UNIVERSITY OF CALIFORNIA

Los Angeles

Causal Inference and Large Language Models from the Causal Invariance Framework

A thesis submitted in partial satisfaction of the

requirements for the degree of Master of Science

in Statistics

by

Emily Frances Wong

2023

ABSTRACT OF THE THESIS

Causal Inference and Large Language Models from the Causal Invariance Framework

by

Emily Frances Wong

Master of Science in Statistics

University of California, Los Angeles, 2023

Professor Hongjing Lu, Chair

Statistics serves as the grammar of all science, and central to the goal of science is understanding cause-effect relationships. Scientists rely on research methodology and statistical tools to uncover causal relationships, and engineers rely on statistical methods to create artificial assistants to aid daily life. Neither statistical learning nor next-word-prediction (used to train artificial general intelligence) are consistent with rational causal learning and reasoning in humans. The present thesis examines the fundamental goals and assumptions made in dominant statistical methods and discusses their implications for statistical inference and commonsense reasoning in artificial general intelligence (AGI). The first section introduces and evaluates a causal alternative to logistic regression, which estimates the causal power (from the causal invariance framework) of treatments among covariates. Causal invariance is defined as the influence of a candidate cause (elemental or conjunctive) that is independent of background causes, with the aspiration of acquiring knowledge that's useable, in the minimalist sense being able to generalize from a *learning context* to an *application context*. The second and final section investigates current benchmark tasks used to evaluate causal reasoning in large language models (e.g., GPT-3, GPT-4), and introduces a stricter test informed by psychological literature on human causal cognition under the causal invariance framework.

The thesis of Emily Frances Wong is approved.

Chad J. Hazlett

Qing Zhou

Patricia Cheng

Hongjing Lu, Committee Chair

University of California, Los Angeles

2023

DEDICATION

It takes a village to raise a child. To Chauncey Jones, for your unwavering belief in me, but most importantly, your kindness. December 8, 1958 to May 13, 2023.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

1. The causal invariance framework was developed by Dr. Patricia Cheng.
2. The CISI model was developed with immense guidance from Dr. Han Du.

Causal Inference and Large Language Models from the Causal Invariance Framework

Statistics serves as the grammar of all science, and central to the goal of science is understanding cause-effect relationships. Most recently, the scientific community was called to action to put an end to a deadly pandemic. After two grueling years, headlines all around the world read "FDA Approves First COVID-19 Vaccine", and public health officials universally recommended the uptake of one of three vaccine schedules. Fundamental to these decisions and recommendations were what scientists learned about the causal relationship between vaccine uptake and COVID-19 infection. Through well-controlled clinical trials, researchers learned that the Moderna vaccine was 95% effective at preventing infection. As it turns out, vaccine efficacy as defined by the Center for Disease Control (CDC, 2012), is a causal probability (Sheps, 1958), equivalent to Cheng's (1997) [preventative] causal power: it is the probability that the vaccine alone prevents infection, regardless (i.e., independent) of background causes, known and unknown. Importantly, this probability implies that if healthy people were to get sick (counterfactually), the vaccine would have prevented illness by 95%. This consideration of counterfactuals is not evident in standard statistical models such as logistic regression.

This recent example is a strong reminder of the critical role causality should play in statistics; it also demonstrates how people naturally reason about causality. However, many current statistical models (e.g., logistic regression) are not causal in nature. As Pearl (2009) has argued, two fundamental questions of causality remain without satisfactory answers: (1) What empirical evidence is required for legitimate inference of cause-effect relationships? (2) Given that we are willing to accept causal information about a phenomenon, what inferences can we draw from such information and how? He suggests that these questions have been without satisfactory answers in part because there exists no semantics or mathematical tools for casting causal questions or

deriving causal answers. This lack of causal representation in basic statistics continues to impact related fields such as artificial intelligence, where models such as deep neural networks rely on noncausal statistical methods (e.g., regression).

The issue of causality has been studied across several fields, including philosophy and cognitive science. From the cognitive science standpoint, all empirical knowledge constitutes human representations of reality (e.g., Hawking & Mlodinow, 2010; Hoffman, 2019; Hume, 1739/1987; Kant, 1781/1965), only made available by basic cognitive constraints such as preference for coherence and thus parsimony (Marr, 1982; Pizlo, 2001). From this perspective, traditional statistical semantics lacks any representation of causality. Given major advances in graphical models such as directed acyclic graphs (DAGs; Pearl, 1995), causal Bayes nets (Pearl, 2001, 2009), and path models (Holland, 1988), some language exists for describing and testing causal relationships. (However, it is important to note that graph surgery approaches that utilize the *do*(*X*) operator must assume no self-selection bias and other confounding, but do not correct for violations of that assumption. To be clear, there is no replacement for true randomized control trials, and the present thesis will not offer one.) Additionally, no existing model directly estimates the *causal probability* of a candidate cause of interest among covariates. It is important to note the differences between causal models from Pearl's (e.g., causal DAGs and Bayes nets) and Cheng's perspective (e.g., causal invariance), and the consequences of these two different approaches. Critically, the two approaches differ in what they are willing to assume about how causal influences contribute to an outcome (e.g., under what conditions two causal factors influence the outcome independently). The next section introduces the issue of causality from the perspective of Cheng's causal invariance framework.

Longstanding Questions About Causality from the Philosophy of Science

A major goal in our daily lives, as lay reasoners and as scientists, is to understand and explain the way in which the world works. Our ability to do so has contributed to the survival of our species as evidenced through our ability to create tools for basic survival (e.g., striking stones in certain ways can create fire), advance medicine, and find solutions to mitigating climate change. Scientists rely on research methodology and statistical tools to uncover causal relationships, and engineers rely on statistical methods to create artificial assistants to aid daily life. The present thesis examines the fundamental goals and assumptions made in dominant statistical methods and discusses their implications for statistical inference and commonsense reasoning in artificial general intelligence (AGI).

To motivate adopting a causal paradigm where appropriate (e.g., causal inference and causal reasoning in AGI), it is useful to first revisit longstanding questions from the philosophy of science to understand what assumptions are central in the causal induction process. Every day, we perceive a variety of lower-order stimuli such as colors and objects, as well as higher-order causal and non-causal relationships. Where do these perceptions come from, and do they provide us direct access to the reality we live in? In 1868, Thomas Huxley wrote: "How it is that anything so remarkable as that of consciousness comes about as the result of irritating nervous tissue is just as unaccountable as the appearance of the genie when Aladdin rubbed his lamp." Since then, we have learned a lot about human neurobiology and how it *correlates with* conscious experiences. However, as argued by Hoffman (2015), the fundamental question remains: How do we go from neural stimulation to a conscious experience of a red apple? This question extends to our higher-order perceptions such as *causality*.

As early as the 18th century, philosopher David Hume raised two issues about causal inferences: 1) causal relations are beliefs formed in the mind based on inherently noncausal data (Hume 1739/1987), and 2) experience is useful only if the future resembles the past (Hume, 1748/1975). So, how do humans go from covariation (registered neurobiologically) to causality, especially given infinitely many possible causal representations? A great deal of psychological evidence indicates that both humans and other mammals are sensitive to the distinction between covariation and causation (e.g., Blaisdell, Sawa, Leising, & Waldmann, 2006; Waldmmann, Hagmayer, & Blaisdell, 2006). This intuition is routinely taught in introduction to statistics courses, where students learn that correlation does not equal causation. This intuitive distinction suggests that our perception of causality is the product of inference and is constrained by *a priori* assumptions. In their review, Ichien and Cheng (2022) offer an answer to a question raised by Hume's two issues: How is it possible to tease apart the influence of a specific candidate cause from those of background causes, in a way that yields causal knowledge that generalizes across learning and application contexts? Cheng's (1997) concept of causal invariance posits that reasoners assume by default that the influences of candidate causes are independent from the influences of background causes. More recent literature (e.g., Bye, Chuang, & Cheng, 2023; Cheng & Lu, 2017; Park et al., 2022) further argues that reasoners *must* assume invariance of *whole causes* (i.e., the candidate causes, whether elemental or conjunctive; Park et al., 2022) across background causes (i.e., independent causal influences between whole causes and background causes) during learning to ensure logical consistency, if the causal relation is indeed the same across the learning and application contexts. To apply a causal relation, reasoners must assume the causal relation holds in the new context (i.e., is invariant). Because which context is where the reasoner "learns" and which is where they "apply" a causal relation is incidental, if the causal

relation is indeed invariant, failure to assume causal invariance during learning would result in a logical contradiction: the inferred causal relation would be *both* the same and *not* the same in the two contexts. In other words, reasoners assume by default that their learned knowledge is useable and generalizable (i.e., the past resembles the future yielding causal invariance). However, if subsequent empirical evidence undermines this assumption (observations that are inconsistent with causal invariance), the deviation from invariance serves as a signal to revise causal models towards greater generalizability and hence useability (Bye et al., 2023; Cheng, 1997; Cheng & Lu, 2017; Cheng, Sandhofer, Liljehom, 2022; Liljeholm & Cheng, 2007; Rescorla & Wagner, 1972; Woodward, 2000, 2006).

Overview of Thesis

This thesis considers the role of causality in statistical inference procedures for discrete outcomes and in relation to the design of large language models. The first section introduces and evaluates a causal alternative to logistic regression, which estimates the causal power of treatments. This new model will allow researchers to directly estimate the causal power of a candidate cause among covariates. The second and final section investigates current benchmark tasks used to evaluate causal reasoning in large language models (e.g., GPT-3, GPT-4), and introduces a stricter test that requires the models to estimate the causal power of a treatment.

Section 1: Statistical Inference

Statistical inference is a critical tool that allows researchers to make important societal recommendations. However, statistical models are not inherently causal. Logistic regression is a standard statistical model for analyzing data involving a binary outcome and a dominant method taught in undergraduate statistics courses. In logistic regression, a binary outcome variable, $\boldsymbol{Y}$, can be predicted by a set of explanatory variables $\boldsymbol{X} = \{x_1, x_2, \ldots, x_p\}$ with corresponding weights $\boldsymbol{\beta} =$

$\{\beta_1, \beta_2, \ldots, \beta_p\}$. As in all regression, coefficients are a measure of covariation between the explanatory and outcome variables, regardless of method of estimation (e.g., least squares or maximum likelihood). To test cause-and-effect relationships, researchers must address the issue of confounding in the data collection process through careful experimental design. However, whether data were collected through experimentation or not, estimated regression weights (expected change in outcome for every one-unit change in the predictor variable) are not always consistent with causal probabilities and may lead to divergent recommendations.

Causal Probabilities

There are various ways of defining a causal effect using probabilistic language (Pearl, 2022). One common approach has been to define a causal effect counterfactually (i.e., imagining what would have occurred given another realization of events; see Tian & Pearl, 2000, for mathematical definition). Tian and Pearl (2000) laid out three probabilities of causation assuming exogeneity (no confounding) and monotonicity (no prevention): 1) probability of necessity (PN), 2) probability of sufficiency (PS), and 3) probability of necessity and sufficiency (PNS).

Definition 1 (PN). Let *X* and *Y* denote two binary variables in some causal model *M,* and *x and* y stand for the propositions *X* = true and *Y* = true, respectively, and *x'* and *y'* for their complements. Probability of necessity (PN) is defined as the probability that some event *y* would *not* have occurred in the absence of event *x* (i.e., *x'*), given that *x* and *y* did in fact occur.

$$PN \triangleq P\left(y'_{x'} \middle| x, y\right)$$

Such a quantity has relevance in epidemiology, artificial intelligence (AI), and legal reasoning. In court hearings, for example, jurors may be required to judge the probability that death would not have occurred in the absence of the defendant's actions, given that death and harm has occurred (i.e., death would have not occurred *but for* defendant's actions).

Definition 2 (PS).

$$PS \triangleq P(y_x|y', x')$$

PS measures the capacity of *x* to produce *y*. Since "production" implies a transition from absence to presence, this definition conditions on situations where *x* and *y* are both absent. PS is often considered in policy, AI, and psychology. In AI, PS plays a major role in the generation of explanations (Tian & Pearl, 2000; Pearl, 2000, 2009). For example, if the goal is to assign moral responsibility, the reasoner must evaluate whether a cause was sufficient to bring out an outcome. Probability of sufficiency is mathematically equivalent to *generative causal power* from Cheng's (1997) causal invariance framework. However, Cheng further defines causal power for *preventative* causation (Equation 3) in addition to generative (Equation 4) by assuming independence in lieu of monotonicity (Cheng, 1997; Pearl, 2022)[1]. By making the independence assumption, causal power also does not require exogenous (i.e., background) variables to be held constant as is required in Pearl's (2022) approach; instead, causal power is estimated independently from background causes and is thus (potentially) generalizable to other contexts with differing backgrounds. In summary, causal power assumes that candidate cause *C* and all other background causes (*B*) influence the effect *E* independently and identically across individuals.

Let "→" indicate "cause" or "generate, "⊣" indicate "prevent" or "cure", *E* indicate effect, *C* indicate candidate cause, and *B* indicate background causes. Furthermore, let $G_B = P(B) * P(B \rightarrow E)$ indicate the probability of background generating the effect. $G_C = P(C) * P(C \rightarrow E)$ indicates the generative, and $P_C = P(C) * P(C \dashv E)$ indicates the preventative causal power of

[1] Cheng's (1997) causal power is identifiable under assumptions of independence and exogeneity (no confounding). Pearl's (2022) derivation of PS (generative causal power) is identifiable under assumptions of monotonicity (no prevention) and exogeneity.

C; $P(C) = 1$ when in the treatment group (cause present) and $P(C) = 0$ when in the control group (cause absent). Assuming *independent causal influences* and *no confounding*, the observed outcome (RHS of Equation 1) is decomposed into influences of its constituent causes, namely, candidate cause *C* and background cause *B* as follows:

$$P(G_B \cup G_C) = G_B + G_C - G_B \cdot G_C \tag{1}$$

Rearranging terms, G_C is:

$$G_C = \frac{P(G_B \cup G_C) - G_B}{1 - G_B} \tag{2}$$

Empirically, $\hat{P}(G_B \cup G_C) := P(E = 1|B = 1 \cap C = 1)$, and $\widehat{G_B} := P(E = 1|C = 0)$. Given that the background causes are always present in the context of the cause, it is implied that $P(E = 1|B = 1 \cap C = 1) = P(E = 1|C = 1)$. Therefore, the generative causal power of *C* can be empirically written as:

$$\widehat{G_C} = \frac{P(E = 1|C = 1) - P(E = 1|C = 0)}{1 - P(E = 1|C = 0)} \tag{3}$$

Likewise, the preventative power of *C* is:

$$\widehat{P_C} = \frac{P(E = 1|C = 0) - P(E = 1|C = 1)}{P(E = 1|C = 0)} \tag{4}$$

The only underlying assumption differing between these two equations is the potential direction of the causal influence of *C* on *E*: potentially generative in the former and potentially preventative in the latter. In the cognitive science and artificial intelligence literatures, the above two equations have been respectively termed the *noisy-OR* and *noisy-AND-NOT* function for the relevant variables. These two functions, and compositions of them, are collectively termed "*noisy-logical*" decomposition functions (Yuille & Lu, 2008). In short, causal power is the *proportional change* from no-cause to cause.

Definition 3 (PNS).

$$PNS \triangleq P(y_x, y'_{x'})$$

PNS is the probability that *y* would respond to *x* both ways, measuring both the necessity and sufficiency of *x* to produce *y*. Under exogeneity, PN and PS can be represented as a function of PNS as follows:

$$PN = \frac{PNS}{P(y|x)'}$$

$$PS = \frac{PNS}{1 - P(y|x')}$$

Note the equivalence between generative causal power and PS. The remainder of the section will focus on a special case of PS, causal power, which is the capacity of *x* to produce or prevent *y*.

Independent Causal Influences

Independence of causal influences means that the influence of one cause on the outcome does not depend on the influence of another cause on the outcome. The way in which this is mathematically expressed—referred to as *decomposition functions* in Cheng (1997)—depends on whether the outcome variable is continuous or discrete.

Continuous outcomes. When the outcome variable is continuous, the contribution of each independent cause is arithmetically summed together (i.e., additive). Consider two independent light sources that illuminate a common spot on a theater stage. One source shines 1500 lumens and the other shines 1400 lumens. Given that their contributions to the amount of light (continuous) falling on that spot are independent, then when both lights are on, there is a total of 2900 lumens on that spot. Under Newtonian assumptions, this notion of independent contributions can be applied to vector addition. Consider a ship headed due North at 12 miles per hour (mph), and a current flowing S 45° W at 4 mph. Given that these two vectors are independent, the actual bearing

and speed of the ship (resulting vector) can be computed by adding those two vectors together. In other words, independent contributions to a continuous outcome are *additive*. Going forward, this method will be referred to as an *additive decomposition function*.

Discrete outcomes. When two events, A and B, are independent, then the probability of either event occurring can be calculated according to the union rule assuming independence:

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B) \quad (5)$$

For causal queries, A can be replaced with "smoking causing lung cancer", and B with "asbestos exposure causing lung cancer" as shown in Equation 1. Given that we must subtract out the intersection, to not count it twice, independent contributions to a dichotomous outcome are *not simply additive*. Going forward, this method will be referred to as a *noisy-logical decomposition function*, to be consistent with Yuille and Lu (2008).

Comparing Logistic Regression and Causal Power

In regression, *independence* also means that the effect of one variable on the outcome does not depend on other variables. This can be expressed in a model without interaction terms between predictor variables, which are allowed to follow any distribution. Consider a study where Y = COVID infection, X_1 = vaccine trial group (treatment or control), and X_2 = age. If we believe that the effect of vaccine trial group on COVID infection is independent from the effect of age, one can specify the model as such:

$$\hat{Y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 \quad (6)$$

Instead of directly modeling the probability of infection, which is either 0 or 1 in observed data, logistic regression transforms the original dichotomous outcome into a continuous one through the logit transformation, yielding *log odds* (see Equation 7)[2].

[2] If we include an interaction term (e.g., x1*x2) into the model above, we are still taking the weighted sum of all predictor variables; in this case, it would be: $Y = b_0 = b_1*x_1 + b_2*x_2 + b_3*x_1*x_2$.

$$Z = \text{logit}\big(P(Y = 1)\big) = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) \tag{7}$$

Asymptotically, log odds follow a normal distribution with support from negative infinity to positive infinity. This transformation allows the use of an additive, linear function (right hand side of Equation 6) to model the initially dichotomous outcome. In the above model, $\hat{Y}$ is the predicted *log odds* of COVID infection, $\beta_1$is the estimated change in *log odds* of COVID infection between the control and treatment groups, and β_2 is the estimated change in *log odds* of COVID infection for every one-unit increase in age. The predicted probability of infection can be recovered by passing the weighted sum through the inverse logit function:

$$P(Y = 1) = \text{logit}^{-1}(Z) = \frac{1}{1 + \exp(-Z)} \tag{8}$$

Lastly, one can estimate the difference in probability between the vaccinated and unvaccinated groups by subtracting the average predicted probability of COVID infection among those who are vaccinated from those who are not; however, neither this difference (ΔP) nor $\beta_1$is a causal probability (i.e., it is not the probability that if a healthy person were to get sick, the vaccine would have prevented illness). Furthermore, estimating the causal strength of vaccine under logistic regression inherently assumes that the influences of the candidate (e.g., vaccine uptake) and background causes are mutually exclusive. As stated in Ichien and Cheng (2022):

> To put the point differently, the additive decomposition function implies that in events where the background causes exert their causal strength, the medication withholds exerting its causal strength, and in events where the medication exerts its causal strength, the background causes withhold exerting their casual strength. Such an absurd state of affairs would involve the medication and the background causes knowing in which patients each

other causes headaches and having the ability to control when they themselves do so. In other words, the medication and the background causes are not acting independently. In other words, if some unknown background cause has led to the outcome, other causes in the model cannot also have an effect.

Hence, logistic regression suffers from the issue of occlusion.

Issue of Occlusion. Occlusion refers to a phenomenon in which a candidate cause cannot influence an outcome when the outcome has already taken place due to background factors. Defining the effect of an intervention as ΔP tacitly assumes that the causal strength of the treatment and the causal strength of other, background causes are mutually exclusive. The following example is taken from Ichien and Cheng (2022). This example illustrates a situation in which a medication may cause headache as a side effect[3]. However, we know that headaches can also occur without medication. Therefore, individuals in the control group (no medication) may have headaches due to unobserved background causes. In this scenario, 12/36 individuals in the control group, and 30/36 individuals in the treatment group have headaches. According to logistic regression, the medication significantly increases the odds of headache, $OR = 10.00$, $p < .001$. Translating these results back into probabilities, the average probability of headache in the treatment and control groups are 83% and 33%, respectively. Thus, the *generative causal strength* of the medicine (candidate cause C), $\boldsymbol{G_{C_{logistic}}}$, as estimated by logistic regression, is simply the difference in probability: $\hat{G}_{C_{logistic}} = \Delta P = 50\%$ (i.e., 18/36). Because estimating the causal strength of vaccine under logistic regression inherently assumes that the influences of the candidate cause C (e.g., vaccine uptake) and background cause (B) are mutually exclusive, the outcome in the treatment group can be decomposed as follows:

[3] Corresponding R code can be found at: https://github.com/emilyfranceswong/Causal-Invariance-Demonstration.

$$P(G_B \cup G_C) = G_B + \boldsymbol{G_{C_{logistic}}}$$

$$= \frac{30}{36} = \frac{12}{36} + \mathbf{\frac{18}{36}}$$

In contrast, causal invariance would estimate the *causal strength as causal power*, $\boldsymbol{G_C}$, to be nearly 81% (computed in accord with Equation 3):

$$P(G_B \cup G_C) = G_B + G_C - P(G_B \cap G_C)$$

$$= \frac{30}{36} = \frac{12}{36} + \mathbf{\frac{27}{36}} - \frac{12}{36} \times \frac{27}{36} = \frac{12}{36} + \frac{27}{36} - \frac{9}{36}$$

Note that $P(G_B \cap G_C)$ does not indicate a conjunctive cause, where both *B* and *C* are necessary for the effect *E*. Instead, $P(G_B \cap G_C)$ is interpreted as the probability that one, *B* or *C*, would have caused E if the other hadn't already done so.

These examples illustrate the tacit assumption made in logistic regression regarding how causal influences combine. Applying an additive decomposition function (i.e., additive model)—where estimated regression weights are the unique contribution of each factor on top of the base rate—to dichotomous data will inherently assume that the causal influences of candidate and background causes are mutually exclusive (as indicated by 0 intersection). In contrast, applying a noisy-logical decomposition function to dichotomous data will assume that the causal influences of candidate and background cases are independent.

Curse of Symmetry. The predicted probability of an outcome obtained through the inverse logit transformation follows a sigmoidal function symmetrical around 0.5 (see Figure 2); consequently, this symmetry will highlight divergent conclusions from logistic regression and causal power. Consider the following scenario taken from Cheng, Sandhofer, and Liljeholm

(2022)[4], assuming random assignment. At a farm, 90% of animals who were not fed grains had red dots on their faces; only 60% of animals who were fed grains had red dots on their faces. At the zoo, 40% of the animals who were not fed grains *and* leaves had red dots on their faces; only 10% of animals who were fed grains and leaves had red dots on their faces. Which treat, grains or leaves, is more effective at curing the red dots? When posed to two-year old children, toddlers overwhelmingly identified leaves as the more effective treat, consistent with the CDC's definition of vaccine efficacy. At a young age, children, like adults, account for the base rate; a 30% difference is more consequential when the base rate is 40% than when the base rate is 90%. Figure 1 highlights this difference. Therefore, when the leaves were introduced, while holding the effect of grains constant, there was proportionally a much greater improvement (75% versus 33%).

However, separate logistic regressions for each context would conclude that the improvement after eating grains alone (farm) and grains with leaves (zoo) are equal, as shown by the black (change in probability) and blue (change in log odds) lines in Figure 2. Given that the model is not fully saturated (i.e., no data where animals only ate leaves), logistic regression would be unable to meaningfully estimate the causal strength of leaves alone.

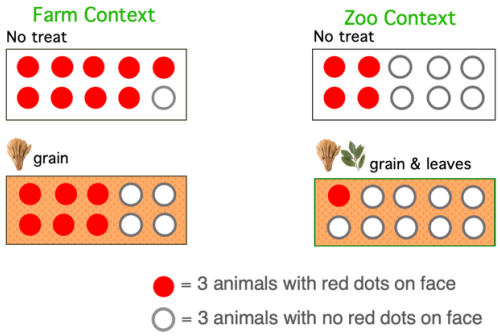

Figure 1: Outcome frequencies at the farm and the zoo before and after treatment (adapted based on Table 1 in Cheng et al., 2022).

[4] This problem was adapted from the original story, which was within-subjects. Regardless of whether the story was within or between-subjects, logistic regression identified grains as the more effective treatment.

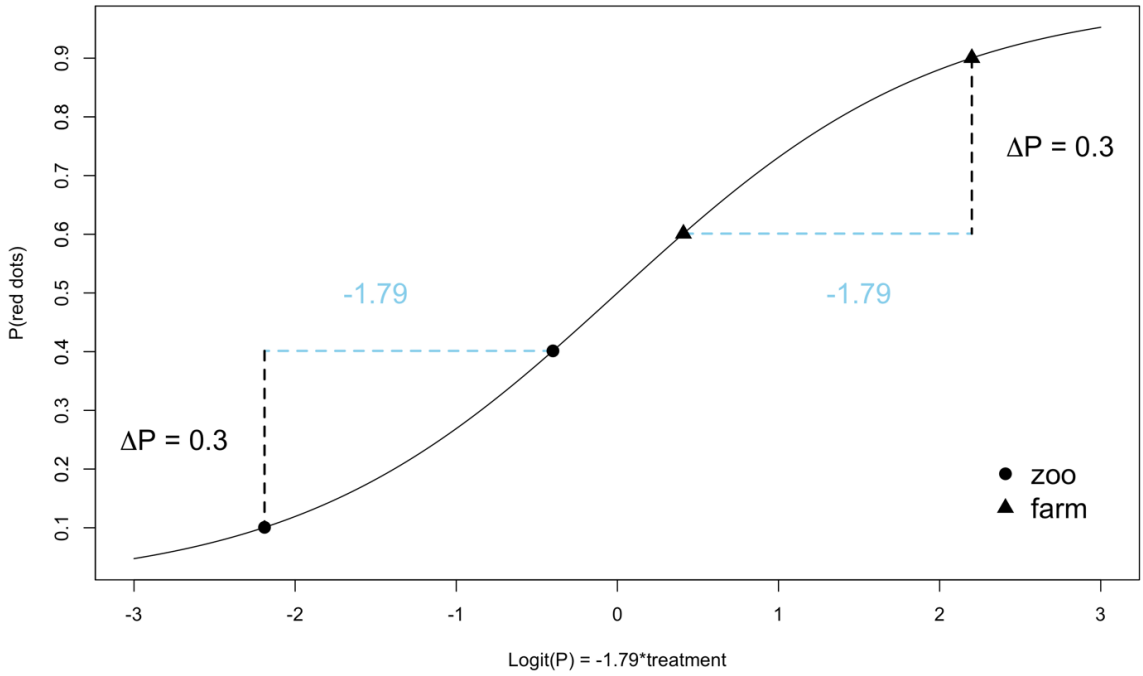

Figure 2: Sigmoidal function where the y-axis is the predicted probability, and the x-axis is the linear model excluding location-specific intercepts. In logistic regression, a 0.3 difference (blue lines) is equally as consequential going from 0.4 to 0.1 as it is from 0.9 to 0.6.

In contrast, the causal invariance approach would first compute the preventative causal power of grains alone (candidate cause C_1) from the farm context according to Equation 4: $P_{C_1} = \frac{1}{3}$, which is constant across contexts. For greater clarity, the following derivations converts the current causal structure (preventative) into a generative one and re-writes the independent union equation as shown in Equation 1. Let *E* indicate the effect (red dots; sick), and *E'* indicate the complement (no red dots; healthy). Therefore, we can replace *E* with *E'* from Equation 3. At the zoo, 60% of animals were healthy without treatment, and 90% were healthy with treatment. Given the union (90%), base rate (60%), and causal power of grains (1/3)[5], one can solve for the causal

[5] Preventative and causal powers are always equal.

power of leaves (candidate cause C_2), P_{C_2}. Consistent with human causal judgments, causal power would conclude that the independent effect of leaves on red dots is greater than that of grains.

$$P\left(G_B \cup P_{C_1} \cup P_{C_2}\right) = 1 - (1 - G_B) \cdot \left(1 - P_{C_1}\right) \cdot \left(1 - P_{C_2}\right)$$

$$\frac{9}{10} = 1 - \left(1 - \frac{6}{10}\right) \cdot \left(1 - \frac{1}{3}\right) \cdot \left(1 - P_{C_2}\right)$$

$$\Longrightarrow P_{C_2} = \frac{5}{8}$$

Simulating data for a fully saturated model assuming independent causal influences (see Figure 3 for data-simulating process), for the farm and zoo separately, we see that the estimated causal strength of grains is not constant across the two locations. At the farm, where animals osnly received grains, the estimated beta weight for "grains" is $b = 1.80$, $OR = 5.99$, $p < .001$. At the zoo, where animals received any combination of grains and leaves, the estimated beta weight for "grains" is $b = 0.55$, $OR = 1.73$, $p < .001$. Analyzing the data together, the effect of grains significantly depended on location, $b_{interaction} = 1.27$, $OR = 3.58$, $p < .001$. In other words, the estimated causal strength of grains under the logistic framework is not independent of context and other causes. However, given data for a fully saturated model in the zoo context, logistic correctly identified leaves as the more effective treatment, $b_{leaves} = 1.27$, $OR = 3.55$, $p < .001$, compared to grains, $b_{grains} = 0.55$, $OR = 1.73$, $p < .001$.

```
Simulate Fully Saturated Model

df_farm = function(sample_size,power1,base_rate){
  n = sample_size/2
  c1 = c(rep(0,n),rep(1,n))                    # grains
  px1 = power1
  pb = base_rate
  p = 1-(1-px1*c1)*(1-pb)                      # union rule
  y = rbinom(sample_size,1,p)                  # bernoulli response variable
  return(cbind(c1,y))
}

# Flip to generative (base rate: 1/10 healthy)
DF_farm = as.data.frame(df(sample_size=400000,base_rate=1/10,power1=1/3))

df_zoo = function(sample_size,power1,power2,base_rate){
  n = sample_size/2
  c1 = c(rep(0,n),rep(0,n)rep(0,n),rep(1,n))  # combination of grains and leaves
  c2 = c(rep(0,n),rep(1,n)rep(0,n),rep(1,n))
  px1 = power1
  px2 = power2
  pb = base_rate
  p = 1-(1-px1*c1)*(1-px2*c2)*(1-pb)           # union rule
  y = rbinom(sample_size,1,p)                  # bernoulli response variable
  return(cbind(c1,y))
}

# Flip to generative (base rate: 6/10 healthy)
DF_zoo = as.data.frame(df(sample_size=400000,base_rate=6/10,power1=1/3,power2=5/8))
```

Figure 3: Simulating data for a fully saturated design at the farm (top) and zoo (bottom).

Summary. This section reviewed the type of outcome data that is appropriate for logistic regression, and how different explanatory factors contribute to it. Though the outcome variable of interest is originally dichotomous (e.g., COVID infection), logistic regression does not directly model the probability, which is either 0 or 1 in observed data. Instead, logistic regression models the *log odds* of the outcome (a continuous value) and decomposes (i.e., models) the potential causal effects additively. Applying an additive decomposition function to dichotomous data tacitly assumes that causal influences are *mutually exclusive*. Secondly, the estimated causal effects under the logistic regression framework do not account for the base rate as people do when reasoning about causal strength. Therefore, if two treatments reduce illness by 30%, logistic regression would

conclude that the two treatments are equally as effective regardless of whether the base rate is 40% or 90%. In contrast, causal power as estimated under the causal invariance framework, decomposes the potential causal effects in accordance with the noisy-logical decomposition function, which assumes that the causal influences are independent of each other; it is interpreted as the probability that the candidate cause would have generated/prevented the effect. As a result (Equation 2), causal power estimates are dependent on the base rate. Therefore, if two treatments reduce illness by 30%, causal power would conclude that a treatment given a base rate of 40% is more consequential than if the base rate were 90% (i.e., proportionally a greater improvement). The following section introduces an alternative approach that 1) assumes independent causal influences between candidate and background causes, 2) directly models the probability, and 3) directly estimates causal power among covariates.

Causal Invariance for Statistical Inference (CISI)

Why Causal Invariance. When the outcome variable is dichotomous, what is reasonable to assume about the intersection term, $P(G_B \cap G_C)$? Given generalization (i.e., useability) as a goal and infinitely many possible causal representations (i.e., problem of underdetermination), the causal invariance framework (Cheng, 1997) posits that the only sensible *default* assumption is for the candidate and background causes exert their influences onto the outcome independently[6]. While some (e.g., Pearl, 2022; Griffiths & Tenenbaum, 2005) have argued that this assumption is untenable and unnecessary, there is strong psychological evidence suggesting that people make this default assumption (e.g., Buehner, Cheng, & Clifford, 2003; Bye et al., 2023; Liljeholm & Cheng, 2007; Cheng et al., 2022). It is important to note that the causal invariance framework does

[6] To the issue of generalizability, estimating causal strength assuming independence, not mutual exclusivity, is the only generalizable solution between the two. To the issue of underdetermination, there are infinitely many ways in which the candidate and background causes can interact but only one way in which they can be independent.

not assert that the influence of different causes on the effect (candidate and background) *must* be independent; it simply states that it is the only sensible *default* assumption for hypothesis revision (Park, McGillivray, Bye, & Cheng, 2022).

Humans seek parsimonious, useable, and thus generalizable causal knowledge (Bye et al., 2022; Cheng, 1997; Park et al., 2022). For learned causal knowledge to be useful, we must be able to generalize our knowledge learned from one context to another. When the knowledge is generalizable from one context to the next, each with differing background causes, then the effect of the candidate cause on the outcome is independent from the effect of the background causes. When learned causal knowledge does not apply in a new context (i.e., when the candidate and background causes interact), the resulting discrepancy signals reasoners to revise their beliefs/knowledge (i.e., hypothesis) towards greater coherence (Bye et al., 2023; Cheng, 1997; Cheng, Novick, Liljehom, & Ford, 2007; Cheng & Lu, 2017; Cheng, Sandhofer, Liljehom, 2022; Liljeholm & Cheng, 2007; Rescorla & Wagner, 1972; Woodward, 2000, 2006). Take, for example, a toddler learning (in the comfort of her dining room) that striking a match produces fire. A few weeks later, the toddler goes on a family camping trip, where she attempts to apply her newly acquired knowledge, except this time, it was outdoors and in the rain. When the toddler strikes the match, she realizes that it does not light; this experience is inconsistent with her current causal model, signaling the toddler to revise her causal model to incorporate air moisture. Her new causal model should now be that striking a match produces a fire when the air moisture is low; this knowledge is then assumed to generalize to a new context with different, other background causes, until proven otherwise. For the same reason that we do not assume the alternative hypothesis as the default, we do not assume by default an interaction between learned causal structures and background causes; there are infinitely many ways the learned structure can potentially interact

with background causes (i.e., problem of underdetermination). Ichien and Cheng (2022) explicate the necessity of assuming independent influences as a default hypothesis.

Real-world consequences. The previous section highlighted logical inconsistency between logistic regression estimates of causal strength (ΔP) and causal power in two hypothetical examples; however, there have been demonstratable consequences of assuming the incorrect decomposition function (*additive* for binary outcome) in real-life public health problems.

The famous Seven Countries Study (Keys, 1980) reports an analysis of large-scale data on coronary heart disease deaths (CHD) across multiple countries that led to a two-decade-long recommendation by U.S. public health authorities and the popular media to adopt a low-fat diet. The low-fat diet recommendation was based on the study's regression analyses indicating that dietary fat is highly correlated with the incidence of CHD, but that holding dietary fat constant, dietary sucrose has no significant effect on CHD. The recommendation led to a prevalence of processed foods at supermarkets that had low fat but high sugar to compensate for flavor—products that likely have contributed to the obesity epidemic and metabolic syndrome worldwide. According to a review of health recommendations (La Berge, 2008), "After 1980, the low-fat approach became an overarching ideology, promoted by physicians, the federal government, the food industry, and the popular health media."

By assuming different decomposition functions, logistic regression (additive decomposition) and vaccine efficacy (noisy-logical decomposition) are not logically consistent with each other (Cheng et al., 2022; Ichien & Cheng, 2022). To our knowledge, there is no published analog of logistic regression that has its flexibility of taking both discrete and continuous variables as predictors but is logically consistent with vaccine efficacy—in other words, an analog that assumes a direct, independent influence of the binary predictors on the binary outcome.

Causal Invariance for Statistical Inference (CISI) Model

Let E = effect, C = candidate cause (1 when present, 0 when absent), B = background causes, $G_B = P(B) * P(B \rightarrow E)$, X_1 indicate the first continuous covariate, X_2 indicate the second continuous covariate, and $G_C = P(C) * P(C \rightarrow E)$indicate the generative causal power (see Equation 3).

Candidate Cause. For a binary candidate cause (e.g., vaccine treatment), the generative causal power, G_C, is estimated as shown in Equation 3.

Background causes. For continuous background causes (X_1 and X_2), their causal influences on the outcome are combined linearly (i.e., additively) via logistic regression. Their composite causal influence is represented as G_B: the causal influence of background causes (B).

$$B = \text{logit}(P) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

$$G_B = \frac{1}{1 + exp(-B)}$$

Probability of outcome. The probability of either the background (B) or candidate cause (C) leading to the effect is computed as follows in accord with the union rule assuming independence:

$$P(G_B \cup G_C) = 1 - (1 - G_B) * (1 - G_C)$$

Bayesian estimation. The Bayesian posterior distributions of each model parameters were estimated via multi-chain Monte Carlo (MCMC) sampling through **rstan**. The burn-in period for each of the two chains was 1100. Sampling terminated after 10^5 iterations. All chains converged within 10^5 iterations ($M_{MPSRF} = 1.00$, $\min_{MPSRF} = 0.999$, $\max_{MPSRF} = 1.002$). The priors for each parameter are as follows (the second parameter in the normal is precision):

$$\beta_0 \sim N(0, 0.00001)$$

$$\beta_1 \sim N(0, 0.00001)$$

$$\beta_2 \sim N(0, 0.00001)$$

$$P_C \sim \text{Uniform}(0, 1)$$

Simulation

The following results demonstrate that the estimate of causal strength derived from logistic regression (regression weight β_C) does not coincide with causal power. In the first scenario, causal power remains constant as base rates increase. In the second scenario, causal power increases with base rate. One hundred iterations were run for each scenario for a total of 200 iterations. In each iteration, data were randomly generated according to the CISI model across five contexts (see Figure 4). Tables 1 and 2 illustrate each of the five contexts in scenarios 1 and 2, respectively. For all contexts, base rates (specified as a probabilities) were converted to log odds (β_0). The data in each iteration were analyzed with both CISI and logistic regression. Total run times were 11.57 hours and 11.80 hours for the first and second scenarios, respectively. Both models accurately estimated the regression weights: β_0, β_1, and β_2 (see appendix). The posterior modes of the estimated CISI parameters are shown in all the following figures.

Data Generation

```
df = function(b0,b1,b2,sample_size,power){
  c = rep(c(0,1),sample_size/2)        # p(c) = 1 or 0
  x1 = rnorm(sample_size)              # some continuous variables
  x2 = rnorm(sample_size)
  b = b0 + b1*x1 + b2*x2               # linear combination with a bias
  gb = 1/(1+exp(-b))                   # pass through an inv-logit function
  gc = power                           # power: gc = 1*p(c->e)
  p = 1-(1-gc*c)*(1-gb)                # union rule
  y = rbinom(sample_size,1,p)          # bernoulli response variable
  return(cbind(c,x1,x2,y))
}
```

Figure 4: R code to generate data according to the model.

Table 1. First scenario where causal power remains constant across base rates.

	Context 1	Context 2	Context 3	Context 4	Context 5
Base Rate	0.05	0.25	0.45	0.65	0.85
Causal Power	8/10	8/10	8/10	8/10	8/10
β_1	0.4	0.4	0.4	0.4	0.4
β_2	0.5	0.5	0.5	0.5	0.5

Table 2. Second scenario where causal power increases with base rate.

	Context 1	Context 2	Context 3	Context 4	Context 5
Base Rate	0.05	0.25	0.45	0.65	0.85
Causal Power	3/8	3/7	3/6	3/5	3/4
β_1	0.4	0.4	0.4	0.4	0.4
β_2	0.5	0.5	0.5	0.5	0.5

As shown in Figure 5, when true causal power remains constant at 0.80 (red) across five increasing base rates, causal power estimates from the CISI model (black) also remain constant. However, logistic regression estimates do not remain constant across base rates, as shown in Figure 6. Two separate linear regressions tested the relationship between base rate and causal strength estimates. Overall, causal power estimates (standardized) did not vary significantly with base rates in scenario 1, b = -0.18, $t(498)$ = -1.16, p =.249. On the other hand, logistic regression estimates (standardized) decreased significantly as base rate increased, b = -1.48, $t(498)$ = -10.30, $p < .001$.

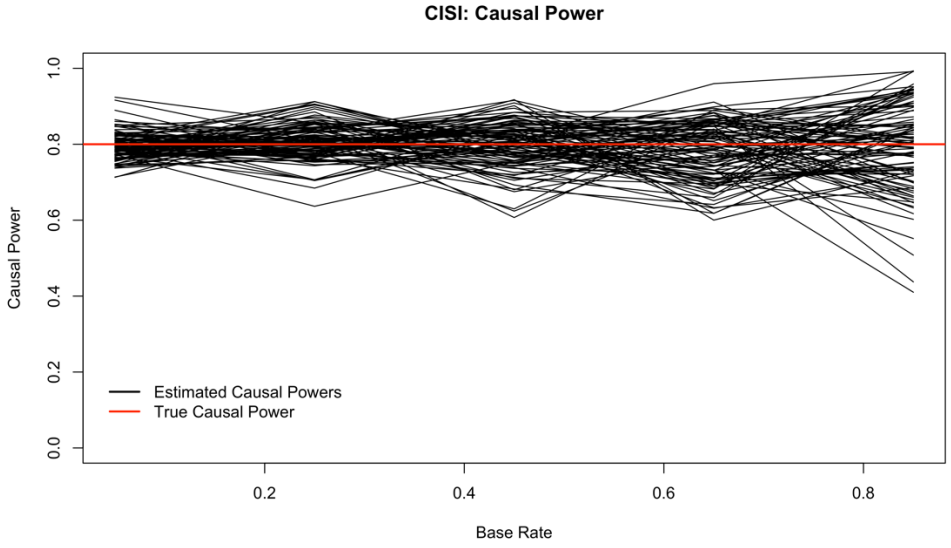

Figure 5: CISI causal power estimates from scenario 1.

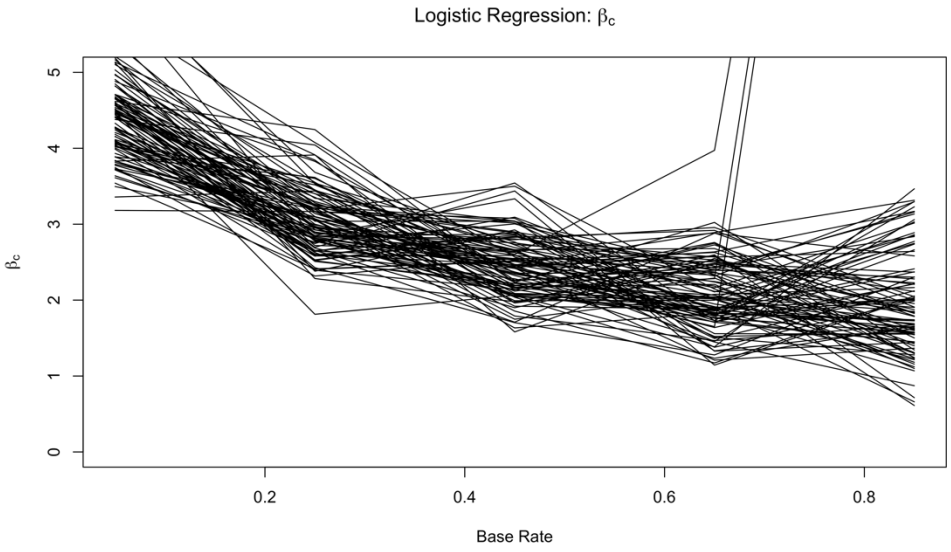

Figure 6: Logistic regression weight estimates from scenario 1.

Figures 7 and 8 show the simulation results from scenario 2, where causal power increases with base rate. As shown in Figure 7, causal power estimates from the CISI model (black) correctly trace the trajectory of the true causal power (red). However, linear regression estimates do not increase accordingly (see Figure 8).

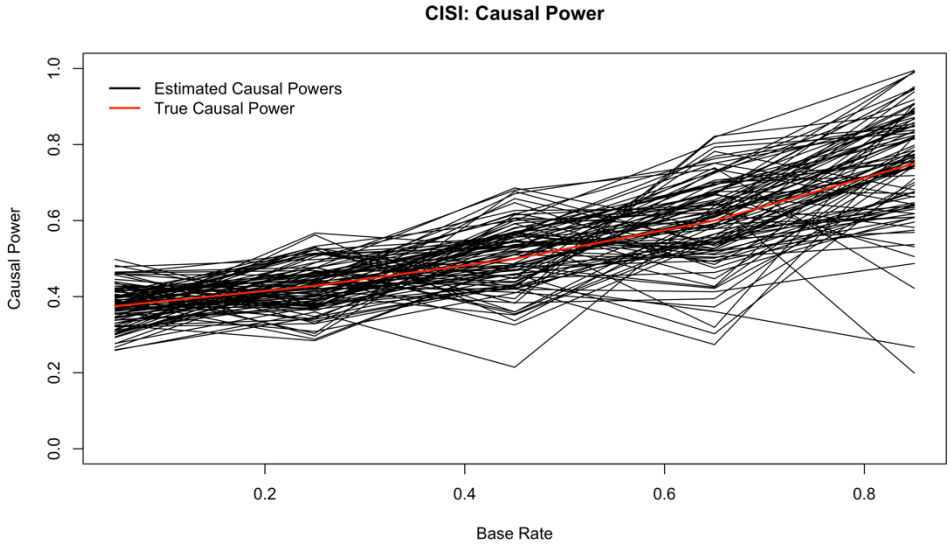

Figure 7: CISI causal power estimates from scenario 2.

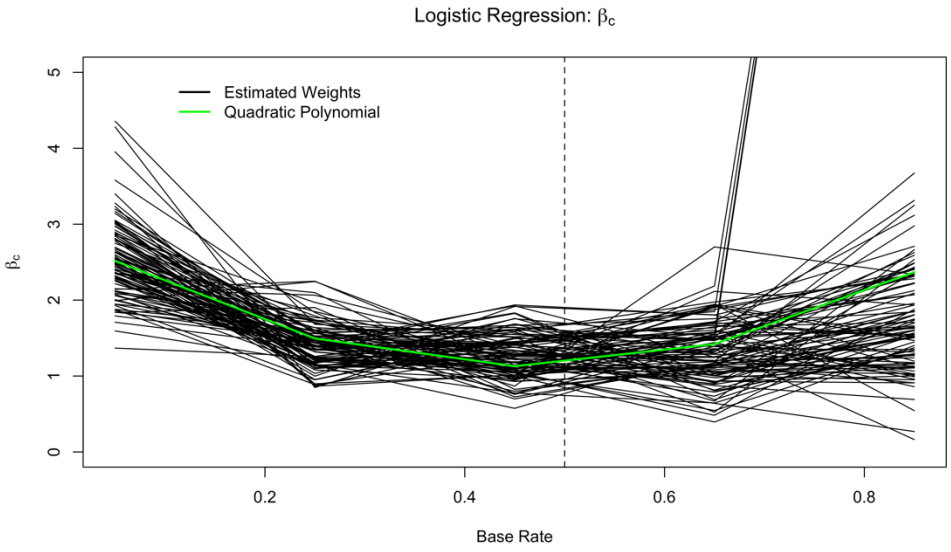

Figure 8: Logistic regression weight estimates from scenario 2.

As for scenario 1, two separate regressions tested the relationship between base rate and causal strength estimates. Overall, causal power estimates (standardized) increased significantly with base rates in scenario 2, $b = 2.79$, $t(498) = 28.70$, $p < .001$. On the other hand, logistic regression estimates (standardized) did not increase linearly with base rate. Instead, there was an

evident quadratic trend (green) symmetric around base rate = 0.5 (dotted line), as expected due to symmetry of the sigmoidal function, $b = 4.93$, $t(497) = 7.81$, $p < .001$. Both methods were able to accurately estimate the regression coefficients (β_0 , β_1 , β_2) (see appendix).

Interim Discussion

This set of simulation results demonstrates the logical inconsistency between logistic regression and causal power. Logistic regression is a standard model used to model binary outcomes. However, by assuming an additive decomposition function, logistic estimates of causal strength do not vary accordingly with true causal power, which assumes a noisy-logical decomposition function. This section introduced and evaluated a causal analog to logistic regression that directly estimates the causal power of a candidate cause among covariates; the estimates from this model vary with true causal power.

Section 2: Artificial Intelligence

For many years, artificial intelligence (AI) researchers have aimed to develop large neural networks (e.g., language models) that exhibit a wide range of general capabilities such as programming, mathematical reasoning, logical reasoning, and planning. Numerous large language models (LLMs) such as ChatGPT have already been implemented in the workplaces to write emails and edit code. Most recently, OpenAI released GPT-4, a LLM with likely more than 175 billion parameters. While less capable than humans in many real-world scenarios, GPT-4 exhibits striking general intelligence comparable to human-level performance on various professional and academic benchmarks, such as scoring in the top 10% on the Bar Exam (OpenAI, 2023). A recent report published by Microsoft Research (Bubeck et al., 2023) discusses the rising capabilities and implications of these models, but also emphasizes the limitations of current artificial general intelligence (AGI) systems. Importantly, future research should pay close attention to model

evaluation and potentially pursue a paradigm that goes beyond next-word prediction; this report emphasizes the need for psychology-based testing in lieu of the current benchmark tests. While GPT-4 is not freely available, this final section of the thesis evaluates GPT-3 on a novel benchmark task that provides a stricter test of causal reasoning, a general ability that supports human creativity, scientific discovery, evaluation of truth, and more (Waldmann, 2017; Mitchell & Krakau, 2023; Mahowald, Ivanova, Blank, Kanwisher, Tenenbaum, & Fedorenko, 2023; Binz & Schulz, 2023).

Language, Transformer, and Thought

Decades of psychological research has established that language is not necessary for thought but serves as a tool to express thoughts and potentially influence them (Gal'perin, 1992; Fedorenko & Varley, 2016; Mandler, 2004). The transformer architecture has revolutionized natural language and computer vision models. Transformer models are deep neural networks that adopt the self-attention mechanism, which allows for advanced contextualization, especially in language (Vaswani et al., 2017); OpenAI's Generative Pre-Trained Transformer (GPT) models capitalize on this architecture. GPT is a LLM trained on copious amounts of unlabeled text data (e.g., Wikipedia), which learns to estimate the probability of any text sequence through next-word prediction. If GPT is taken to be an apt model of language, is it enough for thought? While recent reports suggest so (e.g., Brown et al., 2020; Chowdhery et al., 2022), it is possible that these LLMs have had access to the benchmarking tasks during training (e.g., Srivastava et al., 2022; OpenAI, 2023; Bubeck et al., 2023). Moreover, it is crucial to interrogate the validity of current benchmarking tasks. LLMs such as OpenAI's GPT-3 and Google's Pathways Language Model (PaLM) claim that these models can accomplish a wide range of general reasoning tasks at or beyond human level.

One of these tasks is *causal reasoning*. Currently, the standard causal reasoning task is from the Beyond-the-Imitation-Game (BIG) benchmark. In this task, a language model must identify which of two statements is more sensible; for example: 1) He turned on the windshield wipers because it was raining, or 2) It was raining because he turned on the windshield wipers. Given that transformer based LLMs are trained to predict the likelihood of text sequences, it is no surprise that they will succeed at tasks like this one as it is very unlikely for the second sentence to appear anywhere in natural text. A stricter test of causal reasoning would be those problems discussed in Section 1. The psychological literature has shown that the aforementioned problems can only be solved by causal (e.g., causal power from causal invariance), not associative models (e.g., logistic regression), therefore providing a stricter test of causal reasoning. This final section translates the causal problems from Buehner, Cheng, & Clifford (2003)—similar to the headache problem—and Cheng et al. (2022) into language problems for GPT-3 and ChatGPT-4 to solve. Recall human judgments from the headache (causal power = 27/36; Ichien & Cheng, 2022) and farm-and-zoo (leaves as the more effective treatment; Cheng et al., 2022) examples from Section 1. Figure 9 depicts GPT-3's response to the headache problem[7], and Figures 10-11 depicts GPT-3's responses to the farm-and-zoo problem using both text-davinci-002 and 003 engines[8]. From the generated response for the headache problem, GPT-3's answer seems to match the final proportion of individuals with headaches in the treatment group; it fails to decompose the outcome into its constituent causes. For the farm-and-zoo problem, text-davinci-002 arrived at the correct answer (leaves) by accident and text-davinci-003 arrived at the incorrect answer (grains). Figure 12 shows ChatGPT-4's response to the farm-and-zoo problem; for the same reason logistic regression could not meaningfully estimate the effect of leaves, ChatGPT-4 could not judge which

[7] Answers does not differ between the text-davinci-002 and 003 engines.
[8] Answers do not change when temperature is changed to 0.

treat (grain or leaves) was more effective at reducing red dots. The model's response states that the design is not fully saturated. Given these test examples, it seems unlikely that neither GPT-3 nor ChatGPT-4 can truly engage in causal reasoning.

Headache

```
prompt = "In a clinical trial, individuals were randomly assigned to a control
group or a treatment group. The treatment group received medication and the control
group received nothing. 12 the 36 individuals in the control group reported having
a headache. 30 out of 36 individuals in the treatment group reported having a
headache. If 36 new individuals without headaches were given the medication, how
many would have a headache?"

response = openai.Completion.create(
  engine="text-davinci-003",
  prompt= prompt,
  temperature=0.5,
  max_tokens=256,
  top_p=1.0,
  frequency_penalty=0.0,
  presence_penalty=0.0

)
print(response['choices'][0]['text'])

"30 out of 72 individuals (36 from the treatment group and 36 from the new group)
would have a headache."

"It is impossible to predict how many of the 36 new individuals would have a
headache after taking the medication. This would depend on individual factors such
as age, health history, and other factors."

"30 out of 36 individuals in the treatment group reported having a headache, so if
36 new individuals without headaches were given the medication, it is likely that
30 out of the 36 would report having a headache."
```

Figure 9: GPT-3's responses to the headache problem across three different runs.

Farm-and-zoo-002

```
prompt = "At a farm, 90% of animals were ill. After feeding the animals
grains, 60% were ill. At a zoo, 40% of animals were ill. After feeding
the animals grains and leaves, 10% were ill. Which of the two treatments,
grains or leaves, are more effective and why?"

response = openai.Completion.create(
  engine="text-davinci-002",
  prompt= prompt,
  temperature=0.5,
  max_tokens=256,
  top_p=1.0,
  frequency_penalty=0.0,
  presence_penalty=0.0

)

print(response['choices'][0]['text'])

"The treatment with leaves is more effective because it resulted in a
smaller percentage of animals being ill."
```

Figure 10: GPT-3's response to the farm-and-zoo-problem (engine = text_davinci_002).

Farm-and-zoo-003

```
prompt = "At a farm, 90% of animals were ill. After feeding the animals grains, 60%
were ill. At a zoo, 40% of animals were ill. After feeding the animals grains and
leaves, 10% were ill. Which of the two treatments, grains or leaves, are more
effective and why?"

response = openai.Completion.create(
  engine="text-davinci-003",
  prompt= prompt,
  temperature=0.5,
  max_tokens=256,
  top_p=1.0,
  frequency_penalty=0.0,
  presence_penalty=0.0

)

print(response['choices'][0]['text'])

"Grains are more effective than leaves, because the percentage of ill animals
decreased more significantly when grains were used. At the farm, the percentage of
ill animals decreased from 90% to 60%, while at the zoo, the percentage of ill
animals decreased from 40% to 10%."
```

Figure 11: GPT-3's response to the farm-and-zoo-problem (engine = text_davinci_003).

GPT-4 Farm-and-Zoo

```
prompt = "At a farm, 90% of animals were ill. After feeding the animals grains, 60%
were ill. At a zoo, 40% of animals were ill. After feeding the animals grains and
leaves, 10% were ill. Which of the two treatments, grains or leaves, are more
effective and why?"


"Based on the information provided, it is difficult to definitively determine which
treatment, grains or leaves, is more effective in reducing illness among animals.
This is because we cannot isolate the individual effects of grains and leaves in
the zoo case, as the animals were given both treatments simultaneously.

In the farm case, we know that after feeding the animals grains, the percentage of
ill animals decreased from 90% to 60%. This suggests that the grains treatment had
a positive effect, reducing illness by 30 percentage points.

In the zoo case, the percentage of ill animals decreased from 40% to 10% after the
animals were fed both grains and leaves. While this is a 30 percentage point
reduction in illness, we cannot attribute this change solely to either the grains
or the leaves, as both were provided simultaneously.

To determine the effectiveness of the individual treatments, we would need more
information or a controlled study where the effects of grains and leaves can be
observed independently."
```

Figure 12: ChatGPT-4's response to the farm-and-zoo problem.

Causal Tasks

The following tasks have been administered to human reasoners in psychological studies of causal cognition. Table 3 is taken from Buehner et al. (2003); each of the ten questions will be adapted to be a story-based problem. In each question, there is a medicine in question (A-J). The first subtask requires GPT to identify whether the medication generates, prevents, or has no effect on headaches. The second subtask requires GPT to estimate the causal power of the medication. Probability of headache in the treatment and control groups in each of the ten problems are in accord with Table 3. For example, in Condition A, 18/36 (50%) individuals reported a headache in the control group and 36/36 (100%) reported a headache in the treatment group. Therefore, medicine A has a generative effect on headaches. The last two columns of Table 3 are the mean and median human causal ratings from Buehner et al. (2002), respectively.

Table 3. Ten problems taken directly from Experiment 2 of Buehner et al. (2003).

Design and Results of Experiment 2

Condition	Power	ΔP	$P(e\|c)$	$P(e\|\sim c)$	Causal ratings	
					M (*SD*)	*Mdn*
A	1.00	0.50	36/36	18/36	85.7 (26.5)	100
B	0.75	0.50	30/36	12/36	67.8 (19.1)	75
C	0.75	0.75	27/36	0/36	74.4 (5.7)	75
D	0.50	0.50	18/36	0/36	48.5 (14.1)	50
E	0.75	−0.50	6/36	24/36	−59.7 (31.4)	−75
F	0.75	−0.75	9/36	36/36	−66.5 (18.8)	−75
G	0.50	−0.50	18/36	36/36	−44.5 (21.1)	−50
H	0.00	0.00	12/36	12/36	−0.5 (4.86)	0
I	0.00	0.00	24/36	24/36	0.7 (3.78)	0
J	1.00	−0.50	0/36	18/36	−85.9 (26.6)	−100

Note. The $P(e|c)$ and $P(e|\sim c)$ columns list how many of the sample of 36 patients in each respective group showed the effect in each condition. Preventive ratings are represented by negative numbers.

Subtask 1. In the first subtask, GPT was given the following prompt and asked to state whether medicine X prevented, generated, or had no effect on headaches:

> "You are an employee for a Company that distributes new medicines for preventing allergies. Your job is to review information regarding a possible side effect of the new allergy medicines that are under consideration for distribution. Although these medicines have been found to be clearly effective in preventing allergies, they may cause headaches, prevent headaches, or have no influence at all on headaches. You will see the results of experiments that were conducted to study the influence of these medicines on headaches. For each study, patients were randomly assigned to one of two groups: an experimental group that received the new medicine, and a control group that received a placebo. Based on the data presented, judge whether each medicine has a side-effect on headaches, and if so, whether it causes or prevents them. Your success in the company is highly dependent on your accurate assessment of these side effects.

We conduct a study of medicine [X] and find that: [$P(e|c)$] % of the participants who received medicine [X] (those in the experimental group) have headaches. Likewise, [$P(e|\sim c)$] % of the participants who did not receive medicine [X] (those in the control group) have headaches as well. Recall that participants were randomly assigned to the two groups.

QUESTION: Does medicine X prevent, generate, or have no effect on headaches?"

Subtask 2. In the original procedure from Buehner et al. (2003), if participants indicated that the medication prevented headaches, they were asked:

"QUESTION: How many of 100 people, all of whom have headaches, would not have a headache if given the medicine?"

If participants indicated that the medication had a generative effect on headaches, they were asked the follow-up question:

"QUESTION: How many out of 100 people, all of whom do not have headaches, would have a headache if given the medicine?

If participants indicated that the medication had no effect on headaches, their estimate of causal strength was automatically scored as zero.

Evaluation and Results

GPT-3 (text-davinci-003 engine) was evaluated on its top probability response with a maximum of 100 tokens at temperatures 0, 0.25, 0.5, 0.75, and 1, and ChatGPT-4 was evaluated based on its generated response.

As expected, GPT-3 was poor at correctly identifying the causal structure (~60% accuracy) and estimating the causal strength (see Tables 4 and 5, respectively). For the causal structure task,

it seems GPT-3 was able to correctly identify when the medication had no effect; otherwise, it almost always stated that the medicine prevented headaches even when the rate of headaches was higher in the treatment group (see appendix for exact answers). It is likely that GPT-3 is associating "prevention" with "treatment", "medication", and "headaches".

For subtask 2, GPT-3 was assumed to have correctly identified the causal structure and was thus asked the appropriate causal strength question; therefore, for the two medications that had no effect (H and I), GPT-3 was assumed to have answered 0. If GPT-3 indicated that the question cannot be answered, it was scored as a 0. Otherwise, GPT-3's judgments of causal strength did not coincide with human judgments (see Table 5). Based on GPT-3's responses (see Table A2 in appendix), the model was typically responding with the rate in the treatment group, completely ignoring the base rate. Furthermore, in its explanation of its answer, it sometimes correctly identified the causal structure given the additional context provided in the second sub-task (i.e., explicitly asked how many would or would not have a headache), even when it got the initial causal structure question wrong.

Given that ChatGPT-4 is fine-tuned for conversational tasks, it should perform better on the causal structure task. However, it is likely that it will still struggle with judgments of causal strength. ChatGPT-4 was able to correctly identify the causal structure in all the ten problems. However, Like GPT-3, ChatGPT-4 typically responded with the rate in treatment group in the second sub-task (causal strength). Its responses were correlated $r = 0.185$ with average human and $r = 0.175$ with median human responses. See Table A3 in the appendix for ChatGPT-4's exact responses to the causal strength questions.

Table 4. Accuracy for identifying causal structure (generative, preventative, or no effect).

	Accuracy (%)
GPT-3 (temp = 0)	60%
GPT-3 (temp = 0.25)	50%
GPT-3 (temp = 0.50)	60%
GPT-3 (temp = 0.75)	60%
GPT-3 (temp = 1)	60%

Table 5. Pearson correlations between human judgments and each model.

	Human Judgment (Mean)	Human Judgment (Median)
Causal Power	0.992	1.000
Delta P	0.852	0.813
GPT-3 (temp = 0)	0.155	0.08
GPT-3 (temp = 0.25)	0.158	0.098
GPT-3 (temp = 0.50)	0.114	0.053
GPT-3 (temp = 0.75)	0.298	0.252
GPT-3 (temp = 1)	0.414	0.361

Conclusions

This thesis demonstrates the critical role causality should play in our statistical inference and artificial general intelligence (AGI). The first section highlighted the logical inconsistency between a purely associative model (logistic regression) and a causal one (CISI) such that logistic regression's estimates of causal strength do not generally coincide with causal power. The second

and final section demonstrated that language models are not inherently causal models. While LLMs may be able to recognize causal language (e.g., Yu, Li, & Wang, 2019; Khetan, Ramnani, Anand, Gengupta, & Fano, 2022), it does not reason causally in the way humans do. The final section emphasized the importance of benchmarking in evaluating AGI. While current AGIs perform well on the causal reasoning tasks from the BIG-bench, GPT failed at a stricter test of causal reasoning. Beyond its inability to identify causal structure correctly and reliably, it was unable to estimate the causal strength in a way that was consistent with human ratings. Given that transformer-based LLMs learn through next-word-prediction, they do not have an inherent understanding of causality (see appendix for ChatGPT's response to the question, "Does AI have a representation of cause and effect?"). On the other hand, humans learn about the world through evaluating causal models against empirical evidence and revising such models when appropriate.

Future Research

Statistical Inference. While the CISI model can accurately estimate causal power, it takes significantly longer to run than logistic regression (42 versus 0.7 seconds). Moreover, runtime and convergence can vary when starting values in the MCMC chain are far from the true values. Future work may consider a maximum likelihood estimation in lieu of MCMC sampling. The model's estimation of the covariates' weights (see appendix) may also improve with maximum likelihood since the estimate and time to convergence are highly dependent on the starting values of the Monte Carlo chains.

AGI. While neural networks have been the dominant architecture for AGI systems, it is unclear whether they will ever be sufficient for thought. What has made connectionism (neural networks) so appealing for cognitive modeling has been its flexibility to perform a wide variety of tasks such as computer vision, programming, translation, and more. However, as demonstrated in

this thesis, much of cognition relies on structured representations (e.g., of cause-and-effect relationships) (Fodor & Pylyshyn, 1988; Quilty-Dunn, Porot, & Mandelbaum, 2022). While deep neural networks can be fine-tuned to optimally perform certain tasks, the fine-tuning process is theoretically opaque. Without some form of structured representation, especially for causality, LLMs will likely continue to struggle with commonsense tasks such as causal and moral attribution, explanation, and more. A hybrid architecture such as that introduced by Hummel & Holyoak (1997) achieve both the flexibility of a connectionist system and the structure sensitivity of a symbolic system. Such hybrid systems may provide competitive alternatives to next-word-prediction in achieving commonsense AGI.

Appendix

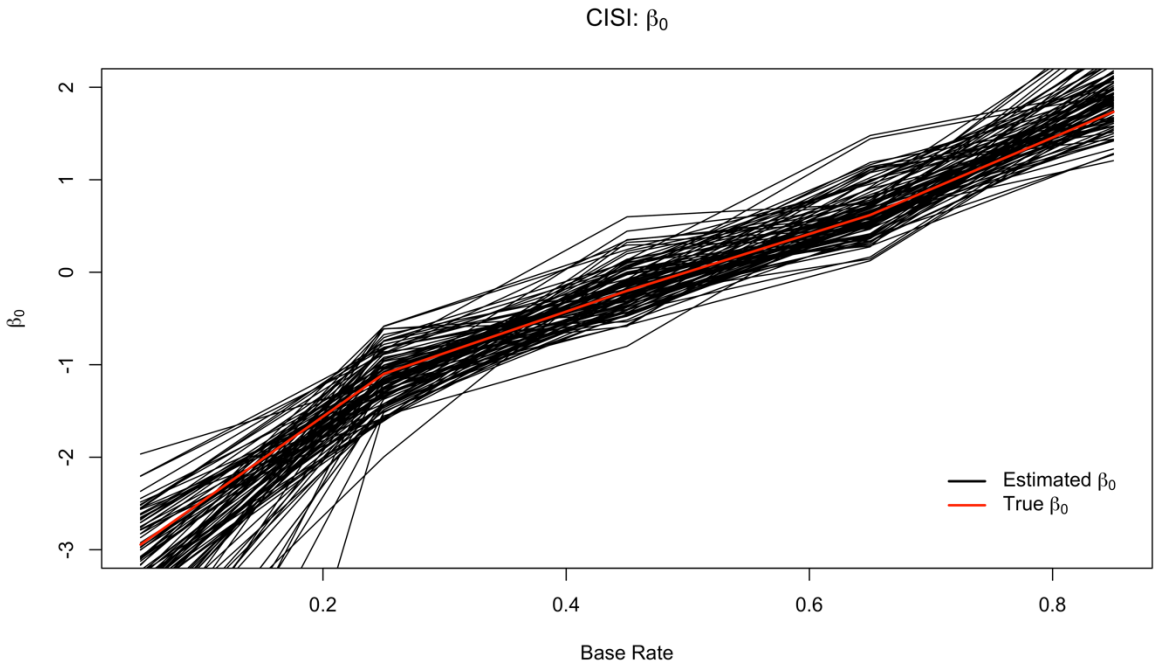

Figure A1: CISI estimates of β_0 in scenario 1.

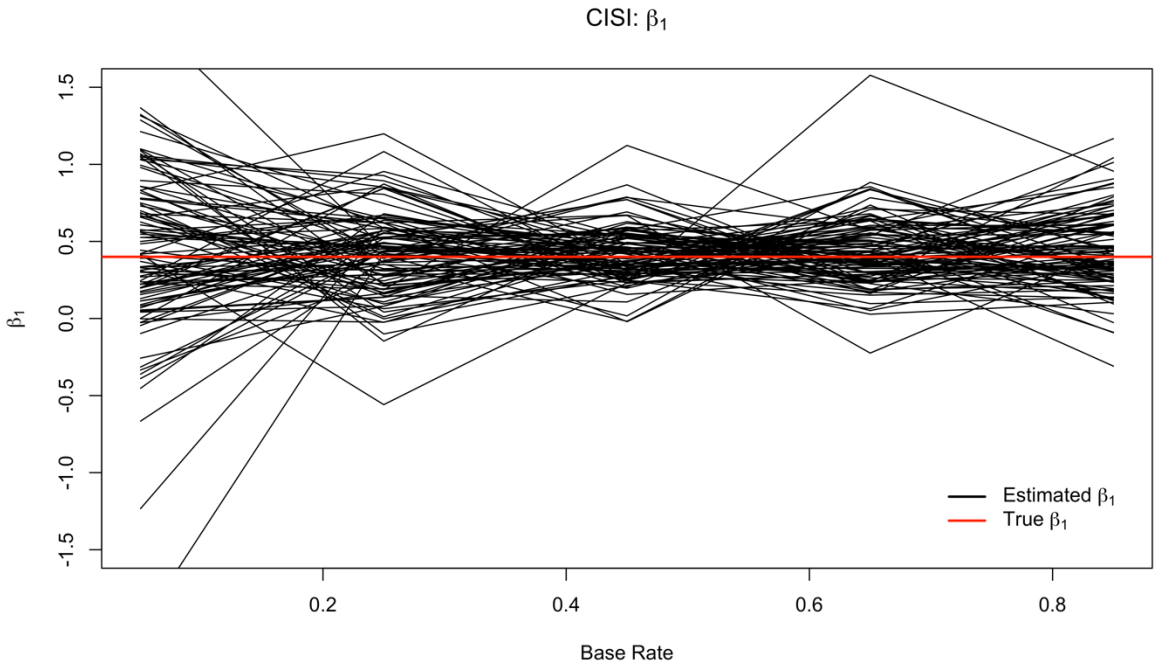

Figure A2: CISI estimates of β_1 in scenario 1.

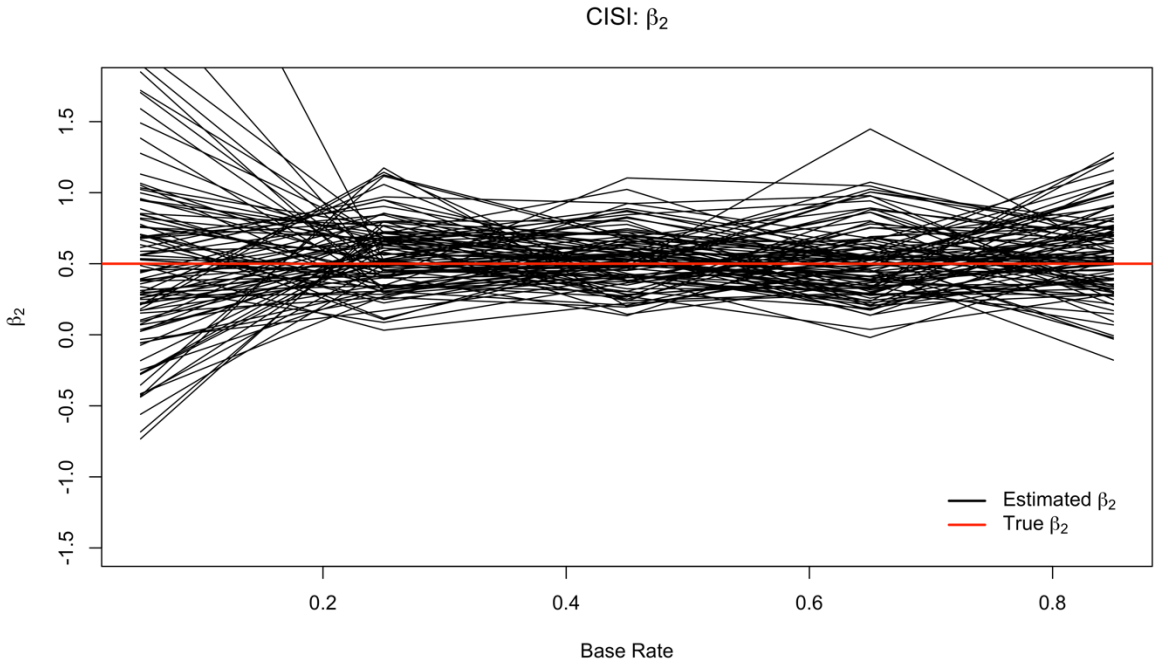

Figure A3: CISI estimates of β_2 in scenario 1.

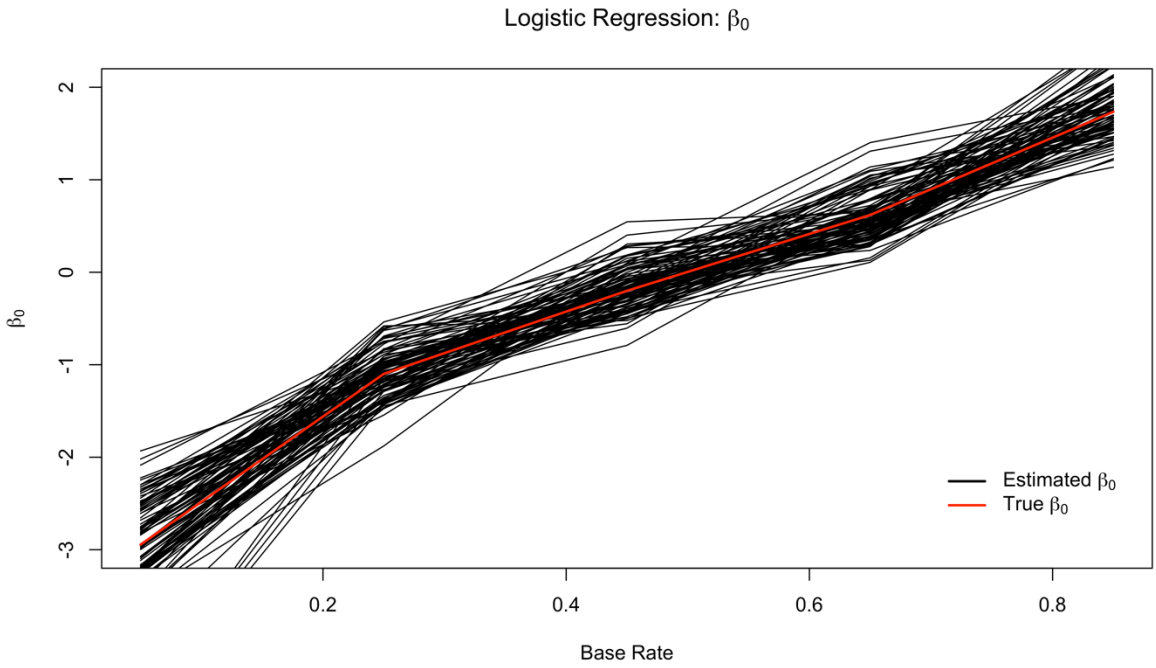

Figure A4: Logistic estimates of β_0 in scenario 1.

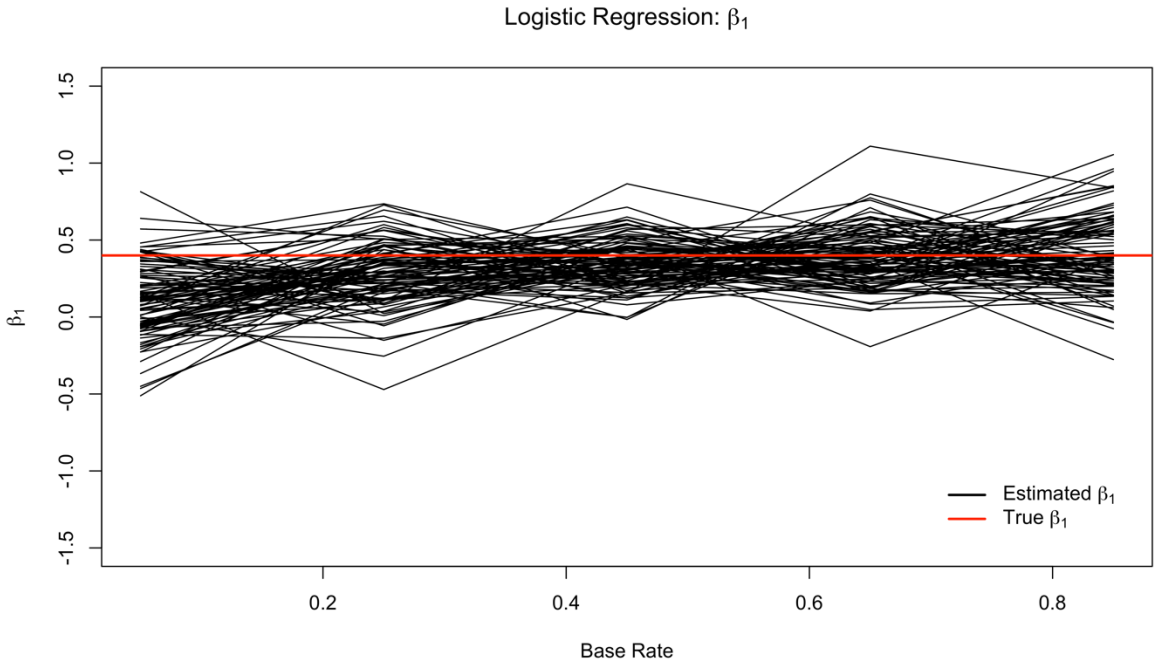

Figure A5: Logistic estimates of β_1 in scenario 1.

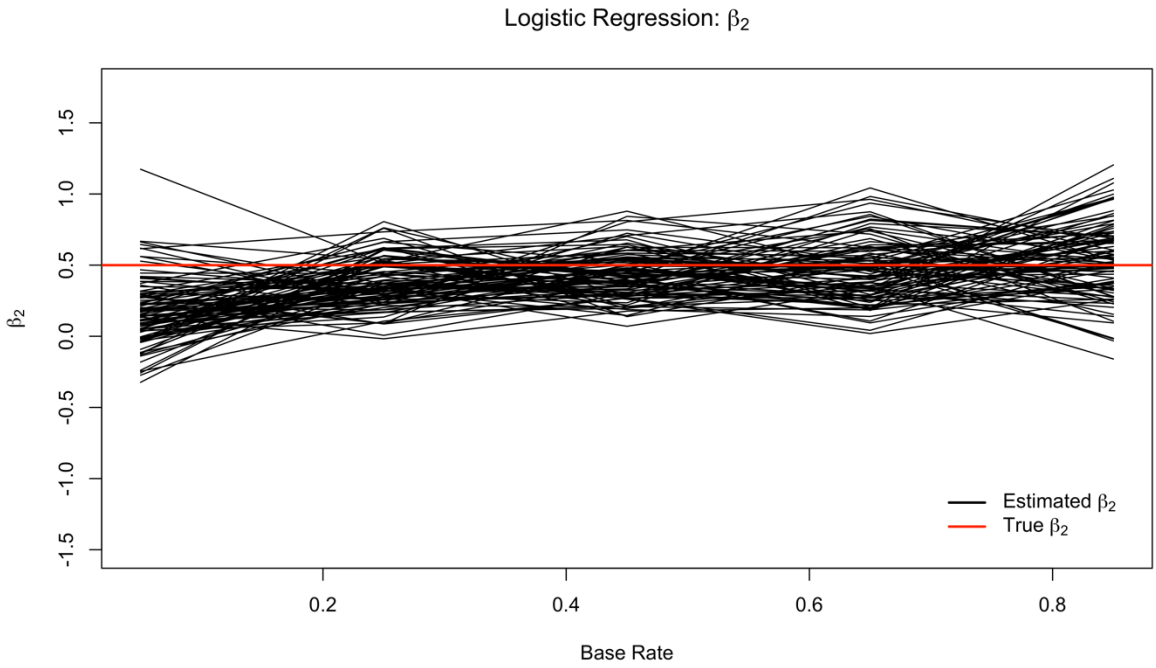

Figure A6: Logistic estimates of β_2 in scenario 1.

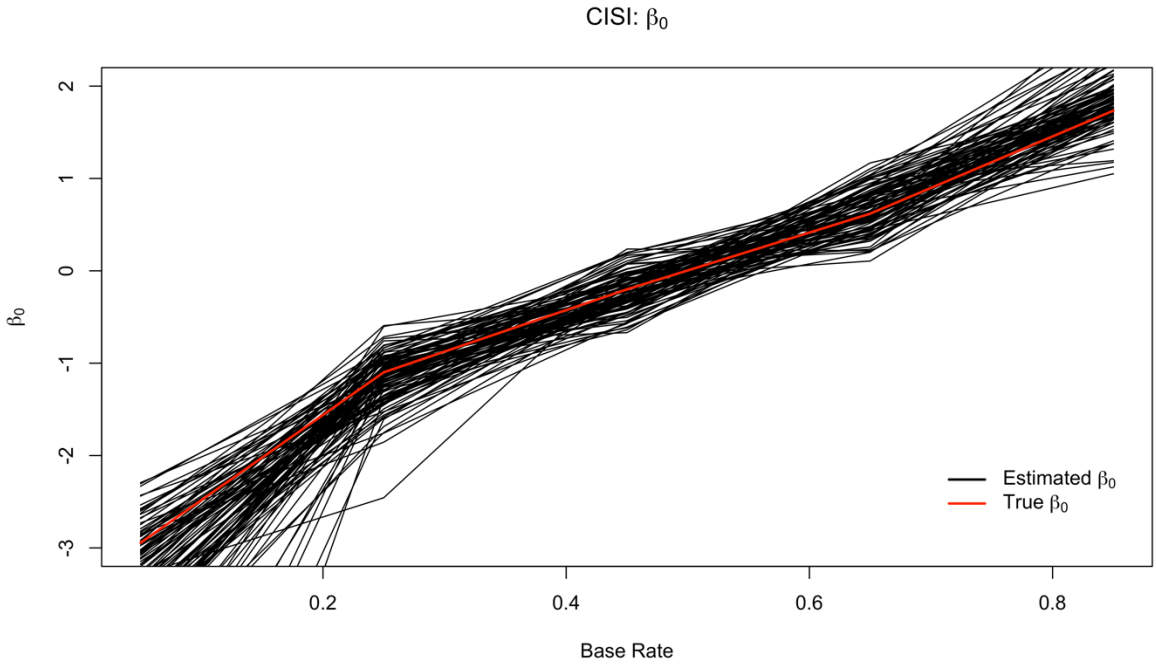

Figure A7: CISI estimates of β_0 in scenario 2.

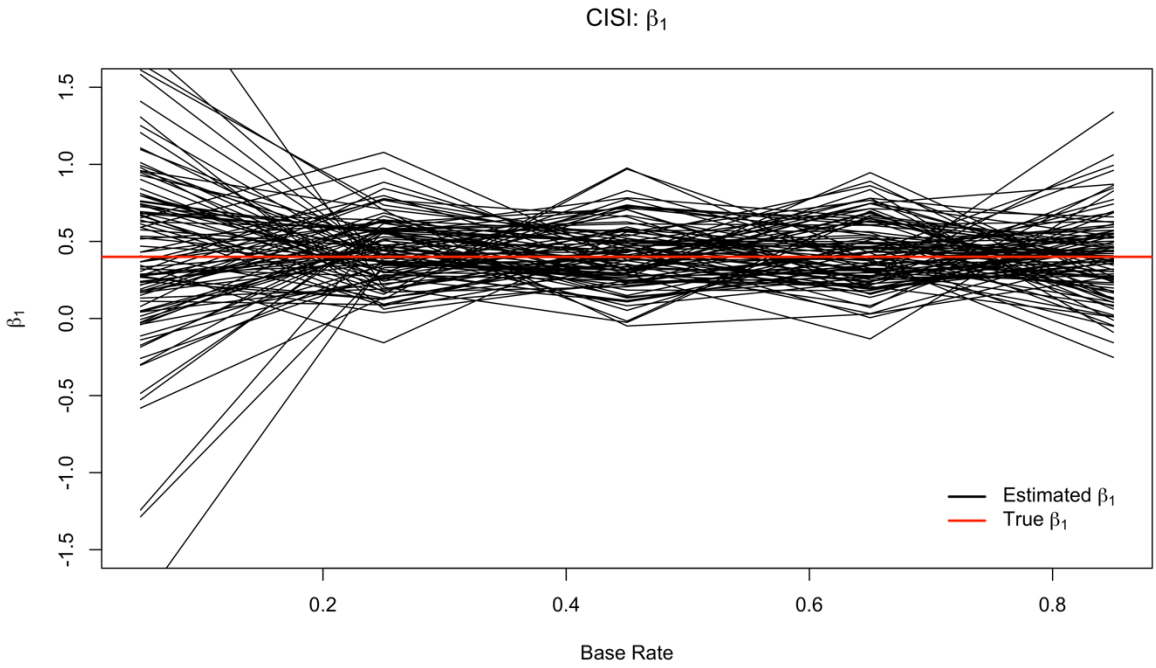

Figure A8: CISI estimates of β_1 in scenario 2.

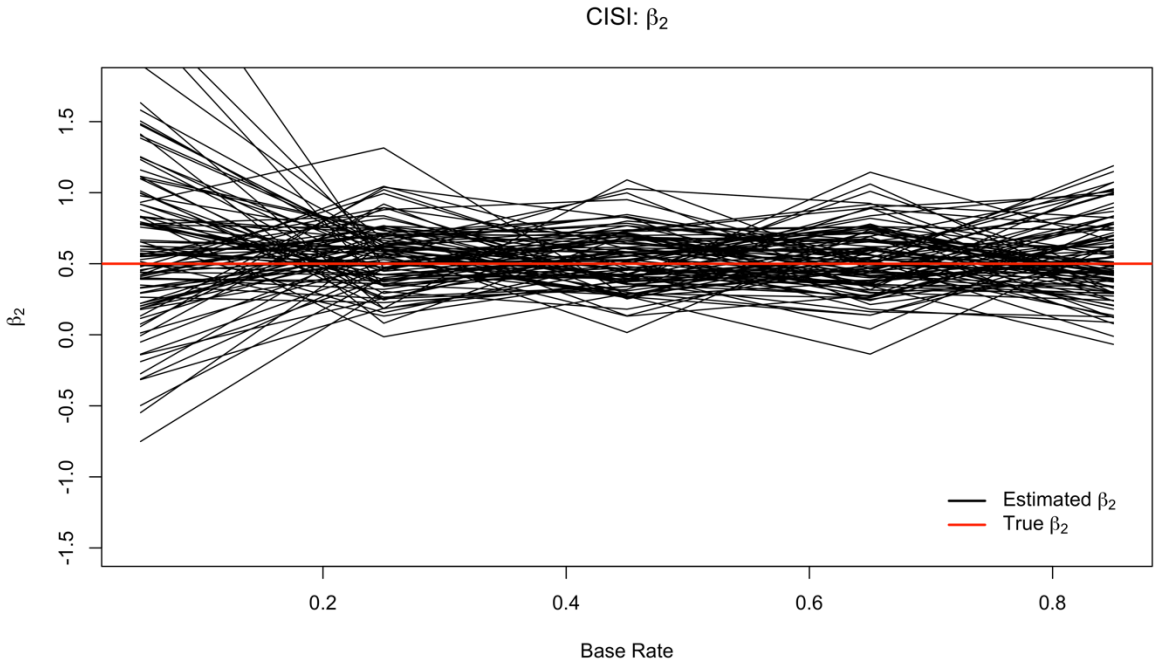

Figure A9: CISI estimates of β_2 in scenario 2.

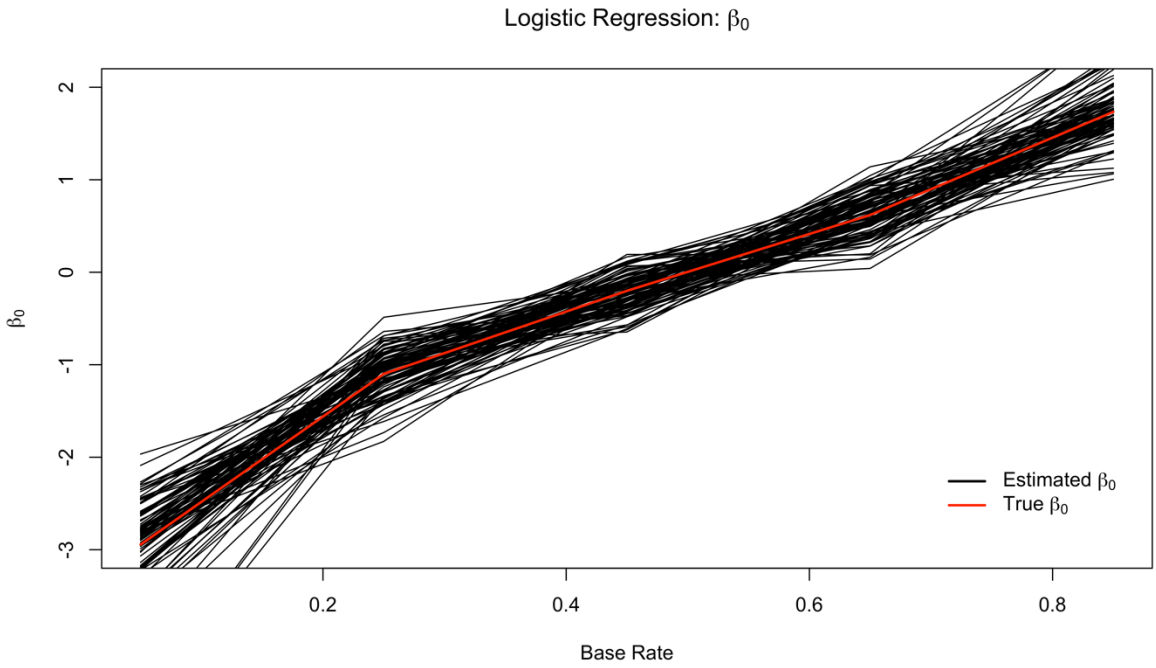

Figure A10: Logistic estimates of β_0 in scenario 2.

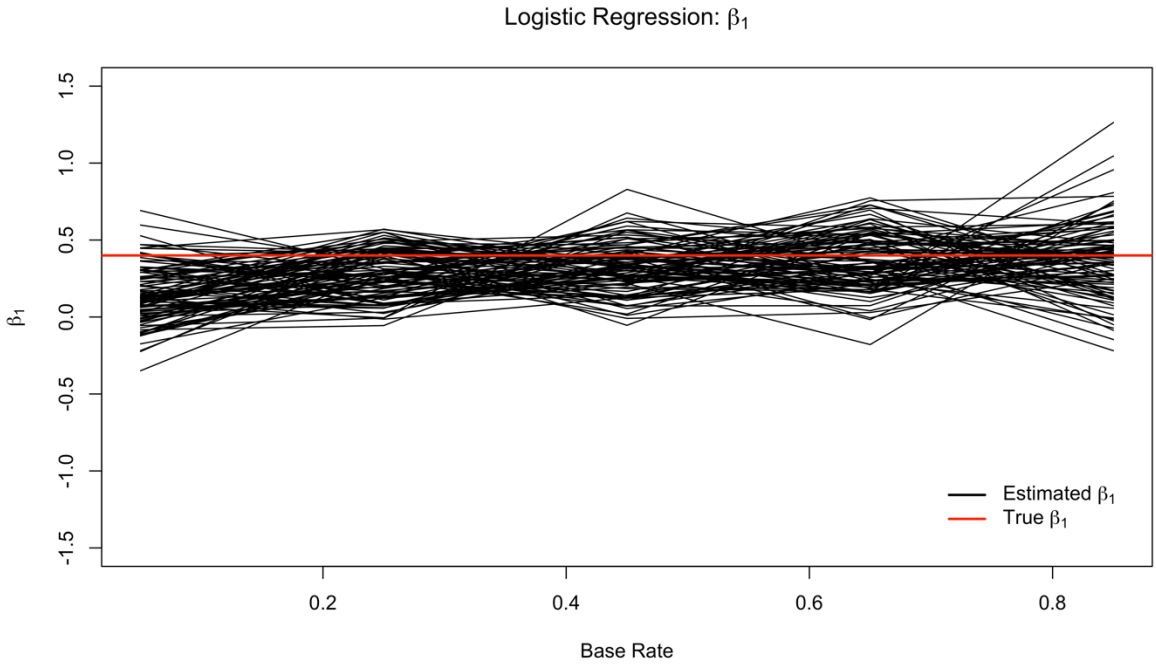

Figure A11: Logistic estimates of β_1 in scenario 2.

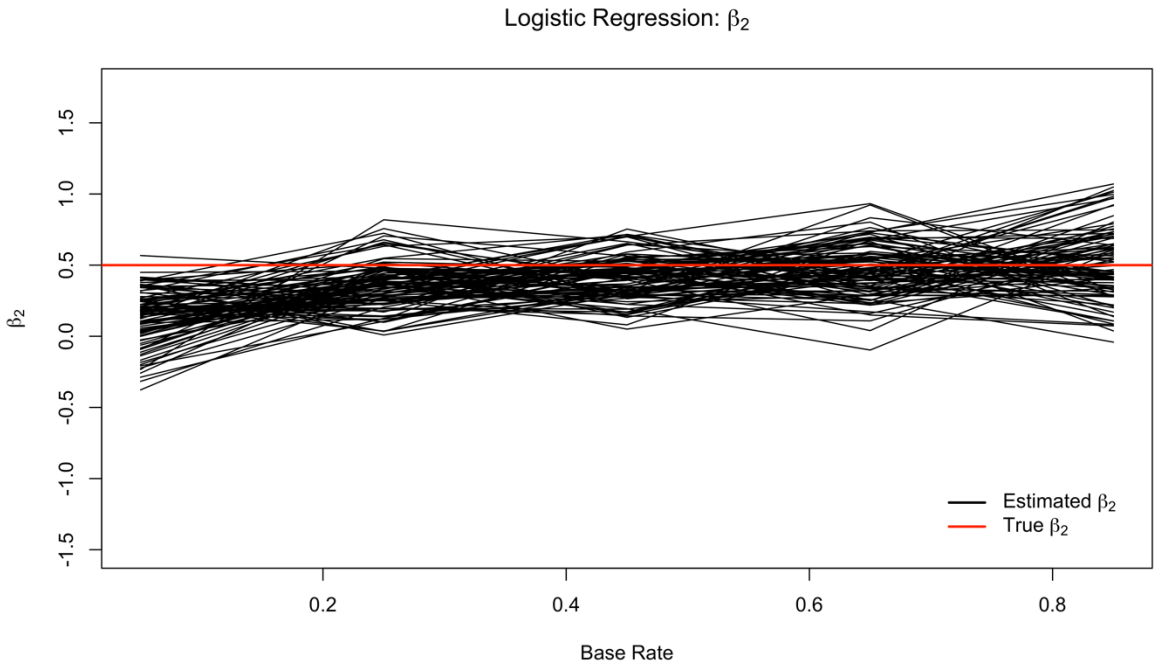

Figure A12: Logistic estimates of β_2 in scenario 2.

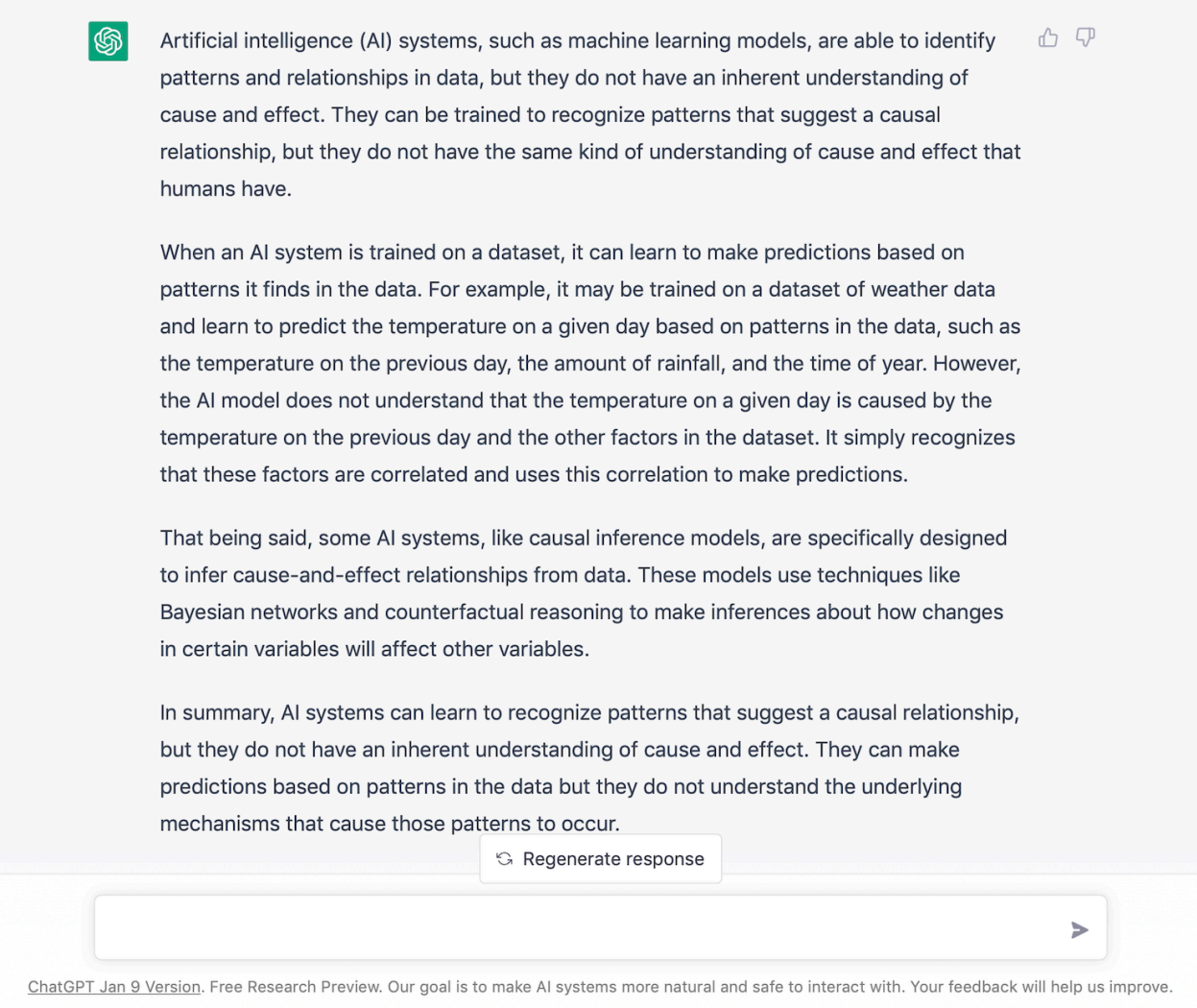

Figure A13: ChatGPT's response to the question, "Does AI have a representation of cause and effect?"

Table A1. GPT-3 causal structure responses across varying temperatures. Red text indicates incorrect answer.

Temp = 0	Temp = 0.25	Temp = 0.50	Temp = 0.75	Temp = 1.0
Medicine A prevents headaches. The results of the study show that 100% of the participants who received medicine A had headaches, while only 50% of the participants who did not receive medicine A had headaches. This indicates that medicine A is effective in preventing headaches.	Medicine A prevents headaches.	Medicine A appears to prevent headaches, as 100% of the participants who received the medicine did not experience headaches, while 50% of the participants who did not receive the medicine did experience headaches.	Medicine A prevents headaches. The fact that 100% of the participants in the treatment group had headaches compared to 50% in the control group indicates that the medicine is preventing headaches.	Medicine A prevents headaches.
Based on the data presented, it appears that medicine B prevents headaches. The percentage of participants who received medicine B and experienced headaches (83%) is significantly lower than the percentage of participants who did not receive medicine B and experienced headaches (33%). This suggests that medicine B is effective in preventing headaches.	Based on the data presented, it appears that medicine B prevents headaches. The percentage of participants who had headaches was significantly lower in the treatment group (83%) than in the control group (33%). This suggests that medicine B is effective in preventing headaches.	Medicine B prevents headaches. The data suggests that the participants who received medicine B had a much lower rate of headaches than the participants who did not receive medicine B. This indicates that medicine B is effective in preventing headaches.	Based on the data presented, Medicine B prevents headaches.	Medicine B appears to prevent headaches. The data shows that 83% of the participants in the treatment group have headaches, while only 33% of the participants in the control group have headaches. This suggests that medicine B is effective at reducing the number of participants who experience headache symptoms.
Medicine C prevents headaches.	Medicine C prevents headaches.	Medicine C prevents headaches.	Medicine C prevents headaches.	Medicine C appears to prevent headaches. The data shows that 75% of participants in the experimental group experienced headaches, while 0% of participants in the control group experienced headaches. This suggests that the medicine is effective in preventing headaches.
Medicine D prevents headaches.	Medicine D prevents headaches.	Medicine D prevents headaches.	Medicine D prevents headaches.	It appears that medicine D prevents headaches, as there was a significantly higher percentage of participants in the control group who had headaches compared to the treatment group.
Medicine E prevents headaches. The data shows that the participants who received medicine E had a lower percentage of headaches than those who did not receive medicine E. This suggests that medicine E is effective in preventing headaches.	Medicine E prevents headaches. The percentage of participants with headaches in the treatment group (17%) is significantly lower than the percentage of participants with headaches in the control group (67%). This suggests that medicine E is effective in preventing headaches.	Medicine E appears to prevent headaches. The percentage of participants who received medicine E and had headaches (17%) is significantly lower than the percentage of participants who did not receive medicine E and had headaches (67%). This suggests that medicine E has a protective effect against headaches.	Medicine E prevents headaches. The results show that the group that took medicine E had a lower proportion of participants with headaches than the group that did not take medicine E. This suggests that medicine E is having a preventative effect on headaches.	Medicine E appears to prevent headaches. The percentage of participants with headaches was significantly lower in the treatment (medicine E) group than in the control group. This suggests that medicine E has an effect on headaches and reduces the likelihood of experiencing them.
Medicine F prevents headaches.	Medicine F prevents headaches.	Medicine F prevents headaches.	Based on the data presented, it appears that medicine F prevents headaches. 25% of the participants who received medicine F did not have headaches, while 100% of the participants who did not receive medicine F did have headaches.	Based on this data, it appears that Medicine F prevents headaches. Since a higher percentage of the participants who received Medicine F did not have headaches compared to the participants who did not receive Medicine F, this suggests that Medicine F is effective at reducing the occurrence of headaches.
Medicine G prevents headaches.	Medicine G has no effect on headaches.	Medicine G prevents headaches. The data indicates that 50% of the participants in the treatment group (those who received the medicine) had headaches, while 100% of the participants in the control group (those who did not receive the medicine) had headaches. This suggests that the medicine is having a protective effect, preventing headaches in some of the participants.	Medicine G appears to have no effect on headaches, as the same percentage of people in the treatment and control groups experienced headaches.	Medicine G prevents headaches. This is evidenced by the fact that the group receiving medicine G had only 50% of participants experiencing headaches, while the group without medicine G had 100% of participants experiencing headaches.
Medicine H has no effect on headaches.	Medicine H has no effect on headaches.	Medicine H has no effect on headaches.	Medicine H has no effect on headaches.	Medicine H appears to have no effect on headaches since both the experimental group and the control group experienced the same percentage of headaches.
Medicine I has no effect on headaches.	Medicine I has no effect on headaches.	Based on the data presented, it appears that medicine I has no effect on headaches.	Medicine I has no effect on headaches.	Medicine I has no effect on headaches.
Medicine J prevents headaches.	Medicine J prevents headaches.	Medicine J prevents headaches.	Medicine J prevents headaches.	Medicine J prevents headaches.

Table A2. GPT-3 causal strength responses across varying temperatures.

Temp = 0	Temp = 0.25	Temp = 0.50	Temp = 0.75	Temp = 1.0
The answer is 0 out of 100 people, all of whom do not have headaches, would have a headache if given the medicine. This is because 100% of the participants who received medicine A (those in the treatment group) already had headaches, and 50% of the participants who did not receive medicine A (those in the control group) also had headaches. Therefore, it can be concluded that medicine A causes headaches.	The answer is 0 out of 100 people. Since 100% of the participants who received medicine A had headaches, and 50% of the participants who did not receive medicine A had headaches, it is clear that the medicine has a side-effect on headaches and causes them. Therefore, if 100 people who do not have headaches were given the medicine, all of them would develop headaches.	The answer is 0 out of 100 people, all of whom do not have headaches, would have a headache if given the medicine. This is because 100% of the participants who received medicine A (those in the treatment group) already had headaches, and the control group (those who did not receive medicine A) had 50% with headaches. Therefore, this medicine has a side-effect that causes headaches.	It is impossible to answer this question without further information. The data from this study only indicates that 100% of participants who received medicine A had headaches, and that 50% of participants who did not receive medicine A had headaches. We cannot infer from this study how many out of 100 people, all of whom do not have headaches, would have a headache if given the medicine.	100% of the patients who received the medicine A would have headaches, so all 100 people in the group not receiving the medicine would also have headaches if given the medicine.
Based on the data presented, it appears that medicine B has a side-effect on headaches, and that it causes them. Therefore, if 100 people who do not have headaches were given the medicine, 83 out of 100 would likely experience a headache.	Based on the data presented, it appears that medicine B has a side-effect on headaches, and that it causes them. Therefore, if 100 people who do not have headaches were given the medicine, it is likely that 83 out of the 100 would develop a headache.	Based on the data, it appears that medicine B has a side effect on headaches. It appears to cause headaches, as 83% of participants who received the medicine experienced headaches, while only 33% of participants who did not receive the medicine experienced headaches. Therefore, if 100 people who do not have headaches were given the medicine, it is likely that 83 of them would experience headaches.	In this case, 83 out of 100 people would have a headache if given the medicine.	Based on the data presented, it appears that taking medicine B may have a side-effect on headaches as there is a higher percentage of participants in the treatment group who have headaches compared to the control group. Therefore, if 100 people who do not have headaches are given medicine B, it is likely that 83 out of the 100 people would have a headache due to the medicine.
100% of the people who do not have headaches would have a headache if given the medicine.	Based on the data presented, it appears that medicine C has a side-effect on headaches, and that it causes them. Therefore, if 100 people who do not have headaches were given the medicine, 75 out of those 100 people would have a headache.	Based on the data provided, it appears that medicine C has a side-effect on headaches, and that it causes them. Therefore, if 100 people who do not have headaches are given medicine C, 75 out of those 100 people would have a headache.	75 out of 100 people would have a headache if given the medicine.	75 out of 100 people would have a headache if given the medicine.
The answer is 50 out of 100 people.	The answer is 50 out of 100 people.	The answer is 50 out of 100 people who do not have headaches would have a headache if given the medicine.	The answer is 50. If 100 people who do not have headaches were given medicine D, then 50 of them would end up with headaches.	The answer is 50. Half (50%) of the participants who received medicine D experienced headaches. Therefore, if you gave medicine D to 100 people who did not have headaches originally, 50 out of those 100 people would now have a headache. This indicates that the medicine does cause headaches as a side-effect.
The answer is 50 out of 100 people. This is because 17% of the participants who received medicine E (those in the treatment group) have headaches, while 67% of the participants who did not receive medicine E (those in the control group) have headaches. Therefore, 50 out of 100 people would not have a headache if given the medicine.	The answer is 50 out of 100 people. This is because 17% of the participants who received medicine E (those in the treatment group) have headaches, and 67% of the participants who did not receive medicine E (those in the control group) have headaches as well. Therefore, 17% of 100 people (17 people) would have a headache if given the medicine, and 67% of 100 people (67 people) would have a headache if not given the medicine. This means that	The answer is 50 out of 100 people. This is calculated by taking the difference between the two groups (67% - 17%) and multiplying it by 100. Thus, 50 out of 100 people would not have a headache if given the medicine.	The answer would be 67 out of 100 people with headaches would not have a headache if given the medicine.	17 out of 100 people would not have a headache if given the medicine.
25 out of 100 people would not have a headache if given the medicine.	25 out of 100 people would not have a headache if given the medicine.	The answer is 25 out of 100 people. This is because 25% of the participants who received medicine F did not have headaches, which means that 25 out of 100 people would not have a headache if given the medicine.	25 out of 100 people would not have a headache if given the medicine.	25 out of 100 people would not have a headache if given the medicine.
The answer is 50 out of 100 people would not have a headache if given the medicine.	The answer is 0 out of 100 people. Since the study found that 50% of the participants who received medicine G had headaches, and 100% of the participants who did not receive medicine G had headaches, it is clear that the medicine does not prevent headaches. Therefore, none of the 100 people would not have a headache if given the medicine.	The answer is 50 out of 100 people would not have a headache if given the medicine.	50 out of 100 people would not have a headache if given the medicine.	The answer is 50. 50 out of 100 people in the study would not have a headache if given the medicine.
The answer is 33 out of 100 people.	The answer is 33 out of 100 people.	The answer is 33 out of 100 people, all of whom do not have headaches, would have a headache if given the medicine.	This experiment does not provide sufficient information to answer the question. The data provided only tells us the number of participants in the experimental and control groups who experienced headaches. It does not tell us how many people who did not have headaches would have a headache after taking the medicine.	The answer is 33 out of 100 people.
The answer is 0 out of 100 people, all of whom do not have headaches, would have a headache if given the medicine. This is because the study found that the medicine had no influence on headaches, as the same percentage of participants in both the treatment and control groups had headaches.	The answer is 0 out of 100 people, all of whom do not have headaches, would have a headache if given the medicine. This is because the study found that the medicine had no influence on headaches.	100 out of 100 people would not have a headache if given the medicine, since the study found that 67% of both the treatment and control groups had headaches.	Since the study found that 67% of the participants who received medicine I have headaches and 67% of the participants who did not receive medicine I have headaches as well, it is likely that the medicine has no influence on headaches. Therefore, if 100 people who do not have headaches were given the medicine, it is likely that they would all still remain without headaches.	It is impossible to say how many out of 100 people would have a headache if given medicine I, as the study only provides us with the results of the experiment which show that 67% of the participants, who were already suffering from headaches, in both the experimental and control groups, experienced the same results after taking the medicine.
The answer is 0 out of 100 people. Since 0% of the participants who received medicine J had headaches, it can be concluded that the medicine does not have any effect on headaches, and therefore would not prevent any of the 100 people from having a headache.	The answer is 0 out of 100 people. Since 0% of the participants who received medicine J had headaches, it can be concluded that the medicine does not have an effect on headaches, and therefore would not prevent any of the 100 people from having a headache.	0 out of 100 people, all of whom have headaches, would not have a headache if given the medicine.	ANSWER: 0	The answer is 0 out of 100 people, as none of the participants in the treatment group (the group that received medicine J) had headaches.

Table A3. ChatGPT-4 causal strength responses.

Medicine	Response
A	Based on the information provided, we can conclude that medicine A has a side effect of causing headaches. In the study, 100% of the participants who received medicine A experienced headaches, whereas only 50% of the participants in the control group (who did not receive the medicine) experienced headaches. If we are given 100 people who do not have headaches, and they are given medicine A, we can expect all of them (100 out of 100) to experience headaches, based on the study's results.
B	Based on the study results, 83% of participants who received medicine B experienced headaches. Since the question asks for the expected number of people out of 100 who would have a headache if given the medicine, we can simply apply the percentage to this number: 83% of 100 people = 0.83 * 100 = 83 people So, out of 100 people who do not have headaches, we would expect 83 of them to develop a headache if given medicine B.
C	Based on the information provided in the study, 75% of the participants who received medicine C experienced headaches. Therefore, if 100 people who do not have headaches are given medicine C, we would expect 75 of them to develop headaches as a side effect.
D	Based on the data from the study, 50% of the participants who received medicine D experienced headaches. Therefore, if 100 people who do not have headaches are given the medicine, we would expect 50 out of the 100 people to experience headaches as a side effect of the medicine.
E	Based on the given data, 17% of participants who received medicine E had headaches, while 67% of those who did not receive medicine E had headaches. The information provided allows us to determine the percentage decrease in headaches for those taking medicine E compared to those who did not take it. To find the percentage decrease in headaches, we can use the following formula: Percentage decrease = [(control group headache rate - treatment group headache rate) / control group headache rate] x 100 Percentage decrease = [(67% - 17%) / 67%] x 100 Percentage decrease ≈ 74.63% If we consider 100 people who initially have headaches, and all of them take medicine E, we would expect approximately 74.63% of them to not have a headache after taking the medicine, based on the data provided. Therefore, out of 100 people who initially have headaches, we would expect around 75 (rounding to the nearest whole number) of them to not have a headache after taking medicine E.
F	To answer this question, we need to consider the effectiveness of medicine F in reducing the occurrence of headaches. We know that 25% of participants who received medicine F experienced headaches, which means that 75% of those who received medicine F did not have headaches. In the control group, 100% of the participants had headaches. Now, let's assume we have 100 people who all have headaches. If we were to give them medicine F, based on the study results, we would expect 75% of them to no longer have headaches. To calculate this, we can multiply the total number of people (100) by the percentage of people who didn't have headaches after receiving medicine F (75%): 100 people * 0.75 (75%) = 75 people Therefore, out of 100 people who all have headaches, we would expect 75 of them to no longer have headaches if they were given medicine F.
G	Based on the study results, 50% of participants who received medicine G had headaches, while 100% of participants in the control group had headaches. This means that medicine G reduced the incidence of headaches by 50% compared to the control group. If we consider a group of 100 people who all have headaches and give them medicine G, we would expect 50% of them (50 people) to no longer experience headaches due to the effectiveness of the medicine.
H	Based on the information provided, it appears that medicine H does not have an influence on headaches. The study found that 33% of the participants who received medicine H (treatment group) had headaches, while 33% of the participants who did not receive medicine H (control group) also had headaches. The percentage of people experiencing headaches is the same in both groups, suggesting that medicine H neither causes nor prevents headaches. Therefore, if 100 people without headaches were given medicine H, we would expect 33 of them to develop headaches, which is the same rate as in the control group.
I	Based on the information provided, medicine I does not seem to have any influence on headaches. In both the treatment group and the control group, 67% of the participants experienced headaches. This means that the medicine neither causes nor prevents headaches, as the occurrence of headaches is the same for both groups. To answer your question: If 100 people without headaches were given the medicine, we would expect 67 of them to develop a headache, as the medicine does not seem to have any effect on the incidence of headaches.
J	Based on the information provided, 0% of the participants who received medicine J (treatment group) had headaches. Conversely, 50% of the participants who did not receive medicine J (control group) had headaches. Given that participants were randomly assigned to the two groups, it can be inferred that medicine J is effective in preventing headaches as a side effect. If there are 100 people with headaches, and they are given medicine J, we would expect 0% of them to have a headache after receiving the medicine, since medicine J has been shown to prevent headaches in the experimental group. Therefore, all 100 people who initially had headaches would be expected to not have a headache after being given medicine J.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.*, 120, e2218523120.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, *311*(5763), 1020-1022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Scott, L., Nori, H., Hamid, P., Ribeiro, T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of experimental psychology: learning, memory, and cognition*, *29*(6), 1119.

Bye, J. K.**,** Chuang, P.-J., & Cheng, P. W. (2023). How do humans want causes to combine their effects? The role of analytically-defined causal invariance for generalizable causal knowledge. *Cognition*, 230, 105303. https://doi.org/10.1016/j.cognition.2022.105303

Centers for Disease Control and Prevention (CDC. (2012). *Principles of epidemiology in public health practice: An introduction to applied epidemiology and biostatistics* (3rd ed.) https://www.cdc.gov/csels/dsepd/ss1978/SS1978.pdf.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Cheng, P.W. & Lu, H. (2017). Causal Invariance as an Essential Constraint for Creating a Causal Representation of the World: Generalizing the Invariance of Causal Power. In M.R. Waldmann (Ed). *The Oxford Handbook of Causal Reasoning* (pp. 65-84). Oxford, England: Oxford University Press.

Cheng, P.W., Novick, L.R., Liljeholm, M. & Ford, C. (2007). In M. O'Rourke (Ed.), Topics in Contemporary Philosophy (Volume *4,* pp. 1 – 32): *Explanation and Causation*. Cambridge, MA: MIT Press.

Cheng, P.W., Sandhofer, C.M., Liljeholm, M. (2022). Analytic causal knowledge for constructing useable empirical causal knowledge: Two experiments on preschoolers. *Cognitive Science, 46*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, *1369*(1), 132-153.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Gal'perin, P. I. (1992). Linguistic consciousness and some questions of the relationship between language and thought. *Journal of Russian & East European Psychology*, *30*(4), 81-92.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334-384.

Hawking, S., & Mlodinow, L. (2010). *The grand design*. New York: Bantam Books.

Hoffman, D. (2015, June). Do we see reality as it is? [Video]. TED Conferences. https://www.ted.com/talks/donald_hoffman_do_we_see_reality_as_it_is

Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. New York: W.W. Norton & Company.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, *1988*(1), i-50.

Hume, D. (1739/1987). *A treatise of human nature* (2nd edition, Clarendon Press, Oxford).

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, *104*(3), 427.

Huxley, T. H. (1890). *Lessons in elementary physiology*. Macmillan and Company.

Ichien, N., & Cheng, P. W. (2022). Revisiting Hume in the 21st century: The possibility of generalizable causal beliefs given inherently unobservable causal relations. *Advances in experimental philosophy of causation*, 7-34.

Kant, I. (1781/1965). *Critique of pure reason*. Macmillan, London.

Keys, A. (1980). *Seven countries: a multivariate analysis of death and coronary heart disease*. Harvard University Press.

Khetan, V., Ramnani, R., Anand, M., Sengupta, S., & Fano, A. E. (2022). Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1* (pp. 965-980). Springer International Publishing.

Liljeholm, M. & Cheng, P.W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science*, *18*, 1014-1021.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Mandler, J. M. (2004). Thought before language. *Trends in cognitive sciences*, *8*(11), 508-513.

Marr, D. (1982). *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Co., San Francisco, CA.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

OpenAI. (2023). GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Pearl, J. (2022). Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 317-372).

Park, J., McGillivray, S., Bye, J. K., & Cheng, P. W. (2022). Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions. *Cognitive Psychology*, *132*, 101432.

Pearl, J. (2022). Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, *19*(2).

Pearl, J. (2001). Bayesian networks, causal inference and knowledge discovery. *UCLA Cognitive Systems Laboratory, Technical Report*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision research*, *41*(24), 3145-3161.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 1-55.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Sheps, M. C. (1958). Shall we count the living or the dead?. *New England Journal of Medicine*, *259*(25), 1210-1214.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., a Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1-4), 287-313.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Waldmann, M. R. (2017). Causal reasoning: An introduction. *The Oxford handbook of Causal Reasoning*, 1-9.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*(6), 307-311.

Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal of the Philosophy of Science, 51*, 197–254.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review, 115*, 1–50.

Yu, B., Li, Y., & Wang, J. (2019, November). Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4664-4674).

Yuille, A. L., & Lu, H. (2008). The noisy-logical distribution and its application to causal inference. *Advances in neural information processing system*s, Vol. 20. Cambridge, MA: MIT Press.