UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Chain Versus Common Cause: Biased Causal Strength Judgments in Humans and Large Language Models

Permalink
https://escholarship.org/uc/item/2w45f1bv

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Keshmirian, Anita
Willig, Moritz
Hemmatian, Babak
et al.

Publication Date
2024

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Chain Versus Common Cause: Biased Causal Strength Judgments in Humans and Large Language Models

Anita Keshmirian (Anita.Keshmirian@gmail.com)
Forward College Berlin; Fraunhofer Institute for Cognitive Systems (IKS); LMU Munich University, Germany

Moritz Willig
Computer Science Department, Technical University of Darmstadt, Germany

Babak Hemmatian
University of Illinois at Urbana-ChampaignUniversity Urbana-Champaign, USA

Ulrike Hahn
Department of Psychology, Birkbeck University, London, UK

Kristian Kersting
Technical University of Darmstadt; hessian.AI; German Research Center for AI, Germany

Tobias Gerstenberg
Department of Psychology, Stanford University, USA

Abstract

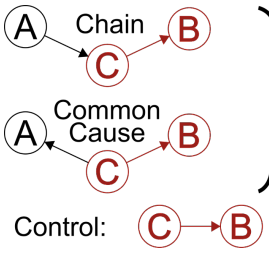
Causal reasoning is important for humans and artificial intelligence (AI). Causal Bayesian Networks (CBNs) model causal relationships using directed links between nodes in a network. Deviations from their edicts result in biased judgments. This study explores one such bias by examining two structures in CBNs: canonical Chain (A→C→B) and Common Cause (A←C→B) networks. In these structures, if C is known, the probability of the outcome (B) is normatively independent of the initial cause (A). But humans often ignore this independence. We tested mutually exclusive predictions of three theories that could account for this bias (*N*=300). Our results show that humans perceive causes in Chain structures as significantly stronger, supporting only one of the hypotheses. The increased perceived causal power might reflect a view of intermediate causes as more reflective of reliable mechanisms. The bias may stem either from our interventions in the world or how we talk about causality with others. LLMs are primarily trained on language data. Therefore, examining whether they exhibit similar biases can determine the extent to which language is the vehicle of such causal biases, with implications for whether LLMs can abstract causal principles. We therefore, subjected three LLMs, GPT3.5-Turbo, GPT4, and Luminous Supreme Control, to the same queries as our human subjects, adjusting a key 'temperature' hyperparameter. We show that at greater randomness levels, LLMs exhibit a similar bias, suggesting it is supported by language use. The absence of item effects suggests a degree of causal principle abstraction in LLMs.

Keywords: Causal Cognition; Mechanistic Reasoning; Large Language Models; Causal Chain; Bias in Causal Judgment, Common Cause; Bayesian networks; Causal argumentation.

Introduction

Representations of causal structure guide our reasoning and shape our views of reality. For instance, keeping variables and probabilities constant, people provide different judgments of a causal Chain, a sequence of causally related events that result in an outcome, than a Common Cause structure where an underlying factor gives rise to multiple effects (Rehder, 2014). Would a mere change in structure lead to systematic differences in how the *causal strength* of a cause is perceived?

Causal Bayesian Networks (CBNs) provide a common approach to causal structure representation, which has been fruitfully applied to similar questions (Pearl, 2009). CBNs are graphs that depict probabilistic interdependencies between variables. The variables (called "nodes") are connected through directed arrows (called "edges") into acyclic structures, indicating their probabilistic associations. Bayes Theorem provides a prescriptive framework for rationally and consistently updating beliefs based on evidence according to such networks, a deviation from any of its axioms leading to demonstrably suboptimal reasoning.



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

$$\begin{aligned} &P(\mathbf{A}, \mathbf{B}, \mathbf{C}) \\ &\quad = P(\mathbf{B}|\mathbf{C})P(\mathbf{C}|\mathbf{A})P(\mathbf{A}) \\ &\quad = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})P(\mathbf{C}) \end{aligned}$$

Figure 1: The joint probability P (A, B, C) in canonical Common Cause and Chain Causal Bayesian Networks is identical. Given C, the causal strength of C → B is independent of the network structure.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Two well-studied CBNs that we will focus on are shown in Figure 1: three-node Chains and Common Cause networks. The joint probability, the probability that the events represented by all three nodes happen in an instance of the causal structure, is equal for the two graphs. This "equivalence class" means that a given dataset would have the same likelihood under both structures, indicating that they cannot be differentiated solely based on observational data. Therefore, any systematic difference in our intuitions of causal strength across them is not due to the networks' overall likelihood. But more central to our study is the conditional independence that leads to this equivalence: the probability of B should not depend on A if we know C. For a given value of C, the likelihood of C→B, hence the causal strength of C for bringing B about, should be the same for the Chain and Common Cause networks in Figure 1.

Humans systematically violate such independence assumptions in causal judgments (cf., Rehder, 2014; Park & Sloman, 2013; Rehder & Burnett, 2005; Mayrhofer et al., 2008). Recently, the direct scope of a cause, i.e., the number of distinct effects generated directly by it, has been studied as a source of perceived causal strength (or lack thereof) not predicted by Bayesian theory (Sussman & Oppenheimer, 2020; Stephan et al., 2023). In the Chain A→C→B, the direct scope of C is one, less than the node's scope in the Common Cause structure A←C→B, which is two. Sussman and Oppenheimer (2020) argued that, depending on the valence of a target effect (B in this case), a broader scope might enhance or diminish perceived causal strength. For positive effects like a drug preventing symptoms ("boons"), a narrower scope increases perceived causal strength: a drug that only works on headaches is expected to have a stronger effect than one that reduces dizziness too. However, for negative effects like causing symptoms ("banes"), a broader scope leads to higher perceived strength: a drug that causes both headaches and dizziness is perceived as stronger than one that causes only headaches. Like most prior research, Sussman and Oppenheimer (2020) limited their comparison to a single structure type: Common Cause networks, with two-variable direct causation as the baseline.

Stephan et al. (2023) found that when scenarios were abstract enough to eliminate prior beliefs in participants (e.g., an alien on Mars eating a red crystal that induces/prevents three vs. one attribute(s)), the effect of scope was the same for boons and banes. They found a "dilution effect" instead: in causes with broader scopes (three effects rather than one), a singular "source" of causal strength was seen as distributed and, therefore, "diluted" across the multiple effects, regardless of effect valence.

Park and Sloman (2013) uncovered another systematicity in people's responses that may be relevant to perceived causal strength in these cases: Subjects' deviations from experimenter-provided causal structures and their subsequent apparent violations of normative reasoning always revolved around *mechanisms* and could be changed with mechanism-directed instructions. For instance, based on whether the same mechanism accounted for the two effects in a Common Cause network or the mechanism for each effect was distinct, perceptions of the causes' influence differed significantly. This finding adds to a large literature in the cognitive sciences demonstrating the centrality of mechanisms to causal reasoning in ways that do not always track normative predictions (Johnson & Ahn, 2017).

For instance, Zemla and colleagues (2017) found that mechanistic information subverts a preference for simplicity in explanations, with subjects going so far as to identify enough causes to make the effect seem inevitable. From a more normative standpoint, Russo and Williamson (2007) have emphasized the importance of mechanistic evidence in discerning alternative non-causal explanations, such as confounding, bias, or chance, which may lead to misleading associations. If a plausible mechanism to explain the correlation is absent, they argue the association is likely coincidental (Russo & Williamson, 2007).

Menzies (2012) argues that mechanistic relevance can be framed as a link in the causal chain connecting the input (the initial cause) and the output (the final effect). Adopting this view, C in the Chain A→C→B may be seen as the *mechanism* that explains A's impact on B. If so, the perceived causal strength of C on B in A→C→B would be *higher* than in a control condition C→B, as the chain would be seen as more likely to represent a *mechanistic* rather than a merely *correlational* connection. In the Common Cause network, on the other hand, the mechanism is vague without elaboration. Therefore, if people accept the given networks as the ground truth (Rehder, 2014) but base their strength judgments on other structural features like the presence of a mediating process, we would expect the Chain condition to have higher likelihood ratings. Figure 2 shows how three experimental conditions would allow this hypothesis to be evaluated alongside the accounts of Sussman and Oppenheimer (2020) and Stephan et al. (2023).

A key question in cognitive science concerns the role of language in causal reasoning, with important implications for the causal capabilities of large language models (LLMs; Binz & Schulz, 2023; Willig et al., 2022). Human beings can communicate causal information via language, but they also develop an understanding of causality through interactions with the world.

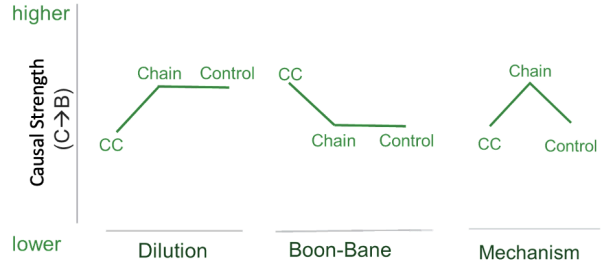


Figure 2: Pattern predictions for the effect of causal structure on perceived causal strength estimates across conditions. The predictions of the different theories for networks with negative-valence contents are illustrated.

Theoretically, LLMs can perform perfect normative reasoning by conducting exact computational operations internally. However, trained almost exclusively on human textual data, we expect LLMs to pick up on biases reflected in language use, not those only learned through experience. We collect and compare LLM answers with the distribution of human judgments to see if language is the primary medium for developing a bias in favor of intermediate Chain nodes (cf. Figure 5). Beyond clarifying the cognitive basis of the bias, this comparison would also contribute to an ongoing debate within the AI community about whether LLMs grasp causality or merely echo causal language without comprehension (Zečević et al., 2023). Many researchers (e.g., Kıcıman et al. 2023, Jin et al. 2023) have recently taken the stance that current LLMs cannot generalize causal ideas beyond their training distribution and/or without strong user-induced guidance. But if a preference for canonical Chain over Common Cause structures emerges across items in experiments with LLMs, that would provide some evidence that LLMs suffer from the influence of human bias for causal principle abstraction.

Human Experiment

Methods

Design We manipulated causal structures between subjects to minimize task demands. We limited our materials to structures with three nodes to minimize working memory demands, as the general explanations we evaluated all demanded degrees of at most two. Stephan et al. (2023) highlighted intuitive familiarity with the variables as a potential explanation for results that differed from Sussman & Oppenheimer (2020). To account for this possibility, we included an adapted version of Stephan et al.'s (2023) Alien scenario that precludes prior knowledge of the causal relations. Our two other items represented more intuitive domains: a widely used causal setup about the Economy (adapted from Rehder, 2014) and a novel scenario about Sex Work Criminalization that represented more prescriptive causal reasoning. The C→B relation was always presented as probabilistic (C can lead to B) to prevent ceiling effects seen in a pilot study. Because Sussman and Oppenheimer's (2020) account deviates from the Dilution predictions only when target effects are negatively valanced, node B in Figure 1 was negatively valanced in all scenarios to provide better experimental contrast. Following Rehder (2014), we instructed participants to consider all presented relations as "single sense," meaning that only the presence of the event represented by a node has an impact on its effect(s). see OSF for the full text of the instructions and the scenarios.

Network Structure For the main manipulation, each participant was assigned to one of the following conditions at random: 1. Chain, where A generates C, which generates B (A→C→B); 2. Common Cause network (CC) where C generates A and B separately (A←C→B), 3. Control, a baseline for (1) and (2) in which A is not included, and C generates B (C→B). For instance, participants in the Chain condition were presented with scenarios such as: "High-interest rates lead to more loan defaults, which leads to more inflation." In contrast, those in the Control condition encountered scenarios like: "More loan defaults lead to more inflation." To examine whether generative causation behaves differently than preventative causal power in these scenarios (Walsh & Sloman, 2011), we compared two additional conditions: 4. CC(P): a Common Cause network where C prevents A and separately inhibits B (A←C→B), 5. Control(P): a baseline for (4) in which A is not included, and C inhibits B (C→B). A random pairing of scenarios was created for within-subject manipulations and then counterbalanced across participants. Scenarios were individually presented on the screen in randomized order. Demographics followed the last vignette.

Dependent Measure The dependent measure was the likelihood of B if C had been present. In the example scenario above, participants were asked to evaluate the likelihood of the statement: "How likely is it that more loan defaults lead to more inflation?". The C→B causal relationship was scrutinized as it remained constant across the network types under examination.

Participants We collected pilot data from 100 participants to calculate the needed sample size using the simulation-based method described by DeBruine & Barr (2021). A target sample size was predetermined using a Monte Carlo simulation via the 'SIMR' package in R (Green & MacLeod, 2016). We estimated the input parameters of our simulation to determine the sample size to have 90% power to detect the main effect of causal structure. Our final estimated sample size was 300.

Three hundred and twenty-nine subjects were recruited through Prolific Academic (https://www.prolific.co/) and compensated at the average rate of $10/hr. Over the simulation, the ~10% increase in sample size was meant to offset anticipated exclusions due to inattentive subjects. We used a data-driven Mahalanobis Distance (Leys et al., 2018) to identify non-human participants and inconsistent or inattentive responses. This step resulted in excluding 6 participants. We replicated the main results, including these subjects. Since our secondary attention check excluded a third of the participants, we limited our exclusion to the data-driven approach explained above to preserve power. However, the key results were replicated after excluding subjects who failed the secondary check (see OSF for details). The final sample of US and UK residents (120 males, 198 females, 5 non-binary) had an average age of 37.29 years (SD = 13.02, range: 18 to 76).

Review of Hypotheses

Figure 2 summarizes the target contrasts examined across theories and CBN types. The OSF preregistration includes all predictions except for the Mechanism explanation. Dilution (Stephan et al., 2023) predicts that causal strength would be reduced if a cause has two direct

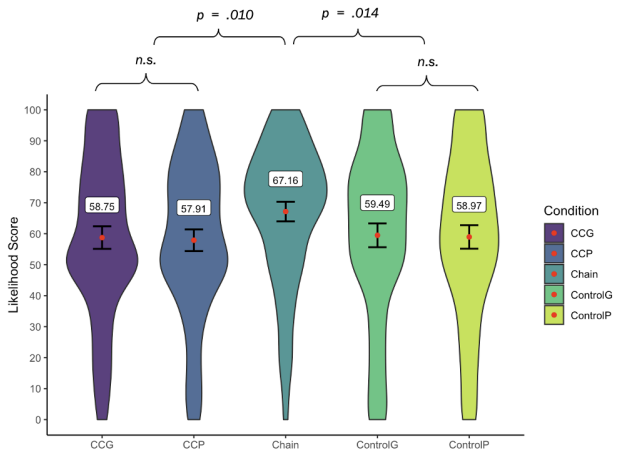


Figure 3. Distribution of responses over conditions. The likelihood in the Chain Condition (middle row) is significantly higher (i.e., higher likelihood score) than in the rest of the conditions. However, there is no difference between Control (CP; CG) and Common Cause (CCP; CCG) networks. Error bars show 95% confidence intervals.

effects rather than one, regardless of other structural features and the effect's valence. Therefore, less causal strength is expected in CC and CC(P) conditions where C has two effects, compared with Chain and Control conditions, which contribute to only one effect. Given that target effects all have negative valence, the Bane-Boon Theory (Sussman & Oppenheimer, 2020) predicts the opposite pattern: CC and CC(P) should receive higher ratings due to their wider direct scopes than the Chain and Control conditions. The Mechanism hypothesis expects higher causal strength in the Chain condition where subjects may regard C as a mechanism (Menzies, 2012), preferred over covariation alone (Ahn & Bailenson, 1996). As no intermediate cause exists in Control and Common Cause conditions, this notion predicts no difference between them.

Result

Given our design's hierarchical structure, we used generalized mixed-effects models. Since the dependent variable was on a 100-point scale, we employed linear mixed-effect models through the 'LME4' package in R (Bates et al., 2015). Structure (Chain vs. Common Cause vs. Control) was a fixed effect, while participants and scenarios served as random effects along with the 'maximal' slopes advocated by prior research (Barr et al., 2013). Figure 3 shows the experimental results. Structure's main effect was significant ($b = 8.78$, $SE = 3.29$, $z = 2.67$, $p = .01$, two-tailed test). However, the direction was in contrast to the predictions of Boon-Bane and Dilution accounts, aligning instead with a view of C as a mechanistic cause. We compared Chains to Control and Common Cause networks in separate models to better interpret the results. Random parameters were included as before.

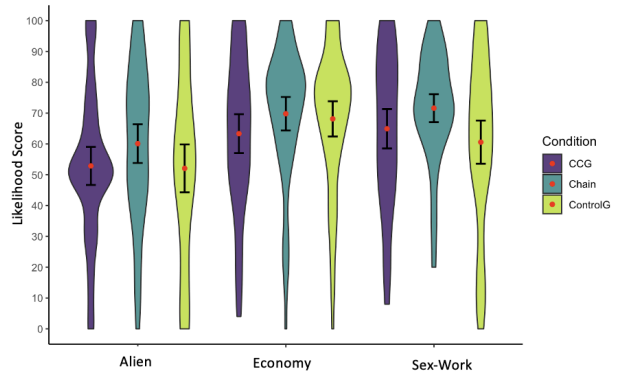


Figure 4. Distribution of responses across conditions for the various items. The likelihood in Chains (in Green) is higher than the other generative conditions across all items. Error bars show 95% confidence intervals.2023)

We calculated contrasts over estimated marginal means using 'emmeans' (Lenth et al., 2018). Pairwise contrasts with the Tukey adjustment for multiple comparisons showed higher ratings for causes in Chains than Control ($b = 7.48$, $SE = 3.01$, $z = 2.48$, $p = .014$, two-tailed test: $BF_{10} = 0.86$) and Common Cause conditions ($b = 8.74$, $SE = 3.34$, $z = 2.61$, $p = .010$; two-tailed test; $BF_{10} = 0.71$). No difference was observed between Control and Common Cause conditions in Generative ($b = 2.26$, $SE = 3.33$, $z = .67$, $p = .91$, two-tailed test) or Preventive ($b = 1.48$, $SE = 3.28$, $z = .47$, $p = .96$, two-tailed test) conditions (See Figure 3). Similar patterns across items suggest that prior knowledge had little influence (Figure 4).

As the lack of Dilution is based on a null effect, a Bayesian mixed-effect analysis was performed to determine its reliability using 'brms' (Bürkner, 2017). The Bayesian Mixed Effect model provided evidence for the null effect in both Generative ($BF_{10} = 0.96$, $CI_{95} = [-1.5, 2.21]$) and Preventative conditions ($BF_{10} = 1.5$, $CI_{95} = [-1.93, 1.74]$). To further ascertain that the effect size we observed was close to zero, an equivalence test was performed using the 'TOSTER' package (Lakens, 2017). The equivalence test confirmed that the distributions of likelihood scores in Common Cause vs. Control are equivalent ($z = 8.9$, $p < 0.001$, two-tailed test).

Discussion

We compared three explanations for changes in perceptions of causal strength based on network structure that are not predicted by normative theory. We limited our comparison to 3-node Chain and Common Cause structures as canonical network formats for which the hypotheses offered mutually exclusive predictions. We found no evidence of causal strength dilution for nodes with more direct effects (Stephan et al., 2023; Sussman & Oppenheimer, 2020) or an increase in perceived strength given negatively valanced material (Sussman & Oppenheimer, 2020).

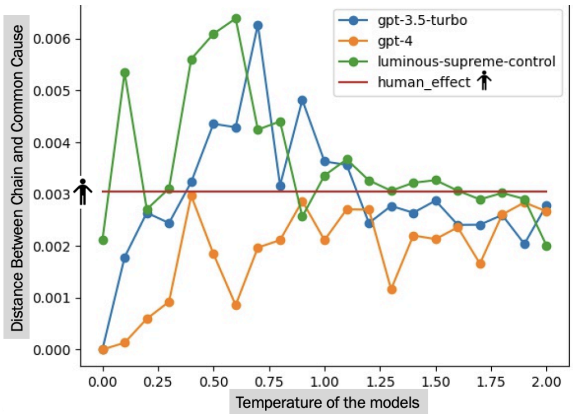


Figure 5. The average effect of switching from CC to CHAIN condition. The effect is measured as the Wasserstein distance between the Chain and Common Cause distributions for various 'temperature' values.

However, the intermediate nodes in canonical chains received significantly higher likelihood ratings than common cause networks. Of the three possibilities examined, this finding is consistent with the Mechanism hypothesis (see Figures 2-3). This hypothesis suggests that an intermediate cause in a canonical chain is a reliable mechanism (Menzies, 2012), preferred over covariational causes (Ahn et al., 1995; Park & Sloman, 2013). In our networks, the influence of C over the final effect is more direct than the initial cause but if seen as a mechanistic explanation, it may be deemed more generalizable to other instances (Johnson & Ahn, 2017). Intermediate causes were missing in the Common Cause networks, precluding such conjectures.

Large Language Models

To examine whether this bias is learned primarily through language rather than interventions on the world, we compared our behavioral data with the responses generated by three recent LLMs: GPT3.5-Turbo (OpenAI, 2022), GPT4 (OpenAI), and Luminous Supreme Control (Aleph Alpha, 2023). Instead of using default hyperparameter values, we systematically adjusted the temperature settings of these models to explore whether deterministic (zero temperature) or increasingly non-deterministic behavior more closely matches human reasoning in this task.

Methods

We queried the models for all conditions and items with temperatures ranging from 0.0 (deterministic) to 2.0 (highly non-deterministic; maximum value permitted by the APIs) in 0.1 steps. The temperature for the Luminous model was scaled by 0.5 to match the GPT models. Luminous did not show fully deterministic behavior at zero temperature (cf. Fig. 5; Wasserstein distance is non-zero). For every condition-temperature-model combination, we sampled 100 answers for the same query.

The queries followed the scenarios used with humans, except for using a template that encouraged the models to give a number between 0 and 100 as a response. The prompt template was as follows: "This is a hypothetical question. *<insert human scenario here>*. Given the information above, in your opinion, on a scale of 1 to 100, how likely is it that *<insert the causal statement examined>*? Please respond only with one single number and no text".

Statistical analysis

We used hierarchical mixed-effect models to discern patterns in the responses. To quantify the similarity between human and AI judgments, we computed Earthmover's distance (EMD, also known as Wasserstein distance; Vaserstein, 1969; Kantorovich, 1960), defined between two distributions *u* and *v* as follows:

$$EMD = inf_{\pi \in \Gamma(u,v)} \int_{R \times R} |x - y| d\pi(x, y) \quad (1)$$

Where $\Gamma(u, v)$ is the set of all probability distributions with marginals *u* and *v*. A key property of EMD with categorical data is its invariance to overall shifts of the distribution, which allows us to focus on the similarity of answer frequencies between humans and models. Other commonly used f-divergent metrics (such as KL divergence; Kullback & Leibler, 1951) do not apply to our comparison, as they vary with the overall mean of the distribution and break down in our scenario of sparsely populated distributions. For instance, a single entry receives all the probability mass with zero temperature, rendering KL divergence meaningless. Nonetheless, to ensure comprehensiveness, we evaluated entropy and KL divergence and have included the results in this repository, where more details about the EMD measure, its implementation through Python's SciPy package, and other simulation information are also stored.

EMD estimates help us gauge the extent to which LLMs' reasoning aligns with human judgment and whether their judgments are as diverse or predictable as those of humans. Note, however, that there are qualitative differences between sampling from *N* different human participants and sampling *N* responses from the latent distribution of a single LLM with varying temperatures. Nonetheless, we find that, with certain temperature settings, human and LLM distributions tend to approach each other.

Results

The mixed effects model with the fixed factor of Condition (Chain, Common Cause, Control) and the random slope of Model (GPT4, GPT3.5, Luminous) highlighted a bias in LLMs similar to the one we found in humans: attributing greater causal strength to the intermediate cause in canonical Chains (μ= 67.59, SD=20.2) than to the corresponding nodes in Common Cause (μ=64.89, SD=19.8; b = 3.02, SE = .27, z = 10.89, p <.0001, two-tailed test) and Control conditions (μ=55.05, SD=17.89; b = 12.15, SE = .28, z = 42.6, p<.0001, two-tailed test).

Model behavior was particularly similar to humans when the temperature parameter surpassed 1.

Figure 5 shows the effect of switching from a Common Cause to a Chain network on dependent scores for each model using Wasserstein Distance averaged per scenario. For comparison, we indicate the average distance in the human data, the red line representing generally higher ratings for Chain than Common Cause structures. For GPT-3.5-Turbo and Luminous, low temperatures correspond to little preference for either condition, but the distance between the condition distributions increases afterward. Both models eventually converge towards a randomly sampled uniform distribution with distance values below the human reference. The most recent LLM, GPT-4, starts with zero distance, but the preference for chains increases with higher temperatures. Generally speaking, temperatures >1.0 match the human data best on average, while too high of a temperature value (>1.9) induces too much variance. With temperatures between 1.0 and 1.9, the observed preference for Chains is remarkably similar to that observed in humans across all three models.

Receiving human instructions and fine-tuning with human feedback may strengthen the alignment of LLM responses with human biases. Such methods commonly result in the models being more 'certain' about their answers (resulting in sharper modes in their answer distributions). We observe this effect in GPT-4, where only a few possible answers are considered in low-temperature settings. Raising the temperature diffuses the modes, lowering the models 'confidence', until the variance in the distribution corresponds to that observed in human participants.

General Discussion

We examined the effect of network structure on causal strength judgments in humans and Large Language Models (LLMs). Human participants and multiple LLMs - GPT3.5-Turbo (OpenAI, 2022), GPT4 (OpenAI, 2023), and Luminous Supreme Control (Aleph Alpha, 2023) - considered intermediate causes in chains to be more potent than causes in simple C→B networks, or those with multiple independent direct effects (i.e., Common Cause). This represents a violation of normative Bayesian reasoning. Varying LLM hyperparameters, we found the closest match for the human bias in variants with higher temperatures, i.e., those incorporating more randomness into the AI response selection process.

Given that all scenarios involved negative effects, the Bane-Boon Theory (Sussman & Oppenheimer, 2020) predicted that the Chain cause with the narrower direct scope would garner lower ratings than the Common Cause condition, which contradicts our finding. Our results also contradict Stephan et al.'s (2023) experiments, which predicted a "dilution" of causal strength in Common Cause conditions and similar ratings for the target cause in Chain and Control cases due to their identical direct scopes. While our scope manipulation differed slightly from both prior studies (comparing degrees of two rather than three with a single-effect baseline), we find this an unlikely explanation for the discordant results given the generality of the previous researchers' explanations.

One alternative explanation for the enhanced causal strength of 'in-between' causes in canonical Chains is their representation as mechanism nodes (Menzie, 2012), observed to have an outsize effect on causal intuitions compared with covariational causes (Ahn et al., 1995; Ahn & Bailenson, 1996; Johnson & Ahn, 2017; Zemla et al., 2017; Russo & Williamson, 2007). While the A→C link in the chain is also part of the mechanistic explanation according to Menzie (2012), the C→B link is the only consistent element across the network types in our study and, therefore, served as the focus of our evaluation. If Menzie's (2012) characterization is accurate, a similar boost in perceived strength should be seen in judgments of the initial cause in a canonical Chain.

A partially divergent explanation is that the middle node in a chain is considered supported by its own cause. This perceived support could arise from violating the normative conditional independence between causes A and B given C. In Chains, the sequence A→B may be perceived as "passing on" some of its causal strength to or suggesting greater regularity for the C→B link. In contrast, a A←C link would not support the independently produced C→B in a Common Cause network. This line of reasoning leads us to propose what we might call the "Causal Relay Hypothesis", which predicts no boost in the perceived strength of the initial cause in a chain (A→C) but enhanced strength for the downstream link from C to B. Future experiments can help differentiate between this and the mechanism account by including judgments of the initial cause. Direct evaluation of whether the intermediate node is seen as a reliable mechanism would also help the comparison. Comparing generative and preventive causal chains can further corroborate such findings.

Whatever the explanation for the biased judgment of intermediate causes in Chains, it is not merely learned from interventions in the world, as large language models trained on human-generated text exhibit similar biases.

The implications for domains where causal reasoning is essential, such as medicine and law, could be far-reaching. Suppose LLM-based decision-support systems in these areas inherit biases like the one observed in our study. In that case, there is a risk of perpetuating errors, especially at higher model temperatures where biases may be more pronounced and align more closely with human reasoning. As we increasingly rely on AI for complex decision-making, it becomes more important to study such biases and mitigate them if needed so that more reliable AI systems can aid human decision-makers without introducing undue risk. In causal cognition, our findings prompt further inquiry into how language shapes or reflects our conceptualization of causality. They suggest that linguistic data, rich with human experiences and inferential patterns, could play a significant role in studying causal biases.

Acknowledgment

AK and UH, part of the Mercator fellowship, received support from the Deutsche Forschungsgemeinschaft (DFG) under Project number 455912038/ "Robust Argumentation Machines" (RATIO). AK also acknowledges the support from Forward College, Berlin, and additional backing by the Bavarian Ministry for Economic Affairs, Regional Development, and Energy as part of a project aiding the thematic development of the Fraunhofer Institute for Cognitive Systems IKS.

BH received funding from the Beckman Institute at the University of Illinois Urbana-Champaign.

MW and KK acknowledge the support of the German Science Foundation (DFG) for the project "Causality, Argumentation, and Machine Learning" (CAML2, KE 1686/3-2) under the program SPP 1999 "Robust Argumentation Machines" (RATIO). They also benefited from the support of the Hessian Ministry of Higher Education, Research, Science, and the Arts (HMWK) for the project "The Third Wave of AI", and the EU Project Tango. The views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible (Grant Agreement no. 101120763 - TANGO)

TG was supported by a grant from the Stanford Institute for Human- Centered Artificial Intelligence (HAI).

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ... & Wagenmakers, E. J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. Advances in Methods and Practices in Psychological Science, 1(3), 357-366.

Ahn, W. K., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. Cognitive psychology, 31(1), 82-123..

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54(3), 299-352.

Aleph Alpha. (2022). Luminous language model [luminous-supreme-control] Available at: https://docs.aleph-alpha.com/docs/introduction/model-card/

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(i01).

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of statistical software, 80, 1-28.

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. Advances in Methods and Practices in Psychological Science, 4(1), 2515245920965119.

Englemann, N., & Waldmann, M. R. (2022). How causal structure, causal strength, and foreseeability affect moral judgments. Cognition, 226, 105167.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493-498.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Zhiheng, L. Y. U., ... & Schölkopf, B. (2023). CLadder: A Benchmark to Assess Causal Reasoning Capabilities of Language Models. In Thirty-seventh Conference on Neural Information Processing Systems.

Johnson, J. T., & Drobny, J. (1985). Proximity biases in the attribution of civil liability. Journal of Personality and Social Psychology, 48(2), 283.

Johnson, S. G., & Ahn, W. K. (2017). Causal mechanisms. In M.R. Waldmann (Ed.), The Oxford handbook of causal reasoning, 127-146. Oxford University Press.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. Management science, 6(4), 366-422.

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86.

Lakens, D. (2017). Equivalence testing with TOSTER. APS Observer, 30.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). emmeans: Estimated marginal means, aka least-squares means (R package, Version 1.4)[Computer software].

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. Journal of experimental social psychology, 74, 150-156.

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 30, No. 30).

Menzies, P., 2012, “The Causal Structure of Mechanisms”, Studies in History and Philosophy of Biological and Biomedical Sciences, 43: 796–805.

OpenAI. (2022). OpenAI GPT API [gpt-3.5-turbo]. Available at: https://platform.openai.com/docs/models/

OpenAI. (2023). OpenAI GPT API [gpt-4]. Available at: https://platform.openai.com/docs/models/

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. Cognitive psychology, 67(4), 186-216.

Pearl, J. (2009). Causality (2nd Edition). New York, NY: Cambridge University Press.

Ramdas, A., García Trillos, N., & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. Entropy, 19(2), 47.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. Cognitive psychology, 50(3), 264-314.

Rehder, B. (2014). Independence and dependence in human causal reasoning. Cognitive psychology, 72, 54-107.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. International Studies in the Philosophy of Science, 21(2), 157-170. DOI: 10.1080/02698590701498084

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. Annual review of psychology, 66, 223-247.

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In: Galavotti, C.M. et al. (Eds.) Stochastic causality. Stanford: CSLI Publ., pp. 157-172.

Stephan, S., Engelmann, N., & Waldmann, M. R. (2023). The perceived dilution of causal strength. Cognitive Psychology, 140, 101540.

Sussman, A. B., & Oppenheimer, D. M. (2020). The effect of effects on effectiveness: A boon-bane asymmetry. Cognition, 199, 104240.

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. Problemy Peredachi Informatsii, 5(3), 64-72.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. Mind & Language, 26(1), 21–52. https://doi.org/10.1111/j.1468-0017.2010.01409.x

Willig, M., Zečević, M., Dhami, D. S., & Kersting, K. (2022). Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. Psychonomic bulletin & review, 24, 1488-1500.