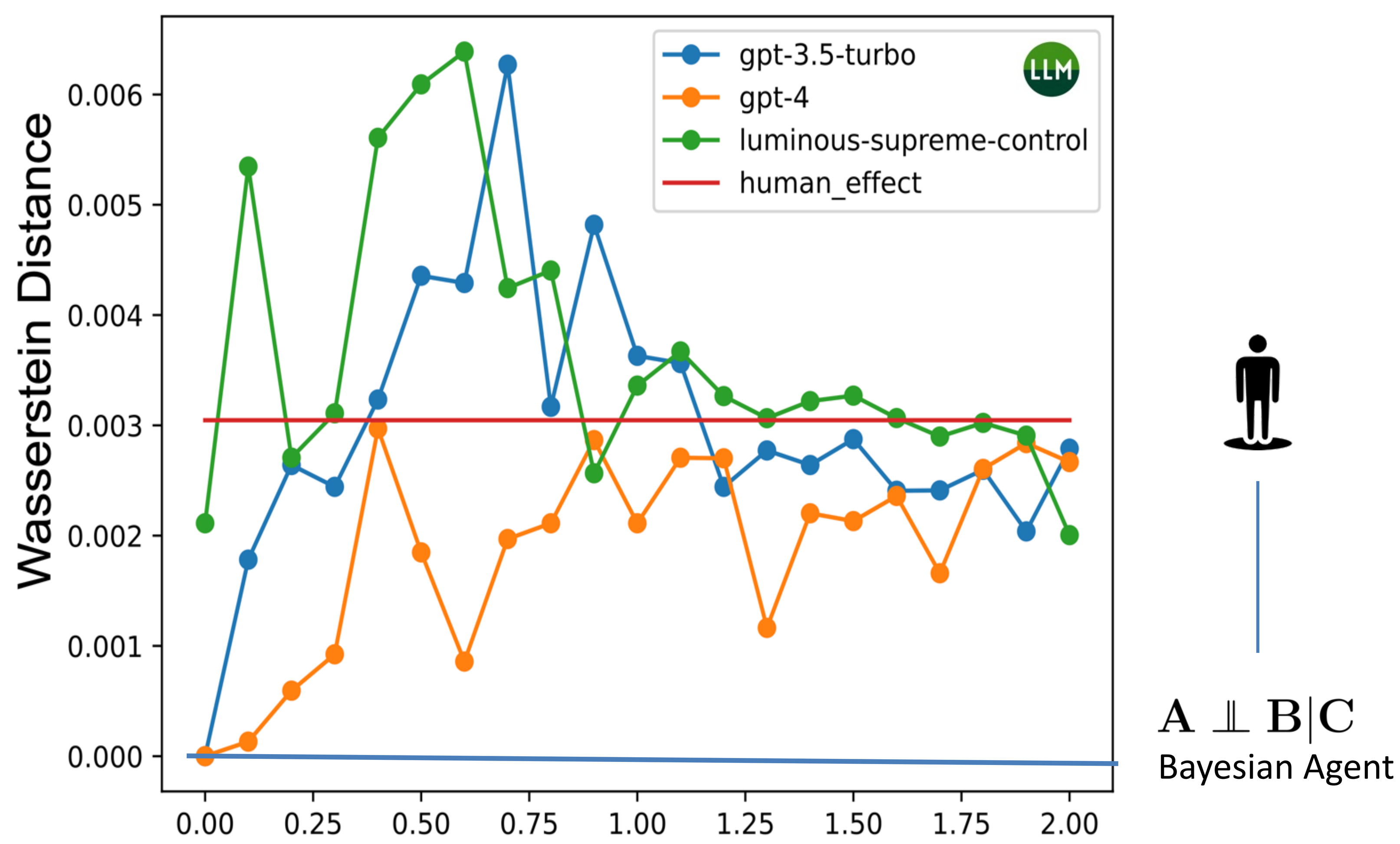
Using **Wasserstein Distance** to compare distributions

Results

Causal Power Judgment Across Structures



Preference for Chains Across Temperature Values



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 similar to 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.