Open Access Library Journal
2024, Volume 11, e11817
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Causal Inference, Fast and Slow

Sergio Da Silva

Department of Economics, Federal University of Santa Catarina, Florianopolis, Brazil
Email: professorsergiodasilva@gmail.com

How to cite this paper: Da Silva, S. (2024) Causal Inference, Fast and Slow. *Open Access Library Journal*, **11**: e11817.
https://doi.org/10.4236/oalib.1111817

Received: June 12, 2024
Accepted: July 27, 2024
Published: July 30, 2024

Abstract

This paper explores causal inference through cognitive psychology, focusing on the dual-processing theory of the mind, which includes fast (System 1) and slow (System 2) thinking. It explains that our fast thinking, geared towards identifying causes, helps us survive but can also lead to incorrect causal inferences. The paper underscores the need for slow, deliberate thinking in accurately determining cause-and-effect, a challenging but essential approach. It outlines established methods for developing precise causal inference frameworks and highlights the need for a balanced approach in research, utilizing both systems for creating effective causal diagrams. It proposes using the "thinking, fast and slow" concept to combine System 1's intuitive reasoning with System 2's thorough causal analysis, improving causal inference in everyday research.

Subject Areas

Artificial Intelligence, Psychology

Keywords

Causal Inference, Causal Diagrams, Causal Networks, Statistical Inference, Data-Driven Decision Making, Dual-Processing Theory of the Mind

1. Introduction

This paper reviews The Book of Why by Judea Pearl and Dana Mackenzie [1], analyzing it from a cognitive psychology standpoint, specifically through the lens of the dual-processing theory of the mind [2]. It is a meta-review, examining the material from the perspective of how our minds process information. According to Simonyi's law, named after the lead developer of Microsoft Word, "everything that can be done can be done meta." The paper examines how our fast thinking (or System 1 thinking), evolved in the Paleolithic for identifying causes to ensure survival, frequently leads to incorrect causal inferences in today's world. There-

fore, for accurate cause-and-effect analysis, slow, deliberate thinking (or System 2 thinking) is essential. However, implementing this approach in practice is challenging.

Evidence-based medicine, which favors data-driven decisions over value-based ones, has outpaced eminence-based medicine. An empirical approach is crucial, though it does not imply that everything is solely contained in the data. However, in our mind's automatic mode, we are susceptible to type I errors or false positives, perceiving patterns in randomness. This phenomenon, known as patternicity, often leads to attributing these random patterns to supernatural causes (agenticity). Hence, it is wise to combine empiricism with skepticism, necessitating slow, deliberate thinking. Nassim Taleb [3] advocates for empirical skepticism in The Black Swan, but he does not discuss its cognitive demands, a topic we explore here.

Slow thinking can also lead to errors, such as type II errors or false negatives, by missing existing patterns. Scientists, who primarily use System 2 thinking, are less likely to be deceived by randomness due to their empirical skepticism. However, they remain susceptible to type II errors. Statistics, an essential tool in modern science, exemplifies this. Statistics show correlations, not causation, which can sometimes result in type II errors. Causation only occurs in statistics when two variables have a deterministic relationship, indicated by a correlation coefficient of either 1 or −1.

A type I error, or a false positive, occurs when a true null hypothesis is incorrectly rejected [4]. It is an error of commission, involving the detection of non-existent patterns. Conversely, a type II error, or a false negative, happens when a false null hypothesis is not rejected [4]. This is an error of omission, where actual patterns are overlooked.

Cognitive psychologists recognize the two mental processes, popularized by Daniel Kahneman as System 1 and System 2 [5] [6]. System 1, older in evolutionary terms, consists of autonomous subsystems and is responsible for domain-specific processing. System 2, in contrast, allows for abstract reasoning and hypothesis use, serving as a domain-general processing mechanism. This distinction highlights evolutionary rationality (System 1 logic) versus individual rationality (System 2 logic). The late development of System 2 enables humans to pursue personal goals, beyond genetic objectives, leading to what is termed the "carbon robot revolution" [7]. This revolution signifies the possibility of overcoming the constraints of natural and sexual selection.

Most evolutionary psychologists reject the concept of a domain-general processing mechanism (System 2) [8] and adhere to the modularity of the mind hypothesis (System 1) [9]. A few cognitive psychologists agree, suggesting that both intuitive and deliberate judgments operate on common principles [2]. Although the majority of evolutionary psychologists dispute the idea of a general-purpose, content-free cognitive architecture, a growing number are starting to acknowledge the two-minds theory [10].

Evolutionary psychologists contend that although specialized adaptations

arise from recurring adaptive problems, humans also encountered novel problems too infrequent for specific adaptations to evolve. They caution against prematurely assuming the existence of a domain-general processing mechanism alongside proven domain-specific mechanisms. The domain-specific mind theory has effectively revealed key mechanisms, and it is yet to be determined if the domain-general mind theory will produce similar empirical findings [10].

The human mind's mechanisms are interdependent, as they share data among each other [10]. For instance, internal data like sight, smell, and hunger combine to assess food's edibility. This lack of information encapsulation in psychological mechanisms contradicts the idea of modularity [10]. Encapsulation would mean each mechanism operates on isolated information, not integrating data from other sources. Additionally, the existence of super mechanisms (or daemons) is suggested, which specialize in coordinating and regulating these interconnected mechanisms.

System 2 strategizes actions to maximize utility based on individual goals, while System 1 focuses on maximizing inclusive fitness from a genetic standpoint. In scenarios beyond evolutionary adaptations, analytical processing by System 2 is required to override System 1 [7]. The conflict between these systems leads to numerous cognitive biases, as explored in the heuristics and biases research of Kahneman and Amos Tversky. These biases hinder an individual's ability to effectively maximize utility.

Cognitive psychologist Keith Stanovich argues that evolutionary psychologists are mistaken in believing that System 1 heuristics, developed during the Pleistocene (or Paleolithic, as some prefer for its specific relation to human prehistory), are still effective for decision-making in today's world. Consequently, it is necessary to depend on System 2 for making logical and probabilistic inferences with different rules. Additionally, we need to sift through the substantial information from our independent modules (System 1) that might impede sound decision-making.

The dual-processing mind, adapted for evolution, developed System 1 in the Paleolithic. At that time, the survival benefit of making type I errors (false positives) outweighed the cost. For example, mistaking a bush's movement for a lion rather than wind increased survival chances, leading us to naturally perceive causation. Modern science needs to refine causal inference, moving beyond the statistical tendency to reject causation. It should recognize and incorporate the inherent survival value of causality identified by fast thinking, and then systematically apply it through slow, deliberate thinking. We hypothesize that during the development of System 2, approximately 50,000 years ago [2], humans adapted their pre-existing brain functions for automatic pattern recognition to also perform deliberate causal reasoning. In Reference [2], we offer a thorough discussion of the cognitive theories presented here and document their empirical support. Next, we explore a comprehensive script for reliable causal inference.

2. The Three-Rung Ladder of Causation

Slow thinking in causal inference should go beyond the quick causality judg-

ments of System 1. This is crucial because System 1's tendency to jump to conclusions leads to errors, especially in contexts far removed from its evolutionary origins. Fast, empirical thinking assumes "what you see is all there is," ignoring silent evidence. Cicero's reference to Diagoras of Melos illustrates this. Diagoras, shown paintings of people saved from shipwrecks by the gods, pointed out the absence of paintings of those who drowned. This concept of silent evidence is extensively discussed in Chapter 8 of The Black Swan by Taleb [3].

Intuition often misses the concept of regression to the mean, leading to non-regressive judgments by System 1, as identified by Francis Galton. Typically, a good performance is followed by a worse one, and vice versa, due to chance rather than changes in ability. Since people naturally think in terms of cause and effect, they fail to recognize this statistical phenomenon. As a result, praising a good performance can mistakenly be seen as causing a subsequent poor performance. Critics of poor performance often incorrectly assume that their criticism causes improvement. This can lead to the false belief that criticism is effective while praise is not. The subsequent performance improvement is actually due to regression to the mean, a random occurrence, not the criticism. There is a strong correlation between praise and poor performance, and criticism and good performance. However, this correlation does not mean one causes the other. Regression to the mean is causeless. Kahneman explores this in Chapter 17 of Thinking, Fast and Slow [6].

To infer causes correctly, we must use System 2 reasoning to follow a detailed process, akin to climbing a three-rung causation ladder (as depicted in **Figure 1**). Successfully doing so could lead to programming this process into machines, potentially achieving "strong artificial intelligence" (human-level intelligence automation). Pearl developed this concept of a structured approach to understanding causation [1].

Figure 1. The three-rung ladder of causation. The first rung of the causation ladder focuses on association, involving observations and queries like "What if I see…". The second rung addresses intervention, encompassing actions and questions such as "What if I do…". The third rung covers counterfactuals, characterized by imagination, reflection, and comprehension, asking "What if I had done…".

Causal learning encompasses three key cognitive skills: observation, action, and causal imagination. Observation is about noticing patterns in our surroundings. Action involves forecasting the outcomes of changes in the environment and selecting actions to achieve specific goals. Causal imagination, or counterfactual thinking, involves grasping the reasons behind events by picturing their possible outcomes.

A successful mammoth-hunting tribe during the Paleolithic likely followed a three-step causation process, as outlined in **Figure 1**. This suggests that their automatic System 1 was naturally able to navigate this process. It implies that specific cognitive modules evolved to facilitate this task. In the first section of Chapter 1 of The Book of Why, Pearl and Mackenzie give the mammoth hunting example and contend that awareness of the causation ladder only emerged 50,000 years ago during the "Cognitive Revolution." We note that this coincides with the theorized development of System 2 [2]. Therefore, Pearl's standpoint does not consider the long history of mammoth hunting since the early Paleolithic, indicating that System 1 played the most significant role in the Cognitive Revolution. Pearl identifies the Lion Man of Stadel Cave, a 40,000-year-old sculpture of a half-man, half-lion made from mammoth tusk, as a key indicator of the Cognitive Revolution's start, highlighting the use of counterfactual reasoning [1]. However, this sculpture only represents the later emergence of System 2 deliberate thinking [2]. The need to hunt a mammoth for its tusk, essential for creating the sculpture, implies an early use of planning and counterfactual thinking, tied to the basic cognitive abilities of System 1.

System 1's causation leaps evolved for survival, allowing us to automatically climb the first rung of the three-rung causation ladder. This fast thinking, however, can lead to errors. For instance, when A precedes B, we often perceive a causal relationship, as David Hume discussed in A Treatise of Human Nature [11] [12]. Consider a cartoon sequence where Bugs Bunny eats a carrot; System 1 easily infers Bugs caused the action (**Figure 2**). This indicates that System 1 automatically evaluates temporal causality. Daniel Dennett highlights this [12], noting that without this inference ability, understanding cartoons would be challenging. However, System 1 thinking might overlook other possibilities, like Daffy Duck being the eater, showing that automatic causation does not always mean legitimate causation. This limitation of System 1 is particularly significant in our current world, far removed from Paleolithic contexts.

Figure 2. Illustration of automatic causation from sequential observation. © Warner Bros. Both images, taken separately, are in the public domain.

Observing a rooster crow before sunrise could lead to a mistaken causal belief. However, slow thinking understands that the crowing does not cause the sunrise. Even System 1 judgments, which are imaginative, imply this understanding. For example, if you had eaten the rooster the previous day, you could still easily imagine it crowing at dawn, indicating knowledge that its crowing is not causative of the sunrise. This example is given by Pearl and Mackenzie [1]. System 1 comprehends both temporal causation and counterfactuals. Responsibility, blame, regret, and credit naturally arise in the mind, showing that System 1 can effortlessly create counterfactuals. Understanding these concepts requires comparing actual events with hypothetical alternatives.

Hume also explored this idea, focusing on how automatic imagination, or counterfactual fast thinking, arises when not witnessing events in sequence, as discussed in An Enquiry Concerning Human Understanding [12] [13]. Hume recognized that people effortlessly and reliably make counterfactual judgments, quickly and easily. This ability to envision alternate realities stems from shared experiences and a common understanding of the world's causal framework. It is understood that counterfactuals imply causation. Hume defined a counterfactual as, "if the first object had not been, the second had never existed." Translation: B would not have occurred if not for A, or A has caused B. The use of counterfactuals in System 1 thinking can be effective, as seen in the previous example. The challenge is how to algorithmize these automatic counterfactuals using System 2 thinking, especially when climbing the three-rung ladder of causation to reach the third rung, where counterfactuals are key.

3. Bayes

Bayes' rule is a valuable tool for the initial rung of the causation ladder but does not assist in reaching the rung two of the ladder. By applying System 2's slow thinking in the first rung, we can make valid judgments while observing facts. This is effectively achieved through statistical inference using Bayesian methods. In contrast to objective frequentist statistics, subjective Bayesian statistics start with an initial belief and incorporate new evidence to update this belief. Often, in cases of big data, the effect of prior beliefs fades, resulting in a single, objective conclusion.

The discussion of Bayesian statistics must consider the intricacies of prior and posterior distributions. Prior distributions represent initial beliefs before observing any data, while posterior distributions update these beliefs in light of new evidence, reflecting a fundamental aspect of Bayesian analysis that combines prior information with the likelihood of observed data to produce updated probabilities.

Bayesian networks automate reasoning from evidence to hypotheses and from effects to causes. Thomas Bayes' question was: when does a hypothesis shift from being impossible to improbable, probable, or almost certain? He answered this using inverse probability. Knowing the cause makes it easy to estimate the ef-

fect's likelihood. However, determining the cause from the effect is more challenging. Bayes' rule solves this. It allows big data to be fed into Bayesian networks, which implies that induction is essentially the reverse of deduction.

In forward probability, we start with a known cause and calculate the probability of its effect. This natural process aligns with System 1 cognition, designed to perceive causes. However, in situations of inverse probability, where we observe effects and infer their probable causes, System 2 cognition is required. System 1 is not equipped to process information flowing in this noncausal direction. This is more challenging because several potential causes must be considered, demanding more deliberate thought. Bayes developed a System 2 technology for accurately computing these inverse probabilities.

Why do forward probabilities match inverse probabilities in Bayes' rule? Consider Pearl and Mackenzie's [1] example in Chapter 3, where 2/3 of 12 customers order tea, and half of tea drinkers also get scones, resulting in 1/3 (= 2/3 × 1/2) ordering both. We can reverse the analysis since data disregards cause-effect distinctions. This leads to 5/12 ordering scones, and 4/5 of them also ordering tea, again yielding 1/3 (= 5/12 × 4/5) for both. It is the same calculation presented differently: the first calculation as $P(\mathrm{S}\text{ and }\mathrm{T}) = P(\mathrm{S}|\mathrm{T})P(\mathrm{T})$ and the second as $P(\mathrm{S}\text{ and }\mathrm{T}) = P(\mathrm{T}|\mathrm{S})P(\mathrm{S})$. Bayes' rule follows: $P(\mathrm{S}|\mathrm{T})P(\mathrm{T}) = P(\mathrm{T}|\mathrm{S})P(\mathrm{S})$. With known values of $P(\mathrm{T})$ and $P(\mathrm{S})$, we can determine the probability of T given S if we know the probability of S given T. Thus, we directly estimate the conditional probability in one direction using our intuitive System 1, which is simpler. For the opposite direction, requiring analytical System 2 thinking, we employ mathematics. Given a consumer's known preferences, you expect her to order tea. However, if she orders scones first, you will likely ask, "Would you like tea with that?" Bayes' rule assigns numerical values to this System 1 intuitive reasoning. The prior probability indicates the likelihood of her ordering tea. If she orders scones first, this probability is revised to reflect the increased chance that she will want tea.

For making personal decisions, understanding inverse probability is essential. Such decisions require the slow, deliberate reasoning of System 2. Depending on the fast, intuitive thinking of System 1 can be risky in today's complex world, which is very different from the Paleolithic. Pearl and Mackenzie provide a striking example that illustrates this [1]. A 40-year-old woman receives a positive result on her mammogram test for breast cancer. If she makes an intuitive System 1 judgment, she will opt for surgery because "what you see is all there is." However, this is a forward probability perspective. To make a well-informed decision, she needs to evaluate the inverse probability, which involves the more analytical System 2 thinking based on Bayes' rule.

In this scenario, the hypothesis of the woman having the disease is represented as D, and the evidence from the test is noted as T. By applying Bayes' rule with the concept of odds instead of probability, we can express the updated likelihood of D as: Updated odds of D = Likelihood ratio × Prior odds of D. The prior odds of having the disease D are calculated as the probability of D (P(D))

divided by the probability of not having D (P(~D)). The updated odds, after the test, are the ratio of the probability of having D given the test result T (P(D|T)) to the probability of not having D given T (P(~D|T)). The likelihood ratio is determined by dividing the test's true positive rate (P(T|D)) by its false positive rate (P(T|~D)). The base rates are provided by the Breast Cancer Surveillance Consortium. For 40-year-old women, the sensitivity of mammograms (P(T|D)) is 73%, and the false positive rate (P(T|~D)) is 12%. Therefore, the likelihood ratio is 6. The prior odds of having breast cancer are also known. About 1 in 700 women aged 40 have breast cancer, making the odds of having it 1/700 divided by 699/700, which equals 1/699. According to Bayes' rule, multiplying 6 by 1/699 gives approximately 1/116, indicating a very small number of true positives. While the probability of a positive test (forward probability) is 73%, the critical inverse probability – the chance of actually having cancer given a positive test – is less than 1%. Remember, the inverse probability varies by individual and depends on context. For example, if you carry a high-risk gene, Bayes' rule lets you incorporate this factor into your assessment. In summary, Bayes' rule is effective in both predictive contexts, as illustrated by the teahouse example, and in diagnostic settings, exemplified by the mammogram scenario.

Bayes' rule, situated at the initial level of the causation ladder, deals with identifying associations by finding patterns in data. It suggests that one event is linked to another if the observation of the first alters the likelihood of observing the second. This base level focuses on making predictions from passive observations. It involves gathering and analyzing data, particularly using conditional probability. This metric quantifies the connection between two events in large datasets. Crucially, Bayes' rule can provide accurate predictions without necessarily offering detailed explanations for these associations.

In Bayesian networks, we input forward probabilities into a computer, and it calculates and provides the inverse probabilities as required. In Chapter 3 of The Book of Why, Pearl and Mackenzie [1] show the evolution of Bayes' rule into Bayesian networks. Bayes' rule, when applied using large conditional probability tables to compute all possible variable states, demands excessive computer storage and processing time. To overcome this, we can limit interactions to only a few neighboring variables, similar to human neural networks. In this hierarchical network, parent nodes (higher neurons) direct arrows to child nodes (lower neurons). Each node shares its belief level about its variable with neighboring nodes. When a parent node communicates with a child, the child updates its beliefs using conditional probabilities. These are the forward probabilities, P(evidence|hypotheses). Conversely, when a child node sends a message to a parent, the parent adjusts its beliefs by applying a likelihood ratio. These are the inverse probabilities. Belief propagation involves repeatedly applying these rules across the network, making Bayesian networks a viable method for machine learning.

Bayesian network enthusiasts represent one of five machine learning groups, the others being evolutionaries, connectionists, analogizers, and symbolists [14].

Bayesian networks differ from others in that they are transparent: each step is traceable, allowing you to see how and why every piece of evidence alters the network's beliefs.

While the notion that induction is the inverse of deduction sparks debate, its positive impact on machine learning is noteworthy. Induction involves reasoning from specific evidence to a general hypothesis, or from an effect to its cause. Conversely, deduction means reasoning from a general hypothesis to a specific conclusion, or from a cause to its effect. Is induction the counterpart of deduction, much like subtraction is to addition or integration is to differentiation? This question, a recent consideration, finds practical relevance among symbolists [14]. In deductive reasoning:

Socrates is a human being.

Humans are all mortal.

Therefore, …

Here, the first statement is a fact, the second a general rule, and we deduce the outcome. In inductive reasoning, we begin with the initial fact and the derived fact:

Socrates is human.

…

Therefore, Socrates is mortal.

Inducing the rule from Socrates alone is challenging, but an algorithm seeks it in similar facts about others. It starts with a basic but limited rule:

If Socrates is human, he is mortal.

Then, employing Newton's principle, it generalizes:

If an entity is human, it is mortal.

Ultimately, the rule emerges:

All humans are mortal.

Inferring that all swans are white based on observing n white swans is akin to making an infinite leap, which, as Hume argued, lacks logical legitimacy. Karl Popper even suggested that induction is unnecessary. According to Hume, induction is essentially our psychological tendency to assume that unobserved occurrences resemble those we have witnessed. While **Figure 2** shows a causal sequence, it is not mandatory. To deal with this, you must create hypotheses about events you have not seen and test them with your own experiences. There is no way to definitively prove a hypothesis; it can only be rejected if falsified or temporarily accepted when not falsified. At n, you propose that all swans are white. If you encounter a black swan at n + 1, your hypothesis is invalidated. The same principle applies to Taleb's Black Swan, which carries more weight than all n white swans combined. If you spot another white swan at n + 1, your hypothesis remains intact, but it does not prove that all swans are white. The appropriate stance here is empirical skepticism; you cannot assert the absence of black swans because absence of evidence is not evidence of absence.

The essence of machine learning lies in predicting previously unseen events.

This happens at the first rung of the causation ladder. We can infer the possibility of encountering a non-white swan from our knowledge of other white bird species that also have non-white variations [14]. Predicting a black swan solely based on white swan observations remains a challenge, leaving the induction problem unsolved. Machine learning takes a different approach by incorporating information about all white birds capable of changing plumage, not just swans. In this case, we must assert that non-white swans are gray. Black swans remain elusive and always manage to evade detection.

Here, causal inference is crucial because data alone cannot replace scientific knowledge. Raw data does not contain information about the consequences of actions. Deep learning programs simply fit functions to data and analyze it without considering a model, missing the predictive strength found in causal models. In contrast, causal models use an "estimand" (defined later), calculated before examining specific data. Data collection occurs once we establish a causal model, formulate a specific query, and derive the estimand. This approach, focused on causality, better predicts black swans compared to relying solely on diverse datasets and statistical methods like Bayes' rule. A key philosophical question is whether causal networks can solve the problem of induction.

4. Mediators, Confounders, and Colliders

Intervention, the next stage in causal queries, surpasses mere association as it involves not just observing but actively altering the current state. This requires generating new information that is not already in the data. A fundamental issue in data science is the assumption that all knowledge is contained within the data. However, no matter the size of the dataset or the depth of the neural network, questions about interventions cannot be answered using only passively collected data.

Therefore, to advance on the second rung of the causation ladder, we first convert Bayesian networks into causal networks. Bayes' rule, which is used for inverse probability, forms the most basic Bayesian network, consisting of just two nodes and a single link. The next level of complexity is a three-node network, which has two links and is known as a junction. Junctions serve as fundamental components for all networks, as they can represent any pattern of arrows within the network. There are three primary types of junctions:

1) $A \rightarrow B \rightarrow C$

2) $A \leftarrow B \rightarrow C$

3) $A \rightarrow B \leftarrow C$

Within these three categories, A and C exhibit correlation but lack a direct causal arrow connecting them. B plays a crucial role in each of these junctions.

In the first chain junction, B serves as a mediator. In the sequence Fire $\rightarrow$ Smoke $\rightarrow$ Alarm, there is no direct arrow from Fire to Alarm, as the Alarm is activated by Smoke, acting as the mediator. In this context, given the presence of B, A and C become conditionally independent at the chain junction.

In the fork junction, B acts as a confounder, as seen in Shoe Size ← Child Age → Reading Ability. While Shoe Size and Reading Ability correlate, giving larger shoes will not enhance reading skills. There is a correlation but no causation between Shoe Size and Reading Ability. To guide interventions, controlling for the shared factor, Child Age, is crucial. Confounding bias arises when a variable affects both the selection for treatment and the outcome of the experiment. The genuine causal effect A → C is mixed with the correlation between A and C induced by the fork A ← B → C. To effectively address confounding bias, interventions must employ deliberate, System 2 slow thinking for the control of confounders. Similar to the chain junction, given B, A and C are conditionally independent at the fork junction.

In the third junction, B acts as a collider, as seen in Talent → Celebrity ← Beauty. Avoid controlling for Celebrity, as it can create a misleading correlation between Talent and Beauty. While both Talent and Beauty contribute to an actor's success, they are unrelated in the general population. Initially, A and C are independent, but conditioning on B makes them dependent. For instance, if you observe Vin Diesel lacking Beauty, inferring Celebrity due to his Talent is flawed; he might also be untalented. Never, ever control a collider!

Imagine arrows as data pipelines. At a collider junction, A and C start independently, but controlling B connects them, unintentionally opening the data flow. In the second rung of the causation ladder, correctly handling the mediator B means not controlling it in a chain junction, maintaining data flow between A and C. Never try to control a mediator! You might inadvertently control for the very variable you intend to measure.

Figure 3 illustrates that controlling for a confounder (C) is essential, while controlling for a mediator (M) or collider (L) is not necessary, because only the confounder C directly affects both the treatment X and the outcome Y. This is the issue of overcontrol.

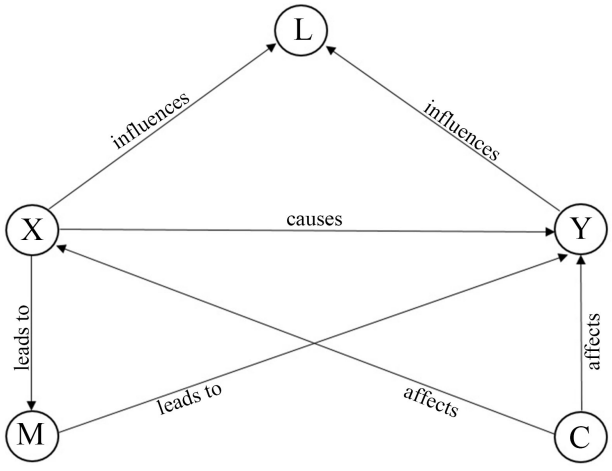

Figure 3. Control is crucial only for a confounder (C), as it directly influences both treatment X and outcome Y. Mediators (M) and colliders (L) should be disregarded.

Chains, forks, and colliders serve as crucial links between the first and second rungs of the causation ladder. They enable us to test causal models, discover new ones, and evaluate intervention effectiveness.

In a chain like A → B → C, the absence of an arrow between A and C means they are independent once their parents are known. A, with no parents, and C, with B as its sole parent, become independent once we know B's value. In a chain A → B → C, B "listens" to A, C listens to B, and A listens to no one. In this metaphor, knowledge is represented as a causal network where variables listen to each other. Causation is defined as follows: if variable Y listens to variable X and changes its value in response, then X is considered a cause of Y.

Reversing the chain's arrows alters the causal interpretation significantly, yet A and C's independence persists. This underscores the importance of crafting well-founded causal hypotheses that can withstand empirical testing and potential refutation. For instance, if the data fail to support the independence of A and C, given B, the model should be reconsidered. However, in this case, distinguishing between the fork A ← B → C and the chain A → B → C is not possible based solely on data, as both imply the same independence conditions with C influenced solely by B. Consequently, a Bayesian network cannot differentiate between a fork and a chain, as it predicts that observed changes in A relate to changes in C but does not provide predictions about A's intervention effects. Therefore, a Bayesian network resides on the first rung of the causation ladder. Nevertheless, Bayesian networks play a crucial role in enabling causal diagrams (explained next) to interact with data through junctions. In a probabilistic Bayesian network, the direction of arrows pointing to variable Y indicates that Y's probability depends on its parent variables, as defined by Y's conditional probability tables. Conversely, in a causal Bayesian network, these tables determine Y's probability in response to interventions in its parent variables.

Multiple paths often connect variable pairs, involving chains, forks, and colliders. A key criterion, d-separation (directional separation), simplifies analysis in complex models. It helps determine if nodes are d-connected (linked by a path) or d-separated (no connecting path). D-separated nodes indicate definite independence, while d-connected nodes suggest potential dependence. D-separation thus predicts expected data dependencies based on path patterns. It is useful for validating models, particularly if observed data does not align with predicted independencies. The d-separation property, which allows the use of path-blocking rules to identify independencies in data, is definitive. In other words, a causal diagram (which we will discuss next) implies no additional independencies beyond those identified through path blocking.

Certainly, getting all feasible information from data is valuable, but we must recognize its limitations. This approach alone cannot advance us past the first rung of the ladder of causation. Without a causal model, progressing from observational (rung-one) data to interventional (rung-two) queries is impossible. But with a robust causal model, we can use observational data to respond to in-

terventional queries. This enables us to predict the outcomes of interventions without conducting actual experiments. We mentally simulate an intervention before deciding its execution in reality [1].

Causal models and statistical methods serve distinct but complementary purposes in research. Causal models are designed to uncover and validate cause-and-effect relationships by constructing frameworks that simulate interventions and predict outcomes under different scenarios. They often use tools like Bayesian networks, which explicitly map out potential causal pathways and assess how changes in one variable can impact others. In contrast, statistical methods, such as regression models, focus on identifying correlations and patterns within the data without necessarily establishing causation. These methods are valuable for understanding associations and making predictions based on observed data. However, they do not inherently account for underlying causal mechanisms. The distinction lies in the fact that while statistical methods can indicate that two variables are related, causal models aim to explain why and how these relationships occur, often requiring careful consideration of confounders, mediators, and other factors that pure statistical approaches might overlook.

5. Causal Diagrams

Causal diagrams, created by linking the three types of junctions in a causal network, follow clear rules: they control for confounders but not for mediators and colliders. A confounder is an unseen factor (U) that obscures the causal link between a treatment (X) and an outcome (Y). Adhering to these rules and appropriately adding or removing arrows from the causal diagram helps in accurately ascending to the second rung of the causation ladder. And up to the third rung. Pearl suggests that these diagrams could reflect how we think about "what-if" scenarios (counterfactuals), advocating their use in machine learning for better decision-making.

Causal diagrams depict our current knowledge, whereas the do-calculus articulates our inquiries. These form the dual languages of causation calculus. The do-operator indicates intervention, as opposed to passive observation [1]. We see $P(\mathrm{Y}\,|\,\mathrm{X})$ when we look at the data in rung one of the causation ladder. Taking into account the do-operator $do(\mathrm{X})$, intervening in rung two entails $P(\mathrm{Y}\,|\,do(\mathrm{X}))$. The observed quantity is the conditional probability of the outcome Y given the treatment X, $P(\mathrm{Y}\,|\,\mathrm{X})$. However, to understand the causal relationship between X and Y, we focus on the interventional probability, $P(\mathrm{Y}\,|\,do(\mathrm{X}))$. As a result, confounding refers to $P(\mathrm{Y}\,|\,do(\mathrm{X})) \neq P(\mathrm{Y}\,|\,\mathrm{X})$. Thus, confounding is not merely a statistical concept. It represents the gap between the causal effect we aim to evaluate and what we actually measure with statistical techniques.

The causal diagram in **Figure 4** illustrates a proper intervention.

To deconfound variables X and Y, it is essential to block all noncausal paths between them while keeping causal paths open. A backdoor path, which intro-

duces a spurious correlation, is any path where an arrow points to X. By blocking all such paths, X and Y are deconfounded. In Figure 4's causal diagram, there are no arrows entering X, indicating no backdoor paths and no need for control measures. The best intervention here is inaction. Additionally, B is not a confounder in the causal path from X to Y through A, as it does not lie on this path. The backdoor criterion clearly determines the deconfounding variables in a causal diagram. The backdoor criterion serves as a practical test to identify confounding.

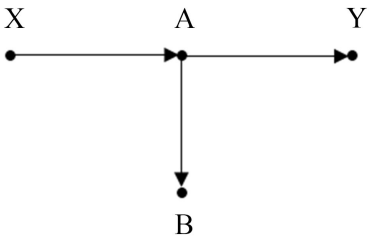

Figure 4. Causal diagram with no backdoors.

Now, examine the M-shaped causal diagram shown in Figure 5.

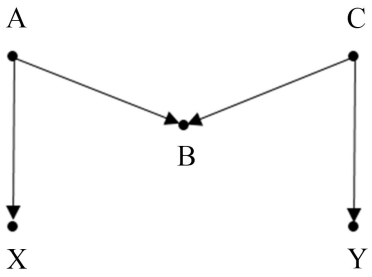

Figure 5. Causal diagram with a collider blocking one backdoor path.

The sole backdoor path in this scenario is naturally blocked by a collider at B, eliminating the need for additional control measures. Believing B is a confounder because it is associated with both X and Y is a mistake. In fact, not controlling for B ensures that X and Y remain deconfounded. It is only when B is controlled for that it turns into a confounder.

Our causal diagrams, also known as directed acyclic graphs, are largely sourced from The Book of Why [1], as seen in Figures 4-21, excluding Figure 10, Figure 14, and Figure 18. These diagrams are closely related to Bayesian networks, but differ in that each arrow in a causal diagram specifically denotes a direct causal relationship or its possibility. This distinction is important because not all Bayesian networks are causal. Software for calculating causal effects using the do-calculus is already available [15].

Causal networks combine diagrams with conditional probability tables, where each node's probability is determined by its parent nodes. This setup calculates

forward probabilities, P(evidence|hypotheses). The main role of a Bayesian network is to address inverse-probability problems, like decoding a message. For instance, it infers the probability of an original message ("Hello world!") from a received one ("Hxllo wovld!") using belief propagation. As new information is introduced, the beliefs at each node adjust dynamically throughout the network. Pearl emphasizes that a key aim of causal inference is to develop an interface that integrates human intuition with this belief propagation process [1].

6. Randomized Controlled Trials

Randomized controlled trials (RCTs), pioneered by R.A. Fisher, are considered the highest standard in clinical trials. In an RCT, individuals are randomly assigned to receive a treatment X or not, and subsequent changes in a variable Y are observed. Randomization serves as a deconfounder by eliminating influences on the treatment variable X (by erasing arrows pointing to X), allowing statisticians to accurately infer causal relationships from X to Y. From the causal perspective, RCTs serve as a man-made instrument for revealing the query $P\left(\mathrm{Y}\,|\,do\left(\mathrm{X}\right)\right)$, and are the principal contribution of statistics to causal inference. Chapter 4 of The Book of Why thoroughly examines RCTs. RCTs isolate variables X and Y from confounding variables U, clarifying the causal relationship between X and Y. RCTs uncover the query $P\left(\mathrm{Y}\,|\,do\left(\mathrm{X}\right)\right)$.

In RCTs, statisticians are uniquely allowed to talk about causes and effects. The phrase "X causes Y" is understood uniformly by both statisticians and those specializing in causal inference within this context. Causal diagrams are somewhat seen as an expansion of RCTs. Yet, Pearl notes that overvaluing RCTs is unnecessary, as other causal inference methods can replicate their results.

RCTs are not always practical. For example, it is ethically and physically impractical to conduct RCTs on the effects of smoking by making people smoke for a decade. In such cases, observational studies are used. For these studies, where randomization is not possible, causal diagrams and the do-calculus are the most reliable methods for accurate deconfounding. Causal inference leverages the experimenter's scientific knowledge, as we can still learn from observational studies where treatment and control groups are not randomly assigned. Confounding, unlike statistical concepts on the first rung of the causation ladder, is addressed in the second rung through intervention. In observational studies, statisticians often advise controlling for all available data, which can be misguided. This is less likely with causal inference, where researchers caution against controlling for mediators and colliders.

The front door adjustment, described in Chapter 7 of The Book of Why and discussed later, enables us to account for unseen confounders while observing natural behaviors outside the lab. This benefits observational studies where participants self-select rather than being randomly assigned as in an RCT. RCTs, in fact, base their validity on deeper causal inference principles. However, the do-operator offers reliable ways to identify causal effects in nonexperimental

studies, questioning the traditional supremacy of RCTs. The do (X = x) operation is fundamental as it reflects a natural property that yields the desired answer. In contrast, randomization is a secondary, artificial technique used to extract that answer [1]. However, randomization offers two key advantages: 1) it removes confounder bias by correctly posing the question to nature, and 2) it allows the researcher to measure their uncertainty accurately.

7. Does Smoking Cause Lung Cancer?

Fisher, once revered, faced skepticism after reviewing observational studies on this question. Given that some lifelong smokers avoid lung cancer while some non-smokers develop it, Fisher claimed the perceived link between smoking and lung cancer was merely coincidental. He suggested that smokers might be "constitutionally" (genetically) predisposed to behaviors detrimental to their health. Figure 6 illustrates Fisher's perspective using a causal diagram.

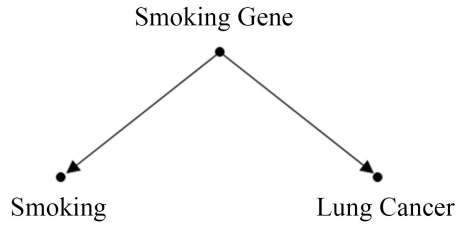

Figure 6. Fisher's causal diagram on Smoking and Lung Cancer.

The Smoking Gene as a hidden third variable could confound the relationship. Fisher's model omits the direct link Smoking → Lung Cancer, often inferred from observational studies. In contrast, the rival theory, depicted in Figure 7, includes this direct link.

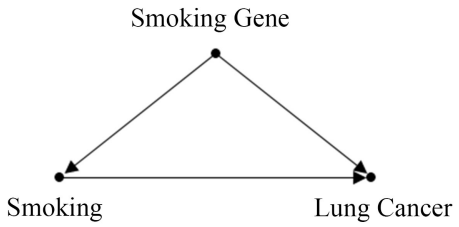

Figure 7. Opposing view's causal diagram linking Smoking to Lung Cancer.

The responsibility fell to the antismoking group to either disprove the existence of any confounding factor or to prove a negative. Jerome Cornfield employed an early form of sensitivity analysis, supported by data from associational studies, to demonstrate to epidemiologists and policymakers that the Smoking Gene by itself could not account for the significant connection between Smoking and Lung Cancer. Cornfield's reasoning indicated that this genetic factor was not enough to justify the significant impact of Smoking on Lung Cancer risk.

In 2008, the Smoking Gene was identified as a single nucleotide polymorphism on chromosome 15, which affects nicotine receptors in lung cells. This genetic variation has two forms. About 11% of people possess two copies of the rarer variant, elevating their lung cancer risk to 77%. This variant also increases nicotine dependence and quitting difficulty, linking the gene to risky smoking behavior. These discoveries warrant an update to the previous causal diagram in Figure 7, as depicted in the revised Figure 8.

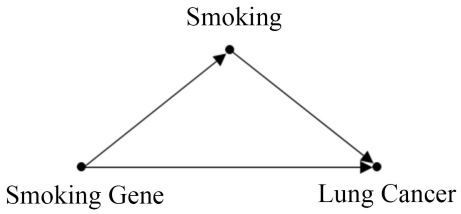

Figure 8. Causal diagram illustrating the effects of the Smoking Gene.

Rather than questioning if Smoking causes Lung Cancer, as it does, we focus on understanding the direct and indirect effects of the Smoking Gene, mediated by Smoking. Epidemiologist Tyler VanderWeele found that the Smoking Gene neither notably raises cigarette use nor causes Lung Cancer independently of Smoking. However, it does significantly heighten Lung Cancer risk in smokers. Fisher lost the debate posthumously, as the direct link Smoking Gene → Lung Cancer needs to be eliminated in Figure 8. For a detailed discussion of this debate, see Chapter 5 of The Book of Why.

Jacob Yerushalmy, supporting Fisher's views, noted that smoking during pregnancy appeared beneficial for underweight newborns' health. Figure 9's causal diagram summarizes his research findings.

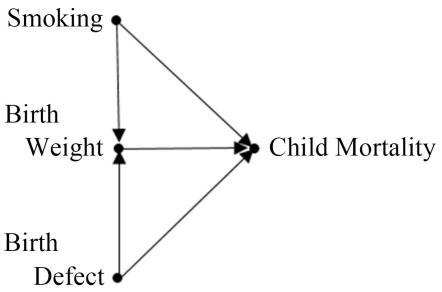

Figure 9. Causal diagram of the birth-weight paradox.

In this case, a statistician overlooked a collider, which was Birth Weight. By focusing solely on low Birth Weight babies, Yerushalmy inadvertently activated a noncausal backdoor path between Smoking and Child Mortality. This path, featuring an incorrectly directed arrow, produced a misleading negative correlation, falsely suggesting that Smoking had a beneficial effect. The paradox is not related to birth weight; it is entirely associated with colliders.

The birth-weight paradox remained unresolved for over forty years after Yerushalmy's publication, long after the smoking-cancer debate subsided. Pearl and Mackenzie attribute this delay to the lack of a causality framework at the time. The concept of collider bias, clarified through a causal diagram, revealed a hidden collider structure in the data selection process.

The smoking debate highlights the survival value of understanding causality. Without the possibility of reliable randomized controlled trials, many lives were lost due to statisticians' lack of proper tools and language for addressing causal questions. Furthermore, the complexity of the discussion arises from the differences in causation between diseases like scurvy and lung cancer, for example. Scurvy is solely caused by a deficiency in vitamin C, making this deficiency both a necessary and sufficient condition. In contrast, lung cancer is likely caused by multiple factors. Therefore, while smoking is a necessary cause of lung cancer, it is not sufficient on its own to cause the disease.

8. Paradoxes

Paradoxes arise from conflicts between statistical reasoning (System 2) and causal intuition (System 1). Human intuition is fundamentally based on causal logic, not statistical. Our minds in automatic mode tend to fall for randomness, perceiving patterns where none exist, known as type I error. Additionally, when examining correlations caused by colliders, like in the previous example, we inadvertently create patterns from what was initially random.

Coin flips are independent events. However, consider this experiment: flip two coins 100 times and only record results when at least one shows Heads. In your 75-entry table, the coin flips seem dependent – if one shows Tails, the other shows Heads. This perceived correlation arises because we excluded all Tails-Tails outcomes, thus conditioning on a collider, and inadvertently created a correlation.

Observing a correlation between Tails and Heads while deliberately controlling for the Tails-Tails collider (as shown in Figure 10) leads to a type I error. This error arises because we attempt to find a stable, causal relationship in the data that, when correctly sampled, is not actually there. Collider bias is a cognitive illusion.

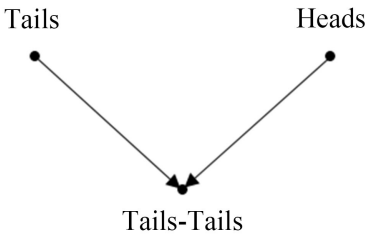

Figure 10. Causal diagram of coin flip correlations with collider control.

Consider the factors of attractiveness and personality in dating choices. If you date attractive but mean, attractive and nice, or unattractive but nice individuals,

but never mean and unattractive ones, you might wrongly conclude that attractiveness correlates with meanness. You form a belief that the attractive people you date are often unpleasant. Yet, this belief is a cognitive illusion resulting from collider bias. Mean behavior is equally common in unattractive and attractive people. However, you will not notice this because you avoid dating those who are both mean and unattractive. This example is similar to the coin flip scenario in Figure 10.

Biostatistician Joseph Berkson found that unrelated diseases could appear linked in hospital patient samples. This is illustrated in Figure 11, where a false positive correlation emerges between Respiratory Disease and Bone Disease due to controlling for Hospitalization. This phenomenon, known as the Berkson paradox, occurs from unintentionally controlling a collider.

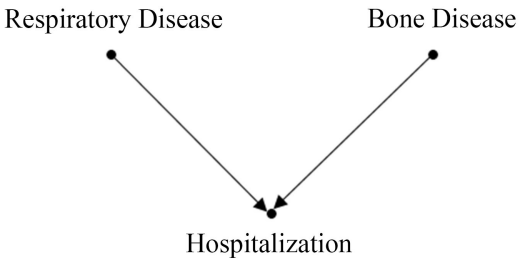

Figure 11. Causal diagram illustrating the Berkson paradox.

On a game show, you choose from three doors: one hides a car, the other goats. You pick Door 1. Host Monty Hall, aware of what is behind each door, opens Door 3, revealing a goat. He then asks if you want to switch to Door 2. Should you change your choice?

You should say "yes" to switching doors. Without switching, your odds of winning the car are 1 in 3; switching increases them to 2 in 3. However, your System 1 might say "no," mistakenly thinking the odds are 50-50 and that switching does not matter. System 1 can mislead you by wrongly assuming a causal connection between your chosen door and the car's location.

The fact that Monty Hall opened Door 3, a collider, creates a false association. To correctly calculate your odds, ignore Monty's choice. The lesson: rely on empirical data analysis. However, System 1 thinking may tempt you to consider the collider, but it is System 2's slower, deliberate thinking that helps you accurately assess probabilities and causality by understanding both the data and the game's rules.

Statisticians adhering to Fisher's principle of focusing solely on data can still be misled by the Monty Hall paradox, as shown in Figure 12's causal diagram. In this game, there is no direct link between Door 1 and the Car Door, indicating that your choice and Monty Hall's car placement are independent events. Door 3, however, is affected by both your choice of Door 1 and the Car Door's location since Monty Hall takes both into account. Thus, Door 3 serves as a collider, with no causal relationship between your chosen door and the Car Door.

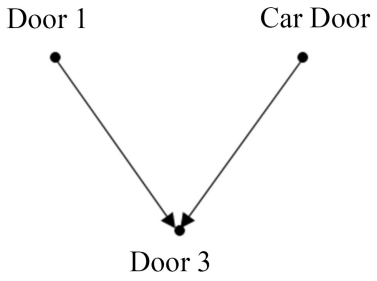

Figure 12. Causal diagram of the Monty Hall paradox.

Since Door 3 was opened, it has become a collider. Learning this affects our probabilities, making them conditional on this information. Conditioning on a collider introduces a false dependence between its causative factors. This is evident in the probabilities: if you chose Door 1, it is twice as likely that the car is behind Door 2 rather than Door 1; if you chose Door 2, it is twice as likely to be behind Door 1. This creates an odd dependence without any causal relationship. This is solely an artifact of Bayesian conditioning.

System 1 is wired for causal reasoning, but not for probability tasks. System 1 equates correlation with causation. For example, if a car behind us mimics our turns, we initially assume it is following us, implying causation. We then instinctively think we share a destination, as fast thinking assumes a shared cause for each turn. However, correlations without causes disrupt this fast-thinking mode. System 1 is incapable of neutralizing colliders.

Pearl argues that relying only on data is flawed, as the same data can result from different data-generation processes [1]. Consider a variation of the game where Monty Hall randomly chooses a door different from yours, as illustrated in **Figure 13**'s causal diagram. An arrow from Door 1 to Door Opened remains, reflecting Monty's need to choose a different door. But with Monty's choice now random, there is no link from Car Door to Door Opened. Thus, focusing on Door Opened does not change the situation: your choice and the Car Door remain independent, both before and after Monty's choice. With the odds now 50-50, switching doors has no benefit.

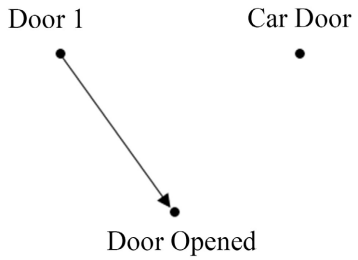

Figure 13. Causal diagram of a modified Monty Hall game.

Statisticians employing model-blind methods and not considering causality are prone to paradoxes, as the same data can lead to correct conclusions in one scena-

rio but incorrect ones in another. The means by which we gather information is just as important as the information we obtain [1]. Chapter 6 of The Book of Why explores more of these paradoxes. The paradoxes highlight the conflict between association and causation, which occupy distinct rungs of the causation ladder. This conflict is intensified because human intuition (System 1) tends to think in terms of causation, while data is governed by probabilities and proportions. Paradoxes occur when we incorrectly apply rules from one domain to the other.

To make valid causal inferences, it is essential to control for confounders. For example, if age (Z) is the confounding variable, we analyze treatment and control groups within each age group. Then, we average the effects, weighting by each age group's representation in the target population. This is how we control for Z. However, mistaking mediators for confounders is another source of error. For instance, when deconfounding, we face decisions like whether to segregate data. With readily available data, age and gender are often chosen for demographic control, due to their accessibility.

However, consider this case. To determine if exercise lowers LDL cholesterol, an observational study was conducted, collecting participants' ages and birth genders. The results revealed a contradiction: while exercise was beneficial for individuals in every age group, it was detrimental for the overall population. This exemplifies Simpson's paradox, a puzzle that highlights confounding factors in data interpretation. Without age segregation, data showed a misleading positive correlation, suggesting exercise increases cholesterol. But, careful analysis revealed older individuals tend to exercise more. Since cholesterol levels also vary with age, age is identified as a confounder in the exercise-cholesterol relationship. Figure 14's causal diagram illustrates this. When age is considered, the correlation reverses, indicating exercise does reduce bad cholesterol, regardless of age.

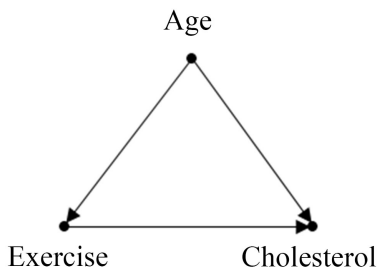

Figure 14. Causal diagram showing Simpson's paradox and its confounding factor.

Simpson's paradox occurs when a trend appears to go in one direction within individual segments of a population, but in the opposite direction when considering the entire population. Resolving this paradox depends on the specific question being addressed. If every individual in a group prefers one option over another, this preference should be reflected in the group's overall data. If the data suggests otherwise, it indicates an error in data processing, particularly in understanding causal relationships.

Lord's paradox is a variation similar to Simpson's paradox: in a school, overall neither boys nor girls gain weight over the year. However, within each initial weight category, boys generally gain more weight than girls. How is this possible? Is not the total weight gain simply the average of the gains in each specific stratum? No, if the composition of the strata changes during the treatment.

The school investigates how two diets affect weight gain. Student weights are recorded at the start and end of the year. They eat in one of two halls, each serving a different diet. Heavier students tend to choose a particular hall, leading to a causal diagram in Figure 15 showing a link from Initial Weight to Diet Choice. Initial Weight also influences Final Weight. Since Gain equals Final Weight minus Initial Weight, the deterministic correlations are −1 and +1. To accurately evaluate Diet's impact on Final Weight, the confounder, Initial Weight, must be controlled.

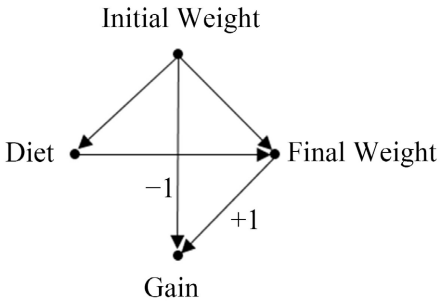

Figure 15. Causal diagram illustrating the control of a confounder in Lord's paradox.

The causal diagram changes, as Figure 16 illustrates, when the school considers the diet's differing effects on girls and boys. Sex is linked to both Initial and Final Weight. Regardless of Sex, a higher Initial Weight generally leads to a higher Final Weight. Now, Initial Weight acts as a mediator instead of a confounder, making control of it incorrect.

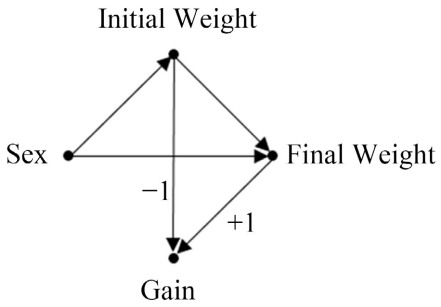

Figure 16. Causal diagram showing the control of a mediator in Lord's paradox.

Simpson's and Lord's paradoxes are detailed in Chapter 6 of The Book of Why. These paradoxes stem from confusion between confounders and mediators. Chapter 9 of the same book discusses the fallacy of conditioning on a mediator in depth.

9. Deconfounding

Galton developed the regression line $Y = r_{YX}X + b$ to show the relationship between a treatment variable X and an outcome variable Y. He did this by drawing a line that best fits through a set of datapoints. The regression coefficient of Y on X, r_{YX}, indicates that for every one-unit increase in X, Y increases by an average of r_{YX} units. However, if a confounding variable Z is present, this coefficient r_{YX} represents only the observed trend, not the average causal effect between X and Y.

Karl Pearson and George Yule found that the partial regression coefficient $r_{YX.Z}$ automatically adjusts Y's trend on X for the confounder Z in the regression plane equation $Y = r_{YX.Z}X + bZ + c$. This means it is unnecessary to separately regress Y on X at each level of Z in linear regressions. Therefore, $r_{YX.Z}$ can indicate the average causal effect, assuming Z is a confounder and not a mediator or collider.

Since data alone cannot clarify Z's role, we need to apply the backdoor criterion in a causal diagram to confirm Z as a confounder. This ensures $r_{YX.Z}$ reflects the average causal effect. The Book of Why delves into this in Chapter 2, providing historical context for regression lines and causal inference, and offers a detailed explanation in Chapter 7.

Even when X's effect on Y varies based on the confounder Z's level, as seen in nonlinear interactions, the backdoor criterion remains applicable. In these nonparametric cases, it is used through extrapolation methods. Linear regression's partial regression coefficients automatically adjust for confounders; however, in nonparametric regression, this adjustment must be explicitly done, either directly using the backdoor criterion or through an extrapolated form of it.

The misconception that partial regression coefficients $r_{YX.Z}$, by adjusting for confounders, contain causal information that unadjusted coefficients r_{YX} do not, is incorrect. For causal legitimacy, two key elements are needed: 1) a path diagram that accurately reflects reality, and 2) the adjusted variable Z must meet the backdoor criterion [1].

During the debate about Smoking and Lung Cancer, the Smoking Gene confounder was undetectable. With causal diagrams, we could have resolved this issue without needing Cornfield's mathematical calibration. As shown in **Figure 17**'s causal diagram, if the Smoking Gene is unknown and thus unmanageable, it is impossible to block the path: Smoking ← Smoking Gene → Lung Cancer through the Smoking Gene using a backdoor adjustment.

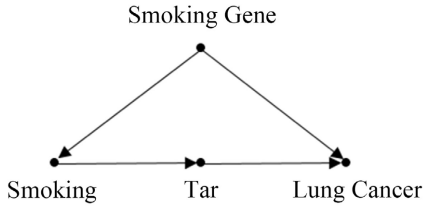

Figure 17. Causal diagram illustrating the front door criterion.

If we think tar in smokers' lungs causes lung cancer, we can apply the front door criterion. The front door is the direct causal path: Smoking → Tar → Lung Cancer. We have data on all three of these variables. The front door adjustment, unlike the backdoor adjustment, involves adjusting for two variables, Smoking and Tar, which are on the front door path leading from Smoking to Lung Cancer, rather than on the backdoor path.

The collider at Lung Cancer is blocking the path: Smoking ←Smoking Gene → Lung Cancer ←Tar. Consequently, we can accurately estimate the average causal effect of Smoking on Tar. While a backdoor adjustment is not feasible, it is unnecessary in this scenario. At the first rung of the causation ladder, we gather data on $P(\text{Tar}\,|\,\text{Smoking})$ and $P(\text{Tar}\,|\,\text{No Smoking})$, and then calculate their difference to determine the average causal effect of Smoking on Tar.

We next estimate the average causal effect of Tar on Lung Cancer. Since we have Smoking data, we can block the backdoor path: Tar ← Smoking ← Smoking Gene → Lung Cancer by adjusting for Smoking. After gathering data at the first rung of the causation ladder, we intervene at the second rung by calculating $P(\text{Lung Cancer}\,|\,do(\text{Tar}))$ and $P(\text{Lung Cancer}\,|\,do(\text{No Tar}))$. The difference between the two represents the average causal effect of Tar on Lung Cancer.

Finally, we can calculate the causal effect of Smoking on Lung Cancer using observational study data from the first rung of the causation ladder. We can express $P(\text{Lung Cancer}\,|\,do(\text{Smoking}))$ in terms of probabilities and this does not require the use of the do-operator. In this case, a randomized controlled trial is not needed. Assuming X represents Smoking, Y Lung Cancer, Z Tar, and U the unobservable Smoking Gene, the front door adjustment then implies

$$P(\text{Y}\,|\,do(\text{X})) = \sum_z P(\text{Z}=z\,|\,\text{X}) \sum_x P(\text{Y}\,|\,\text{X}=x,\text{Z}=z) P(\text{X}=x).$$

The left side of the equation asks, "What effect does X have on Y?" The right side provides the estimand, the method to answer this query. Estimand, derived from Latin, means "what needs to be estimated." Noticeably, the right side only includes do-free probabilities and excludes U. Therefore, we can calculate the causal effect of Smoking on Lung Cancer using just data, effectively deconfounding U without its data.

If a backdoor adjustment were feasible, it would imply

$$P(\text{Y}\,|\,do(\text{X})) = \sum_u P(\text{Y}\,|\,\text{X},\text{U}=u) P(\text{U}=u).$$

In Figure 17, Tar acts as a shielded mediator. If individuals with the Smoking Gene are more prone to Tar formation and those without it are more resistant, an arrow from Smoking Gene to Tar must be added in the causal diagram (Figure 18). Since Tar is no longer shielded, a front door adjustment becomes unfeasible.

The front door adjustment allows for controlling unseen confounders, including unnamed ones, much like RCTs. However, its advantage over RCTs lies in observing individuals in natural, field settings rather than in a laboratory.

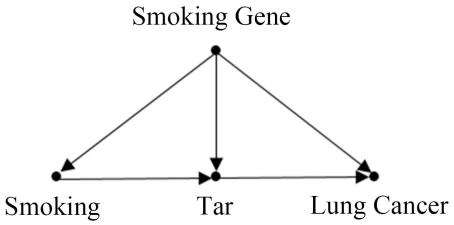

Figure 18. Causal diagram showing a blocked front door path.

10. The Do-Calculus

The primary aim of backdoor and front door adjustments is to estimate the impact of an intervention using data that do not rely on a do-operator. Successfully removing the do-operators enables the use of observational data for assessing causal effects, advancing from the first to the second rung in the causation ladder. There exists a method to determine beforehand if do-operators can be eliminated in a specific causal model. Hence, the do-calculus model can substitute an experiment by converting a do quantity into a see quantity. If this method indicates that do-operators are irremovable, it implies that our assumptions are insufficient for deriving causal effects from observational data alone, necessitating the use of RCTs.

When both backdoor and front door adjustments fail to enable successful intervention amid confounders, an alternative exists. The fully automated do-calculus allows for customizing the adjustment method to suit any given causal diagram. The objective is to determine the impact of variable X on Y, beginning with the target sentence $P(\mathrm{Y} \mid do(\mathrm{X}))$. The key step is to remove the do-operator from this sentence, resulting in only standard probability expressions, such as $P(\mathrm{Y} \mid \mathrm{X})$ or $P(\mathrm{Y} \mid \mathrm{X}, \mathrm{Z}, \mathrm{W})$. This removal process must accurately represent the physical intervention implied by do(X), achieved through a series of valid deductive steps.

In the do-calculus, there are three foundational rules for valid manipulations. Rule 1 highlights that observing a variable W, which is independent of Y (given other variables Z), does not affect Y's probability distribution. For instance, in the chain Fire $\rightarrow$ Smoke $\rightarrow$ Alarm, knowing the mediator Z (Smoke) renders W (Fire) irrelevant for Y (Alarm). Thus, Rule 1 implies

$$P(\mathrm{Y} \mid do(\mathrm{X}), \mathrm{Z}, \mathrm{W}) = P(\mathrm{Y} \mid do(\mathrm{X}), \mathrm{Z}).$$

This implies that once we remove all incoming arrows to X, Z will obstruct any path from W to Y. Although X is not present in our example, Smoke (Z) effectively blocks every path from Fire (W) to Alarm (Y).

Rule 2 states that if Z blocks all backdoors from X to Y, then do(X) is the same as see(X) when considering Z. This means if Z satisfies the backdoor criterion, then

$$P(\mathrm{Y} \mid do(\mathrm{X}), \mathrm{Z}) = P(\mathrm{Y} \mid \mathrm{X}, \mathrm{Z}).$$

Simply put, Rule 2 indicates that after controlling for all confounders, any persisting correlation is a true causal effect.

Rule 3 allows for the removal of do(X) from $P\left(\mathrm{Y}\mid do\left(\mathrm{X}\right)\right)$ when there are no causal paths from X to Y. If no paths exist from X to Y, Rule 3 asserts:

$$P\left(\mathrm{Y}\mid do\left(\mathrm{X}\right)\right)=P\left(\mathrm{Y}\right).$$

Essentially, Rule 3 implies that if an action on the treatment variable X does not influence the outcome Y, then Y's probability distribution remains unchanged. Do something that has no effect on Y, and the probability distribution of Y will stay the same.

Rule 1 permits adding or removing observations. Rule 2 enables switching between observation and intervention. Rule 3 allows adding or eliminating interventions.

As observed, the primary aim of the axiomatic do-calculus, similar to back-door and front door adjustments, is to validly deduce the impact of an intervention $P\left(\mathrm{Y}\mid do\left(\mathrm{X}\right)\right)$ using data without a do-operator, such as $P\left(\mathrm{Y}\mid \mathrm{X},\mathrm{Z}\right)$. Interestingly, a front door adjustment formula, a do-operator-free expression, can be derived through multiple applications of do-calculus rules, using a particular causal diagram as input. This formula estimates causal effects through methods other than controlling for confounders. Moreover, if a causal effect can be estimated from data, a series of steps applying these three rules will remove the do-operator.

To better understand causation, it is more useful to respond to causal queries than to initially define causation. Definitions often necessitate breaking down concepts to simpler forms. Causation, however, is not easily simplified to just basic probabilities, which are at the rung one of the causation ladder [1]. With the introduction of the do-operator, however, we can offer a definition: we say X causes Y if $P\left(\mathrm{Y}\mid do\left(\mathrm{X}\right)\right)>P\left(\mathrm{Y}\right)$. Chapter 7 in The Book of Why presents the do-calculus.

Graphical models are now widely used in epidemiology. Yet, most econometricians, for example, are still doubtful about using graphical analysis tools [16] [17]. However, some have extended and applied causal diagrams and the do-calculus to areas such as economic optimization, equilibrium, and learning [18] [19], and also to social and behavioral methods [20] [21].

A frequent criticism of causal diagrams is the assumption that a simple graphic can fully capture the complex interactions of multiple variables and their joint effect. The actual goal of these diagrams, however, is not to prove causality between X and Y or to identify Y's root cause from the beginning. It is simply to encode plausible causal knowledge in a mathematical language, combine it with empirical facts, and respond to causal queries that have practical significance. Discovering causality is much harder, often impossible. Therefore, causal diagrams are best used for exploration. We form hypotheses about causal connections and predict variable correlations. If these predictions conflict with actual data, it suggests our assumptions were wrong. We cannot draw causal conclu-

sions without a causal hypothesis, which implies we cannot answer a question regarding rung two of the causation ladder based just on information from rung one.

Note that we are not simply assuming what we want to prove, which would make the causal reasoning circular. This is because, following causal analysis, we get unique information, allowing us to extract the non-obvious from the obvious. Causal analysis involves more than just data; it requires integrating knowledge of the processes that generate the data, leading to insights not originally present in the data. A causal diagram must be justified based on scientific reasoning. One objective of causal inference is to develop a more intuitive human-machine interface, enabling the user's intuition to be incorporated into the belief propagation mechanism [1].

11. Instrumental Variables

An instrumental variable Z can serve similarly to a front door adjustment for assessing X's impact on Y when controlling or obtaining data on a confounder U is not feasible. This method is especially effective in cases similar to the causal diagram shown in **Figure 18**, where front door adjustment is not feasible.

While instrumental variables were used before the advent of causal diagrams, the introduction of these diagrams has enhanced our understanding of how instrumental variables operate. First, let us assume the variables are numerical and their relationships are linear. In the causal diagram shown in **Figure 19**, an intervention increasing Z by one unit results in X increasing by a units. Z qualifies as an instrumental variable because there is no direct path U → Z, ensuring Z and X are deconfounded and the Z → X relationship is causal. Consequently, a can be estimated from the slope r_{XZ} of the regression line between X and Z.

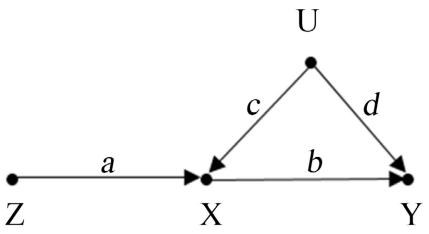

Figure 19. Causal diagram showing an instrumental variable Z.

Additionally, Z and Y are deconfounded due to the collider at X, which blocks the indirect path Z → X ← U → Y. Consequently, the slope r_{YZ} of the regression line from Y on Z represents the causal effect on the direct path Z → X → Y, quantified as ab.

Then, dividing the equation $ab = r_{YZ}$ by $a = r_{XZ}$ yields $b = r_{YZ}/r_{XZ}$, representing the causal effect X → Y. Thus, we infer about b, from the second rung of the causation ladder, using correlations r_{XZ} and r_{YZ} from the first rung. Instrumental variable methods are also applicable to nonlinear variables,

yielding range estimates instead of specific point estimates.

The solution in **Figure 19** was simple due to the collider obstructing the indirect path. When this blockage is absent, assessing the direct and indirect effects of an intervention becomes crucial, as detailed in Chapter 9 of The Book of Why.

In the 1853-4 London cholera outbreak, Dr. John Snow unintentionally used an instrumental variable (Z) to clarify the causal relationship between water purity (X) and cholera (Y), unaffected by the unobserved confounder of unhealthy air (U). He distinguished between two water companies, one upstream and one downstream of the sewers. In areas served by both, the air quality was constant, eliminating confounders (U) for the instrumental variable Z. This demonstrated that the link between water purity (X) and cholera (Y) was indeed causal.

After carefully considering the causal intuition from System 1, we conclude that there is no connection between U and Z. This intuition is represented and explained in the causal diagram using System 2. We rely on causal intuition (System 1) for responding to causal queries, with this intuition being encapsulated, clarified, and detailed in the causal diagram using System 2. Instrumental variables are valuable as they reveal causal insights beyond the do-calculus, making them highly useful in observational studies. They also aid in RCTs, particularly when noncompliance occurs, like when participants are assigned a drug but do not take it. However, the do-calculus offers greater flexibility than instrumental variables as it does not require assumptions about the causal model's function types. Additionally, causal diagrams are essential for effectively applying instrumental variables methods, which have limited scope on their own.

Econometric textbooks cover instrumental variables [22]-[24], but econometricians often resist adopting causal diagrams [25], struggling with the concept of causality [26]. Causal diagrams provide a graphical yet mathematically robust method for causal inference. Analyzing these diagrams can be labor-intensive, making them suitable for computer program automation. The online tool DAGitty allows users to identify generalized instrumental variables in diagrams, reporting the estimands found [27]. Additionally, BayesiaLab offers another diagram-based software for decision-making (https://www.bayesia.com/).

AlphaGeometry combines a neural language model with a symbolic deduction engine to prove complex geometry theorems. This dual approach mirrors the "thinking, fast and slow" concept [6]. One part offers quick, intuitive insights, while the other delivers methodical decisions. We propose applying this neuro-symbolic method for creating causal diagrams in research. It balances System 1's intuitive causal reasoning with the thoroughness of causal inference using System 2.

12. Counterfactuals

To step up the top rung of the causation ladder, data alone is insufficient. The possibilities of what might have been are never observed. Data cannot predict outcomes in counterfactual scenarios, which negate existing facts. Counterfac-

tuals conflict with data, which are factual by nature. Yet, knowledge extends beyond data. For instance, the laws of physics represent counterfactual assertions [1]. Rung one deals with the observable world, rung two with a possible world that can be observed, and rung three with an unobservable world that contradicts what is seen [1].

It is pointless to inquire about the causes of things if you cannot envision their outcomes. This is what causal imagination entails. Before, we examined the impact on either a whole population or a typical individual from that population, by assessing the average causal effect. Now causal inference enables us to generate counterfactuals for an individual. This System 2 technology complements the mind's System 1 counterfactual generation while avoiding cognitive biases. We employ both observational and experimental data to understand counterfactual scenarios. Causal diagrams are used to depict causes on an individual level.

A counterfactual, also known as a potential outcome, is the value of an outcome Y for an individual u if a certain condition X had occurred (X = x), denoted as $Y_{X=x}(u) \equiv Y_x(u)$. Table 1 shows hypothetical data on employee salaries, $S_{ED=i}(u)$, education levels (ED), and years of experience (EX) for a company. A common counterfactual query is, "What would Alice's salary be if she had a college degree?" In this scenario, education levels are coded as i = 0 for high school, i = 1 for college, and i = 2 for a graduate degree. We aim to find the potential salary outcome $S_1(\text{Alice})$ for a college education.

Table 1. Hypothetical employee data.

Employee u	EX(u)	ED(u)	$S_0(u)$	$S_1(u)$	$S_2(u)$
Alice	6	0	81,000	?	?
Bert	9	1	?	92,500	?
Caroline	9	2	?	?	97,000
David	8	1	?	91,000	?
Ernest	12	1	?	100,000	?
Frances	13	0	97,000	?	?

In Table 1, each employee has just one observable potential outcome. A statistician views the missing data, marked by question marks, as regular variables and would apply interpolation methods. For instance, using a matching technique, if Bert and Caroline share the same years of experience (EX(u)), then $S_2(\text{Bert}) = S_2(\text{Caroline}) = 97000$ and $S_1(\text{Caroline}) = S_1(\text{Bert}) = 92500$.

To answer the counterfactual question, a statistician uses these matched data pairs. Yet, no statistical method can transform data into potential outcomes, as this depends on whether education (ED(u)) leads to experience (EX(u)) or vice versa. This causal information is not available in Table 1.

An alternative statistical method uses the linear regression $S = 65000 + 2500EX + 5000ED$, where the intercept is the average starting salary for an employee with no experience and a high school diploma. The model adds

$2500 for each year of experience and $5000 for each extra educational degree (up to two). However, this approach has a flaw: it overlooks the dependency of experience on education (ED → EX). Attending college for four years, for instance, could otherwise contribute to work experience. Unlike the matching method, acknowledging this opportunity cost results in $S_1(\text{Caroline}) > S_1(\text{Bert})$.

To properly tackle counterfactual questions, one should employ a structural causal model. We assess statements like "had X been x" in the same manner as interventions do(X = x), by removing arrows in a causal diagram or altering equations in a structural model. Therefore, prior to analyzing the data in Table 1, it is crucial to first examine the causal diagram shown in Figure 20.

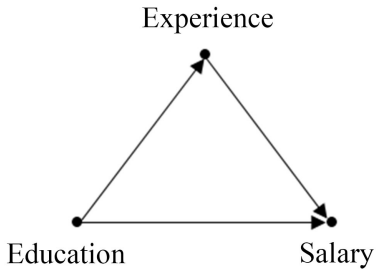

Figure 20. Diagram illustrating how education and experience affect salary.

If EX → ED, then EX acts as a confounder. However, if ED → EX, EX serves as a mediator (Figure 20). Our analysis starts with $S = f_S(\text{EX}, \text{ED})$. We then extend our approach to consider unobserved factors, represented as U_S, that affect salary. Therefore, $S = f_S(\text{EX}, \text{ED}, U_S)$. Galton noted that "regressions are cause blind," so we continue by building on our previous linear regression equation $S = 65000 + 2500\text{EX} + 5000\text{ED} + U_S$. This equation becomes structural when we incorporate our causal conjecture $S = f_S(\text{EX}, \text{ED}, U_S)$.

To finalize the model, we incorporate the equation $\text{EX} = 10 - 4\text{ED} + U_{\text{EX}}$, calculated using Table 1 data. This data indicates employees with only a high school diploma average ten years of experience. Additionally, each year of education beyond high school (up to two years) corresponds to a four-year reduction in experience. Hence, this equation explicitly considers the opportunity cost, which was previously overlooked by the statistical methods.

There is an arrow from Experience (EX) to Salary (S), but not from Salary to Experience in Figure 20. Although there is a strong correlation between S and EX, the coefficient of S is zero, indicating no causal relationship between Experience and Salary in our analysis. Notably, in our structural causal model, Education lacks a causal arrow, ruling out any equation like $\text{ED} = f_{\text{ED}}(\text{EX}, S, U_{\text{ED}})$.

To estimate Alice's salary using structural causal equations, we undertake three steps. Initially, we estimate variables $U_S(\text{Alice})$ and $U_{\text{EX}}(\text{Alice})$ using data from Table 1 for Alice and other employees. Next, we apply the do-operator for the counterfactual hypothesis $\text{ED}(\text{Alice}) = 1$. Lastly, with this data, we

compute Alice's revised salary.

In the first stage, positioned at the first rung of the causation ladder, we input $\mathrm{EX}(\mathrm{Alice}) = 6$ and $\mathrm{ED}(\mathrm{Alice}) = 0$ into $\mathrm{EX} = 10 - 4\mathrm{ED} + \mathrm{U}_{\mathrm{EX}}$ to determine $\mathrm{U}_{\mathrm{EX}} = -4$. Subsequently, using the same data and setting S to 81,000, we feed these into $\mathrm{S} = 65000 + 2500\mathrm{EX} + 5000\mathrm{ED} + \mathrm{U}_{\mathrm{S}}$ to calculate $\mathrm{U}_{\mathrm{S}} = 1000$.

In the second phase, at the second rung of the causation ladder, we focus on variable $\mathrm{ED}(\mathrm{Alice}) = 1$. Moving to the third phase, at the third rung, we input $\mathrm{U}_{\mathrm{EX}} = -4$ and set ED = 1 into equation $\mathrm{EX} = 10 - 4\mathrm{ED} + \mathrm{U}_{\mathrm{EX}}$, resulting in EX = 2.

We then incorporate $\mathrm{U}_{\mathrm{S}} = 1000$ into $\mathrm{S}_{\mathrm{ED}=1}(\mathrm{Alice}) = 65000 + 2500\mathrm{EX} + 5000\mathrm{ED} + \mathrm{U}_{\mathrm{S}}$ to ultimately derive $\mathrm{S}_{\mathrm{ED}=1}(\mathrm{Alice}) = 76000$, which results in a value less than the $85,000 estimated by the linear regression method. The linear regression method produces spurious correlations because it overlooks key causal hypotheses. In our analysis, we factored in the opportunity cost in the causal narrative, leading to a lower estimate of the counterfactual salary.

In this example, we used a fully specified structural causal model. Had it been only partially specified, the counterfactual outcome would be expressed as a probabilistic range, such as "there is an 80% - 90% chance that the salary will be $76,000". A complete structural causal model, which includes both a causal diagram and its underlying functions, enables us to respond to any counterfactual query. This section's example is adapted from Chapter 8 of The Book of Why, where we identified the opportunity cost. Causal analysis demands subjective judgement rather than solely relying on objective statistics. In this case, the focus was on its application in the field of economics.

Counterfactuals, or "what-if" scenarios, can be algorithmically systematized, enabling machines to emulate human retrospective thought. This involves using algorithms to analyze real-world data and generate insights about hypothetical situations. Unlike abstract metaphysical logic, this process relies on structural causal equations and diagrams, which use clear rules for drawing and omitting connections. These methods closely resemble how System 1 processes counterfactuals. Although empirical evidence cannot disprove such hypotheticals, we can still form highly reliable and consistent judgements about potential outcomes. Counterfactual reasoning is essential not only in scientific inquiry but also in moral decision-making [1]. Hence, equipping AI with a causal reasoning module is a vital step towards achieving strong AI.

2023 was the year AI became mainstream, predominantly led by ChatGPT. The potential of this AI generation prompts questions: can it achieve consciousness and become a threat, or will it plateau? Based on our discussion, this AI generation has not advanced to the second rung of the causation ladder, making the outcomes clear. We need to integrate causal inference into the next AI generation to move closer to strong AI. Today's machine learning techniques efficiently turn finite sample estimates into probability distributions. However, we still need to derive cause-effect relationships from these distributions [1]. A cy-

borg is a being whose human capabilities are enhanced by mechanical elements within their body. This includes technology to restore or augment functions, like a hearing aid improving hearing. By this definition, using technology, like relying on a smartphone for memory, gradually makes us more cyborg-like. Embracing tools like ChatGPT broadens this, enhancing our cognitive abilities beyond specific senses or skills. This blurs the line, leading some to mistakenly believe that conscious AI is imminent.

We proposed that automatically ascending the three rungs of the causation ladder is risky. But how exactly does relying on System 1 for causal inference lead to failure? Many evolutionary psychologists believe that System 2 faces a vast number of choices, leading to a combinatorial explosion. For instance, making 100 deliberate decisions in the first minute and another 100 in the second results in 10,000 possible combinations after just two minutes, and one million combinations after three minutes ($100 \times 100 \times 100$). Unlike System 2, a computer avoids this combinatorial explosion, as it is programmed for specific tasks, thus narrowing its decision-making scope [2].

System 1, shaped by evolution, comes with survival and reproduction programs, enabling spontaneous causation ladder climbing. Yet, System 2 causal inference is essential. The adapted mind, designed for Paleolithic survival and reproduction, is not tailored for truth-seeking. In today's world, truth-finding is crucial for survival. We must use System 2 to make causal inferences. This is necessary because many modern decisions were not present in our ancestral environment, leaving us without evolutionary training or data on their frequency. Additionally, we need to sift through the vast information from our independent modules (System 1) to prevent them from hindering sound causal inference.

The Turing test measures a machine's human-like intelligence by its ability to play an imitation game, where an average human interrogator cannot correctly identify the machine more than 70% of the time after five minutes of questioning. Pearl proposes a modified version, a mini-Turing test, focused specifically on causal reasoning as a benchmark for achieving strong AI [1]. For example, the mini-Turing test is resistant to deception by a mere list of scripted questions and answers, which are insufficient to replicate human intelligence. This is due to a combinatorial explosion: even a small set of variables can lead to an astronomical number of potential questions. To succeed in the mini-Turing test, machines must be equipped with an efficient environmental representation system and an effective answer-extraction algorithm [1]. The mini-Turing test was developed to enable computers to make causal inferences effectively. This also helps us understand the functioning of System 2's causal inference process. Pearl presents a causal diagram example that enables machines to successfully pass the mini-Turing test, thereby safely ascending the three levels of the causation ladder.

Imagine a prisoner facing execution by a firing squad. The court issues an execution order (O), which is relayed to a captain (C). The captain then signals soldiers A and B to shoot. Both soldiers are obedient and skilled marksmen; they

shoot only on command, and the prisoner dies (D) if either fires. Every variable (O, C, A, B, D) is binary, representing true or false.

Figure 21(a) displays a causal diagram addressing a first-rung causal query, focusing on association – how one fact informs us about another. Ascending the causation ladder, Figure 21(b) addresses an intervention question (second rung). Progressing to the third rung, Figure 21(c) explores a counterfactual scenario.

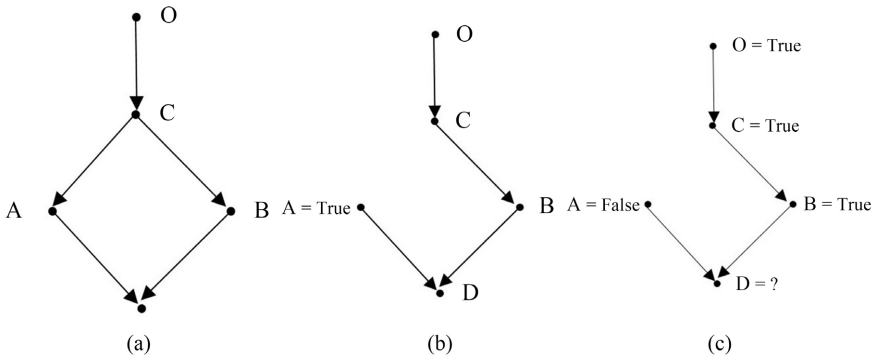

Figure 21. Causal diagrams for the firing squad scenario. (a) First level of causation (observation): A and B indicate the actions of soldiers A and B. (b) Second level (intervention): Soldier A chooses to fire; the link from C to A is removed, and A is set to true. (c) Third level (counterfactual): observing the prisoner's death, we question the outcome if Soldier A had not been fired.

Consider the query: if the prisoner is dead, does this mean a court order was issued? From Figure 21(a), a computer would deduce that the soldiers would only fire on the captain's command. The captain, in turn, would command only if he had a court order. Hence, the computer concludes the answer to the query is yes. In another scenario, if it is known that Soldier A fired, what does this imply about Soldier B? Using Figure 21(a), the computer would ascertain that Soldier B must have also fired, regardless of A's actions causing B's.

Next, consider the query: if Soldier A fires on his own, without the captain's command, is the prisoner dead or alive? The computer, using the causal diagram in Figure 21(a), finds this question unanswerable. To address this, we use Figure 21(b), teaching the computer the distinction between observing an event and affecting it. We tell the computer: "When you cause an event, eliminate all incoming arrows to that event and proceed with the analysis, ignoring those arrows as if they never existed". By erasing all arrows to variable A and setting A to true, the computer can now respond. It concludes that the prisoner is dead.

Finally, consider the query: if Soldier A chose not to shoot, would the prisoner still be alive? To analyze this, we introduce the computer to the counterfactual scenario in Figure 21(c)'s causal diagram. We remove the arrow to A, freeing it from C's influence, and set A to false, maintaining its real-world history. In this imagined scenario, the computer concludes that the prisoner would still die from Soldier B's shot, proving its ability to pass the mini-Turing test.

Why are counterfactuals on the third rung and interventions on the second

rung of the ladder of causation? Consider this scenario: we know the fire escape was blocked (X = 1) and Judy died (Y = 1). What is the probability that Judy would have survived (Y = 0) if the escape was not blocked (X = 0)? Without knowing the actual outcome (hindsight), $P\left(Y_{X=0}=0\right)$ and $P\left(Y=0 \mid do\left(X=0\right)\right)$ are identical. However, since a do-expression cannot be encapsulated $P\left(Y_{X=0}=0 \mid X=1, Y=1\right)$, counterfactuals are ranked higher than interventions in the causation ladder.

13. Concluding Remarks

The first rung of the causation ladder focuses on association, involving observation and queries like "what if I see…", "how are variables related", and "how would seeing X change my belief in Y?". The second rung centers on intervention, encompassing actions and questions such as "what if I do…", "what would Y be if I do X", and "how can I make Y happen?". The third rung addresses counterfactuals, characterized by imagination, retrospection, and understanding, asking "what if I had done…", "was it X that caused Y", "what if X had not occurred", and "what if I had acted differently?" [1].

Evolution through natural and sexual selection has honed our minds to intuitively navigate the rungs of observation, intervention, and imagination, akin to System 1 thinking. Humans surpass data in understanding causes and effects, as we automatically grasp them through our intuitive System 1 thinking, while data lacks this ability. However, our minds are tailored more for survival than for uncovering truth, leading us to be satisficers, not maximizers, as Herbert Simon noted. This concept from Simon [28] sparked the bounded rationality approach, which laid the groundwork for behavioral economics. Fast thinking, therefore, should not be relied upon for accurate causal inferences. Yet, by deliberately pondering over our innate causal inference abilities using System 2, we can safely and effectively ascend the causation ladder.

This paper emphasizes the importance of employing slow thinking for accurate determination of cause-and-effect relationships, and outlines methods for developing precise causal inference frameworks. It underscores the critical roles of System 1 and System 2 in understanding and navigating the complexities of causal inference. The paper delves into numerous instances where the interplay between these two cognitive systems either clarifies or complicates our grasp of causation. System 1, operating on intuition and rapid processing, often leads us to perceive causal relationships where none exist (type I errors), as exemplified by the coin flip experiment and attractiveness and personality in dating choices. It equates correlation with causation, which can be misleading in complex situations such as the Monty Hall problem. In contrast, System 2, with its slow and deliberate reasoning, is essential for accurately assessing probabilities and causality, especially in situations laden with cognitive illusions such as collider bias and paradoxes arising from conflicting statistical reasoning and causal intuition.

This review highlights the importance of employing both systems judiciously to navigate the landscape of causal inference, emphasizing the need for a ba-

lanced approach that harnesses the intuitive rapidity of System 1 and the analytical rigor of System 2. We propose using the "thinking, fast and slow" concept to enhance the creation of causal diagrams in everyday research. This method combines System 1's intuitive causal reasoning with the comprehensive causal scripts derived from System 2's causal inference. Ultimately, causal inference emerges as a quintessential System 2 technology, adeptly leveraging our innate System 1 tendency to discern causality, yet meticulously steering clear of the simplistic extremes of statistical inference that deny causality in favor of mere correlation.

Funding

This work is supported by CNPq [Grant number: PQ 2 301879/2022-2], and Capes [Grant number: PPG 001].

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Pearl, J. and Mackenzie, D. (2018) The Book of Why: The New Science of Cause and Effect. Basic Books.

[2] Da Silva, S. (2023) System 1 vs. System 2 Thinking. *Psychology International*, **5**, 1057-1076. https://doi.org/10.3390/psych5040071

[3] Taleb, N.N. (2010) The Black Swan: The Impact of the Highly Improbable. 2nd Edition, Random House.

[4] Neyman, J. and Pearson, E.S. (1933) The Testing of Statistical Hypotheses in Relation to Probabilities a Priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, **29**, 492-510. https://doi.org/10.1017/s030500410001152x

[5] Evans, J.S.B.T. (2008) Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, **59**, 255-278. https://doi.org/10.1146/annurev.psych.59.103006.093629

[6] Kahneman, D. (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux.

[7] Stanovich, K.E. (2004) The Robot's Rebellion. University of Chicago Press. https://doi.org/10.7208/chicago/9780226771199.001.0001

[8] Tooby, J. and Cosmides, L. (1992) The Psychological Foundations of Culture. In: Barkow, J.H., Cosmides, L. and Tooby, J. Eds., *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford University Press, 19-136. https://doi.org/10.1093/oso/9780195060232.003.0002

[9] Over, D.E. (2003) Evolution and the Psychology of Thinking: The Debate. Psychology Press.

[10] Buss, D.M. (2019) Evolutionary Psychology: The New Science of the Mind. 6th Edition, Routledge.

[11] Hume, D. (1739) A Treatise of Human Nature. Second Edition, John Noon.

[12] Dennett, D. (2017) From Bacteria to Bach and Back: The Evolution of Minds. W.W. Norton & Company.

[13] Hume, D. (1748) An Enquiry Concerning Human Understanding. Andrew Millar.

[14] Domingos, P. (2015) The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake our World. Basic Books.

[15] Tikka, S. and Karvanen, J. (2017) Identifying Causal Effects with the *R* Package Causaleffect. *Journal of Statistical Software*, **76**, 1-30. https://doi.org/10.18637/jss.v076.i12

[16] Heckman, J. and Pinto, R. (2014) Causal Analysis after Haavelmo. *Econometric Theory*, **31**, 115-151. https://doi.org/10.1017/s026646661400022x

[17] Imbens, G.W. and Rubin, D.B. (2015) Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press. https://doi.org/10.1017/cbo9781139025751

[18] Cunningham, S. (2021) Causal Inference: The Mixtape. Yale University Press.

[19] White, H. and Chalak, K. (2009) Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, **10**, 1759-1799.

[20] Morgan, S.L. and Winship, C. (2007) Counterfactuals and Causal Inference. Cambridge University Press. https://doi.org/10.1017/cbo9780511804564

[21] Kline, R.B. (2016) Principles and Practice of Structural Equation Modeling. 3rd Edition, The Guilford Press.

[22] Greenland, S. (2000) An Introduction to Instrumental Variables for Epidemiologists. *International Journal of Epidemiology*, **29**, 722-729. https://doi.org/10.1093/ije/29.4.722

[23] Bowden, R.J. and Turkington, D.A. (1985) Instrumental Variables. Cambridge University Press. https://doi.org/10.1017/ccol0521262410

[24] Wooldridge, J.M. (2019) Introductory Econometrics: A Modern Approach. 7th Edition, Thomson South-Western.

[25] Pearl, J. (2014) Trygve Haavelmo and the Emergence of Causal Calculus. *Econometric Theory*, **31**, 152-179. https://doi.org/10.1017/s0266466614000231

[26] Chen, B. and Pearl, J. (2013) Regression and Causation: A Critical Examination of Econometrics Textbooks. *Real-World Economics Review*, **65**, 2-20.

[27] Textor, J., Hardt, J. and Knüppel, S. (2011) DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*, **22**, 745. https://doi.org/10.1097/ede.0b013e318225c2be

[28] Simon, H.A. (1956) Rational Choice and the Structure of the Environment. *Psychological Review*, **63**, 129-138. https://doi.org/10.1037/h0042769