The Role of Causality in Making Connections Between Human Judgmental Bias and Machine Learning Bias

Wen Zhang

Department of Mathematics and Statistics
McGill University, Montreal

February 2021

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Science

Contents

Abstract

Research on human cognitive bias has a long tradition, and with the recent advancements in deep learning, we present a novel problem of whether machines make mistakes similar to human judgmental biases. It is worth noting that recent theoretical developments have revealed that even the most advanced deep learning algorithms have not acquired human-level intelligence, and causality is deemed to be a revolutionary power due to the fact that causal reasoning is a built-in capacity for humans. In this thesis, we perform a comprehensive literature review of causal inference methods in the structural and functional model framework as well as existing causal approaches to explaining both human judgmental bias and deep network vulnerability. As a preliminary investigation into this new area, our study is not sufficient to establish the connection between human bias and machine bias. We hope this thesis provides a good starting point for further interdisciplinary research on this problem.

Résumé

La recherche sur le biais cognitif humain est une longue tradition, et avec les récentes avancées sur l'apprentissage profond, on vous présente un nouveau problème, celui où les machines font des erreurs de façon similaire au biais du jugement humain. Il est intéressant de noter que les récents développements théoriques ont révélé que même les algorithmes d'apprentissage profond les plus avancés n'ont pas obtenus un niveau d'intelligence égal à l'humain, et le fondement est voué à devenir un pouvoir révolutionnaire dû au fait que le raisonnement causal est une capacité innée de l'humain. Dans cette thèse, nous faisons un résumé compréhensif littéraire des méthodes d'inférences causales dans les systèmes de modèles structurel et fonctionnel, comme les approches causales existantes, pour expliquer le biais du jugement humain ainsi que la profonde vulnérabilité du réseau. Comme c'est une enquête préliminaire dans ce nouveau domaine, notre étude n'est pas suffisante pour établir le rapport entre le biais humain et le biais machinal. Nous espérons que cette thèse vous fournira un bon point de départ pour une recherche interdisciplinaire plus approfondit sur ce problème.

Acknowledgments

I wish to express my sincere gratitude and appreciation to both of my supervisors, Professor Masoud Asgharian and Professor Louigi Addario-Berry, for their guidance, support and encouragement throughout my graduate study at McGill University. I would also like to thank my supervisors as well as the McGill Department of Mathematics and Statistics for the financial support of my graduate study. The year of 2020 is a challenging time for all of us, and this work would not be possible without the aforementioned people.

I am also grateful to my parents and friends for their love and support. In particular, I would like to thank Dongliang and Myriam for their professional suggestions, and Louis-Charles for the emotional support during this long-term quarantine.

Chapter 1

Introduction

In their theoretical work and pioneering investigation on cognitive psychology, Nobel Laureate Daniel Kahneman and his collaborator Amos Tversky introduce the notion of cognitive bias [KT72] based on numerous replicable empirical experiments they conducted. By definition, a cognitive bias is a systematic error that stems from heuristics and affects people's judgments. When making decisions under uncertainty, the events of interest are often perceived using terms such as odds or subjective probabilities. Accordingly, the human brain relies on prior beliefs to create "mental shortcuts" [Kah11], which reduce the complicated task of calculating probabilities into making intuitive estimations. However, Kahneman and Tversky [TK74] show that the probabilistic judgments produced with the aid of these "mental shortcuts" are often biased and are sometimes clear violations to the logical world.

Inspired by Kahneman and Tversky's work on human cognitive bias, we are interested in the bias made by machine learning (ML) algorithms and its potential connections with human cognitive bias. Among various ML models, the design of artificial neural networks (ANNs) was directly motivated by the architecture of biological neural networks in human brains. Although ANNs were initially created to learn and make decisions in a human-like manner, the resemblance

between biological and artificial neural networks is rather limited, especially after the change of focus of ANNs into solving task-oriented problems. Moreover, no study to date has elucidated how human brains work or how general intelligence systems reason and learn with high efficiency. Yet, to a certain extent, an ANN still captures the internal representation and functionality of biological neurons [Kri15], such as establishing connections between neurons, transferring signals and triggering neurons with activation functions. Therefore, for this study, it is of particular interest to investigate neural network models and the potential biases they lead to.

Recent advances in machine computational power have significantly boosted the development in deep learning (DL). For the past five years there has been a rapid rise in the use of DL models to perform complex predictive tasks, and deep neural networks (DNNs) often outperform classic ML algorithms in many domains such as image recognition, natural language processing (NLP) and chess game. Nevertheless, one of the main issues regarding our knowledge of DNNs is a lack of understanding of how machines come to certain decisions. Much like biological neural networks, DNN models are perceived as black boxes that don't exhibit human-comprehensible decision rules, and the relative influence of an input feature on the model prediction remains hidden under the network architecture. This black-box issue is often described as the "interpretability" problem [GBY+18] in ML community, and it makes troubleshooting and investigating machine biases even more perplexing. In his critique of explainable artificial intelligence (XAI), Miller [Mil19] defines the interpretability of a ML model as "the degree to which an observer can understand the cause of a decision". This points us towards the interpretation of ML models from a causal perspective. It is believed by many that causality is the key to explaining the vulnerability of both human and machine intelligence.

This thesis begins by reviewing some seminal works in cognitive psychology

and discussing recent advances in causality. These studies motivate us to examine machine learning bias from the causal perspective and to identify underlying connections between machine learning bias and human cognitive bias. In each of the following chapters, we investigate various aspects of causal inference and machine learning. It is worth pointing out that these theoretical components are not explicitly connected to each other, but we believe that together they facilitate studies on machine learning bias.

This chapter is organized as follows. In Section 1.1, we review some typical psychological studies on human cognitive bias from a probabilistic point of view and discuss the existing causal explanations for cognitive bias. In Section 1.2 we introduce existing studies on causality conducted by researchers from two different domains and compare the different aspects they focus on. Section 1.3 gives a brief overview of the future of machine learning that is powered by theories of causation.

1.1 Selective Literature Review of the Probabilistic Aspects of Human Judgmental Bias

This section introduces some typical judgmental biases in human cognition. The pervasiveness of these illusions reveals that humans are not good at making intuitive statistical judgments. In their ground-breaking paper of 1974 [TK74], Tversky and Kahneman note that even professional researchers with years of experience are victims of cognitive bias. Although knowledge and heuristics endow people with the ability to think and reason rationally, they are relatively weak in reducing systematic errors when people make decisions under uncertainty.

1.1.1 The Problem of Small Sample Size

One typical bias observed by Kahneman and Tversky [TK74] is people's "insensitivity to sample size". When a survey is conducted, it is always crucial to have a sample size that is sufficiently representative of the population of interest. Kahneman [Kah11] provides an example considering the "small schools movement" in the U.S. funded by the Gates Foundation. Results from many related studies [Mei02][HS03] indicate that students in smaller schools perform better than students in larger schools. However, Wainer et al. [WZ06] point out that most of the worst-performing schools are also small, and the causal relationship between small schools and better performance is illusory. This example shows that small samples are more likely to produce extreme results in more than one direction, and it is simply because small samples are often not representative of the population.

This problem of small sample size is often overlooked by people, as suggested by a number of psychological experiments in [TK74], and Kahneman highlights two underlying factors: one is that people tend to be oblivious to the credibility of information and focus on only the content; the other is that people often try to explain observed patterns using cause-and-effect relationships, while in many cases causation should not be applied.

1.1.2 Irregularity and Randomness

Another illusion stemming from people's causal thinking is the expected association between irregularity and randomness. The formal definition of randomness has been proved to be difficult [Nic02][Bel99], but when we observe a finite sequence of events that seems non-uniform, it is intuitive to infer to some extent that the events are random. On the other hand, upon observing a sequence that did not seem random, participants in Kahneman et al.'s experiments [TK74] found it difficult to accept that the events were drawn from a random process.

A very famous example regarding this intuitive belief is the gambler's fallacy, also known as the Monte Carlo fallacy. In 1913, during a roulette game at the Monte Carlo casino in Monaco, the ball fell in black 26 times consecutively. Within this sequence of uniform results, lots of gamblers believed that the next outcome would be red, since the previous outcomes seemed extremely unbalanced. However, their misconception of randomness led to them losing more money.

Based on simple calculation and the assumption that each row is independent, it is straightforward to verify that both sequences of events in Table 1.1 have the same probability, which is 0.5^{10}, but only the first sequence seems random. This seemingly balanced sequence implies local representativeness of a random process, and the local randomness is often directly perceived as randomness of the entire process [Nic02]. The participating gamblers failed to realize that no matter what the previous outcomes looked like, the next draw will be completely independent of the previous outcomes, and the colour red and black will always have the same probability to occur.

Balanced	B	R	B	B	R	B	R	R	B	R
Unbalanced	B	B	B	B	B	B	B	B	B	B

Table 1.1: Table to "balanced" and "unbalanced" outcomes (B: black, R: red)

To relate the above example to probability theory, when analyzing a random process, it is worth noting that irregularity and true randomness in outcomes need to be viewed in different scopes. By the law of large numbers, the overall outcome of a random process will get infinitely close to a balanced result, hence the tendency of irregularity is ensured. What leads to the above cognitive bias, as explained by the law of small numbers [Bel99], is the belief that this irregularity should be observed in small samples as well. For small samples, extreme distributions are more likely to be observed (Section 1.1.1), and the unbalanced pattern in the outcomes of a random process should not be surprising at all.

As a counterpart of the gambler's fallacy, the hot-hand fallacy was first introduced by Gilovich, Vallone and Tversky [GVT85] in 1985. It originates from these researchers' investigation into the well-known phenomenon of "streak shooting" in basketball games. Essentially, a player who has scored continually within a short period of time are often considered to be on a streak. As a result, this player will receive more pass from his teammates and more defence power from the opposing team, since everyone believes that there is a good chance that this player will score again shortly. Interestingly, unlike the gambler's fallacy, in which people predict that future outcomes will be inverted due to the previous extreme results, people in this case believe that the prior pattern will continue, and it is again difficult for them to accept that the overall process is random.

Contrary to many basketball coaches who believe the hot-hand phenomenon, Gilovich's team [GVT85] argue that the hot-hand phenomenon is merely a cognitive illusion and another example of human's intuition leading to the misconception of local representativeness. In an effort to interpret the coexistence of these two conceivably contradictory fallacies, Gilovich et al. [GVT85] suggest that the false belief of the law of small numbers explains them both. In the gambler's fallacy, people expect the outcomes to be balanced as randomness is presumed for the entire process. While in the hot-hand fallacy, there is no assumption of randomness. By assuming the local representativeness of the streaking pattern, it becomes intuitive for people to conclude that the scorings are dependent and are hence no longer from a random process.

More recent evidence indicate, however, that the explanation using the law of small numbers is not conclusive. In his book on rational thinking, Gigerenzer [Gig00] shows that one single rule is not adequate for explaining two contradicting phenomena at the same time. This has led authors such as Ayton [AF04], Burns [Bur01] and Huber [HKS10] to investigate into the difference between the gambler's fallacy and the hot-hand fallacy. A more acceptable explanation has been proposed

[AF04] that, people tend to overestimate the influence of human performance. This conclusion by Ayton better acknowledges the intuitive belief of causal relationship between the player's performance and the hot streak, and it also helps address the limitation of the representativeness conjecture.

1.1.3 Base Rate Fallacy

Another paradoxical result that has been outlined in Kahneman and Tversky's preliminary work [TK74] is the base rate fallacy. The authors' experiments [*ibid.*, Section 1.1] have revealed that when trying to assess the probabilities of two disjoint events, subjects who were not given extra information tend to utilize the base rates of the two events properly. However, when given extra information that potentially supports one of the two events, subjects perceived this event to be more likely, even when the base rates suggested otherwise.

In Kahneman and Tversky's experiment, participants were given the following information [Kah11]:

> An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail."
>
> Is Steve more likely to be a librarian or a farmer?[a]
>
> [a] *Source:* Thinking, Fast and Slow by Daniel Kahneman

Their result shows that almost all the participants chose "librarian", although the number of male farmers is more than 20 times larger than that of male librarians in the U.S.. The description of Steve matches the stereotype of a librarian, and its predictive power is obviously overestimated by the participants. The authors suggest that the prior probability, or the base rate, is neglected as a result of applying the representativeness heuristic.

To articulate how the base rate is intuitively neglected, a toy probabilistic problem can be constructed from the above experiment under certain assumptions. Suppose that 2% of the U.S. male population are farmers, while librarians constitute 0.1% of the male population. Among the librarians, 80% of them are shy and withdrawn, hence the stereotypical impression of the personality of librarians prevails. On the other hand, we assume that 50% of the farmers have similar introverted personalities described in the former statement. For those who are neither a farmer nor a librarian, 50% of them have the described personalities. These assumptions are consistent with the fact that the description is closely associated with a stereotypical librarian, but not a farmer.

Let F be the event that Steve is a farmer, let L be the event that Steve is a librarian, and let N be the event that Steve is neither a farmer nor a librarian.

Let *stereotype* be the event that Steve is said to have a stereotypical personality of a librarian as described in the previous statement.

Integrating the assumptions we have that,

$$P(F) = 0.02 \qquad P(L) = 0.001 \qquad P(N) = 0.979$$

$$P(stereotype \mid F) = 0.5 \quad P(stereotype \mid L) = 0.8 \quad P(stereotype \mid N) = 0.5$$

Using Bayes' Theorem,

$$\begin{aligned} & P(F \mid stereotype) \\ &= \frac{P(stereotype \mid F) \times P(F)}{P(stereotype \mid F) \times P(F) + P(stereotype \mid L) \times P(L) + P(stereotype \mid N) \times P(N)} \\ &= \frac{0.5 \times 0.02}{0.5 \times 0.02 + 0.8 \times 0.001 + 0.5 \times 0.979} \\ &\approx 0.020 \end{aligned}$$

$$
\begin{aligned}
& P(L \mid stereotype) \\
&= \frac{P(stereotype \mid L) \times P(L)}{P(stereotype \mid F) \times P(F) + P(stereotype \mid L) \times P(L) + P(stereotype \mid N) \times P(N)} \\
&= \frac{0.8 \times 0.001}{0.5 \times 0.02 + 0.8 \times 0.001 + 0.5 \times 0.979} \\
&\approx 0.0016
\end{aligned}
$$

The above calculation shows that given Steve has the stereotypical personalities of a librarian, it is much more likely that Steve is a farmer than a librarian. More recent evidence [GH95][HLHG00] suggest that people make more reasonable inference when provided with probabilistic information such as *P(L)*, *P(stereotype|L)* instead of textual information. This provides a possible solution to address the base rate neglect problem in human judgments.

1.1.4 Conjunction Fallacy

Let A and B be two events, which are not necessarily independent. According to the laws of probability, we have

$$P(A \cap B) = P(A) \times P(B \mid A) \tag{1.1}$$

$$P(A \cap B) = P(B) \times P(A \mid B)$$

Since $0 \leqslant P(B \mid A) \leqslant 1$ and $0 \leqslant P(A \mid B) \leqslant 1$, it is straightforward to verify that

$$P(A \cap B) \leqslant P(A)$$

$$P(A \cap B) \leqslant P(B)$$

These two simple inequalities essentially state that the probability of a conjunction of two events cannot exceed the probability of either single event. How-

ever, this conjunction rule is easily violated when people make predictions under intuitive heuristics, and even professionalists have fallen for this trap.

In [TK83] Kahneman and Tversky performed a study involving 115 professional analysts during the Second International Congress on Forecasting in 1982. The participants were divided into two groups. One group was asked to predict the probability that the U.S. and the Soviet Union will terminate their diplomatic relations soon. The other group was asked to assess the probability that Russia will invade Poland and that the U.S. and the Soviet Union will terminate their diplomatic relations soon. Despite that the resulting probabilities from both groups were low, the estimates from the second group were significantly higher than those of the first group, with p-value < 0.01.

It is clear that the conjunctive event from the second group has a lower probability, i.e.,

$$\begin{aligned} & P(\textit{Russia invades Poland} \cap \textit{Suspension of diplomatic relations}) \\ \leqslant\ & P(\textit{Suspension of diplomatic relations}). \end{aligned}$$

Several possible explanations concerning this error have been provided by Kahneman and Tversky [TK83]. First, the event that Russia invades Poland was perceived as a possible cause of the termination of diplomatic relations. This enhanced the causal and correlational beliefs [JAR82][CC67] of the conjunctive event, which therefore appeared to be more representative than any single event. Second, by adopting the availability heuristic [Kah11], people often retrieve with ease the probability of a conditional event $P(\textit{effect} \mid \textit{cause})$ rather than the joint probability $P(\textit{cause} \cap \textit{effect})$, while it can be easily verified from (1.1) that $P(\textit{cause} \cap \textit{effect}) \leqslant P(\textit{effect} \mid \textit{cause})$. Consequently, $P(\textit{effect} \mid \textit{cause})$ has an anchoring effect [Kah11] on the estimation of $P(\textit{cause} \cap \textit{effect})$, causing the latter to be erroneously high.

1.1.5 Causal Interpretations of Human Judgmental Biases

Following their initial work on human judgmental bias, Tversky and Kahneman [TK77] further investigate into their heuristics and biases theory from a causal perspective. They note that people's judgments of probabilities are somewhat affected by causal reasoning. In particular, their subsequent psychological experiments on base rate neglect [TK15] show that data with more causal significance are usually overweighed by human subjects, and that [*ibid.*, p. 57] "causal inferences have greater efficacy than diagnostic inferences". However, there is still considerable ambiguity in their research with regard to how causal interpretations are associated with probabilistic inference. The authors claim that people merely rely on causality as an intuitive heuristic rather than a rational inference paradigm.

A more recent study by Krynski and Tenenbaum from MIT [KT07] provides an alternative framework to interpret people's judgments under uncertainty. In contrast to Kahneman and Tversky's heuristics and biases view that is based on traditional statistical inference, Krynski and Tenenbaum suggest the use of Pearl's structural causal model [Pea00] for the representation of people's cognitive inference and decision process. Their study is motivated by the fact that traditional statistical inference only works well in an ideal case with a small number of variables and sufficient amounts of observational data, while most real-world judgmental tasks involve many causally related variables with a limited number of observations. Several experiments have been conducted by Krynski and Tenenbaum [KT07], and the results suggest that people's judgments under uncertainty are guided by the causal models they construct intuitively. More importantly, the authors' experiments reveal that erroneous probabilistic judgments often occur when the statistics provided in psychological experiments do not fit in the intuitive causal models conceived by subjects, and that these judgmental biases can be reduced if a clear mapping from the statistics to the underlying causal structure

is available to people.

1.2 Comparative Analysis of Research Interests in Current Literatures on Causation

To establish the connection between human cognitive bias and machine bias, it is of interest to investigate whether machines have similar learning and reasoning mechanisms as humans do. In Section 1.1.5, it has been demonstrated that humans are able to think and reason in terms of cause and effect by instinct, while most of the popular ML algorithms are mere exploitations of data correlation. In recent years, there are growing appeals for applying causal inference to ML studies. In fact, a number of works [Pea19][Sch19] have suggested that causal modeling has potential to take ML developments to the next level. Before proceeding to details of some important work that have been strongly endorsed, we first examine the difference between the main themes of research on causation in two communities that are both intrigued by the implementation of causality in their own fields of specialization.

Statisticians, econometricians and epidemiologists

On the one hand, statisticians, econometricians and medical researchers are generally interested in distinguishing between statistical association and causation. In his philosophical analysis of causality, Hume [Hum00][Hum03] suggested three criteria for examining a presumed causal relationship, including contiguity in space and time, temporal precedence (the cause must precede the effect in time) and covariation. Although these criteria have been proved to be insufficient [Hol86], they remain one of the early contributions to the probabilistic theories of causation.

One of the first systematic studies on the connection between probabilistic correlation and causation was carried out in 1956 by Reichenbach [Rei56]. Reichen-

bach proposed the *Common Cause Principle* (RCCP), which states that if two variables X and Y are statistically correlated, that is, if $P(X \cap Y) \neq P(X) \cdot P(Y)$, then either there exists a causal relationship between X and Y, or there exists another variable Z, which is a common cause of X and Y. The RCCP highlights the difficulty of distinguishing causation from association (statistical dependence) in the bivariate case, as it cannot be determined from the observational data of X and Y whether X causes Y, Y causes X, or there exists a third variable Z causing X and Y [Sch19]. Based on the RCCP, Reichenbach introduced [*ibid.*, p. 157] the notion called screening off. Basically, a variable Z is said to screen X off from Y if $P(Y \mid X, Z) = P(Y \mid Z)$. Inspired by the second law of thermodynamics, Reichenbach proposed [Hit97] three sufficient necessary conditions for inferring that an event X causes Y: (i) temporal precedence of X, (ii) X and Y are positively associated, i.e. $P(Y \mid X) > P(Y \mid \neg X)$, (iii) there does not exist an event Z preceding or happening simultaneously with X, such that Z screens X off from Y.

In the same decade of 1950s, there has been a well-known debate among researchers on whether associational data can provide evidence for the claim that smoking causes lung cancer, most notably, the studies by Doll and Hill [DH50] [DH52] and the corresponding criticisms from Fisher [Fis58a] [Fis58b]. It has later been pointed out in Holland's comment [Hol86] that Fisher's argument is an example of "confusion between attributes and causes", as Fisher hypothesized a certain genetic attribute that causes the smoking behaviour among subjects. In a major contribution to public health studies following his initial work on the causal relationship between smoking and lung cancer, Hill [Hil55] [Hil65] highlights nine criteria for identifying causations from associations in observational studies, although there has been some disagreement regarding whether Hill's criteria are applicable in certain areas [PG06] [War09].

In econometrics, Granger [Gra69] suggests a probabilistic notion of causation

for time-series data, while the use of time-series data is simply equivalent to the temporal precedence condition in Hume's and Reichenbach's theories. In Cox's analysis on causality [Cox92], Granger's notion of causation is considered as "a statistical association that cannot be explained as in fact a dependence on other features". Holland [Hol86] restates Granger's causation as a more generalized notion based on conditional independence, that X is a Granger cause of Y if Y is dependent on X when conditioned on another variable Z. Holland also points out the limitation of Granger's causation, that a Granger cause can become a spurious cause if Z is changed.

It is worth noting that most statisticians and econometricians who have been investigating causation based on associational data assume that the precursors are already known in the data, and the temporal precedence condition must therefore be satisfied to infer causation in traditional statistical analysis. This limitation has been addressed by Pearl and Verma [PV95], who have presented the inductive causation (IC) algorithm. Based on the model minimality assumptions powered by the principle of Occam's razor, the IC algorithm can identify the simplest pattern of conditional independences from observational data, and hence distinguish causal relationships from spurious associations with no need of extra chronological information.

Computer scientists and machine learning researchers

On the other hand, computer scientists, especially experts in machine learning, have been investigating whether causal inference can be implemented in machine algorithms. Their studies on causality have greater emphasis on determining whether X causes Y or Y causes X, that is, the problem of identifying causal directions from empirical data pairs. This has been viewed by many as a rather challenging problem, because there is no third variable that provides information on conditional independences. Various approaches have been proposed to solve

this problem by finding patterns of asymmetry between the two variables.

First, many researchers have suggested the use of functional causal models (FCMs) to simulate the underlying data-generating process from cause to effect. Essentially, an FCM is formulated as [GZS19]

$$Y = f(X, \epsilon; \boldsymbol{\theta}), \tag{1.2}$$

where the effect Y is generated from a function $f \in \mathcal{F}$ of the cause X and a noise term ϵ that is independent of X, and $\boldsymbol{\theta}$ is the set of parameters of f. By fitting the FCM to both candidate directions while assuming no hidden confounders, the direction that admits the independence between X and ϵ is the plausible causal direction. To ensure the uniqueness of the independence $\epsilon \perp\!\!\!\perp X$ in only one direction, Hyvärinen et al. [HP99] highlight that further assumptions need to be made on f. Shimuzu [SHHK06] proposes an acyclic model (LiNGAM) $Y = f(X) + \epsilon$, which assumes f is linear and ϵ is an additive noise term that has a non-Gaussian distribution with non-zero variance. Shimuzu's LiNGAM approach does not require the assumption of temporal precedence, and it guarantees the uniqueness of causal direction based on Hyvärinen's independent component analysis theory [Hyv99]. However, there exist causal relationships that are nonlinear in the physical world, and the linear additive noise models do not have the flexibility to model these data generating processes. Zhang and Hyvärinen [ZH10] [ZH12] develop a more generalizable model called post-nonlinear (PNL) causal model $Y = f_2(f_1(X) + \epsilon)$, which assumes f_1 is nonlinear, and f_2 is nonlinear and invertible. Another approach that accounts for nonlinear causal relationships is the nonlinear additive noise model $Y = f(X) + \epsilon$ developed by Hoyer et al. [HJM$^+$09]. It is considered a special case of the PNL model since it is equivalent to the PNL model with f_2 being an identity function. Stegle et al. [SJZ$^+$10] propose a novel approach called probabilistic latent variable model. Their model $Y = f(X, \epsilon)$ assumes that the

error term ϵ is unobserved, is not necessarily additive and has a standard normal distribution. Furthermore, Stegle's latent variable model does not make assumptions on the function f, and it instead uses a non-parametric Bayesian approach to determine the plausible causal direction. All of the aforementioned approaches are designed for continuous-valued observational data, while Peters et al. [PJS11] show that additive noise models can also be applied to discrete data, in which X and Y either are integer values or have finitely many states.

Second, an increasing number of studies have found that algorithmic information theory (Kolmogorov complexity) can help reveal the asymmetry between X and Y as well. Particularly, in [JS10] and [LJ13] the authors hypothesize that a necessary condition for X causing Y is that P_X and $P_{Y|X}$ are algorithmically independent. In other words, the information on P_X does not lower the algorithmic entropy of $P_{Y|X}$ and vice versa. Based on this notion of algorithmic independence, Janzing et al. [JMZ$^+$12] propose the information-geometric method called IGCI for determining if X causes Y or vice versa. The IGCI algorithm is designed for the deterministic case ($Y = f(X)$) without a noise term, and f is assumed to be bijective with no explicit assumption on its class. Vreeken [Vre15] reported on a data-centric method called ERGO, which detects the asymmetry between X and Y by approximating Kolmogorov complexities from data, and the author shows [*ibid.*, Section 6.2] that ERGO also works well when X and Y are high-dimensional.

1.3 Causality and the Future of Machine Learning

The past decade has witnessed numerous breakthroughs in the field of machine learning (ML), whereas researchers' opinions on whether machines can achieve human-level intelligence are divided. In an effort to closely investigate the intelligence that a cognitive system possesses, Pearl [Pea19] summarizes a conceptual causal framework with 3 levels classified by different capabilities of answering

queries. Pearl [*ibid.*, Section 2] suggests that a system is capable of answering questions from level i ($i = 1, 2, 3$) only if it has acquired information from level j, where $j \geqslant i$.

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief inY?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

Figure 1.1: From Figure 1 of [Pea19] (permission to reuse in a thesis obtained from the Licensed Content Publisher ACM(Association for Computing Machinery)), the three-level causal hierarchy.

The bottom level of Pearl's framework is Association, which involves only observational data. The queries essentially focus on inference of statistical relationships using the language of conditional probabilities, such as $P(y \mid x) = p$. Pearl [PM18] argues that ML algorithms, including most of the well-developed deep neural networks in recent years, are still restricted to this level. Admittedly, adding more layers to a neural network enhances the model flexibility, so that the model can fit more complicated nonlinear data. However, the learning process is solely data-centric. No matter how complex the deep learning architecture is, only correlational relationships can be learned from the passively collected data.

The second level is Intervention. It enables inference based on experimental data with the help of the do-operator $do(\cdot)$ proposed by Pearl [Pea00]. A typi-

cal expression used at this level is $P(y \mid do(x), z)$ [Pea19], which represents the probability of $Y = y$ given that X is set to the value x by external intervention and an observation $Z = z$ is obtained afterwards. Clearly, without information on designed manipulations, the associational observations from the first level cannot help answer interventional queries, no matter how much observational data are available.

The top and most powerful level is Counterfactuals. Pearl [Pea19] claims that a cognitive system at this level is capable of imagining a fictitious world consisting of events that have not happened in reality. A representative counterfactual query is $P(y_x \mid x', y')$ [*ibid.*, Section 2], which represents the probability that $Y = y$ would have happened if $X = x$ had been observed, while in reality there are only observations of $Y = y'$ and $X = x'$. Contrary to those at the Intervention level where data can be obtained by taking actions or changing the current settings, systems at the Counterfactuals level do not benefit from previous data or information to be obtained in the future. Instead, Pearl suggests [*ibid.*, Section 2] that retrospective reasoning and the semantics of Structural Causal Models (SCM) are the essential components for solving counterfactual problems. In Section 2.1, a detailed investigation into SCM is provided.

Having established that the above causal framework is directional and hierarchical, Pearl [PM18] points out that contemporary machine intelligence unfortunately remains at the bottom level of association. Most machines operate based on observational data and the logic of probability, while humans are capable of comprehending all three levels of causality in Figure 1.1. Despite numerous breakthroughs in deep learning in the past two decades, the underlying limitations of ML algorithms have been discussed by a great number of authors in literature. First, the existing algorithms have many problems in robustness. A typical and nearly omnipresent problem of overfitting reveals the very nature of any ML models, that a model may have a perfect fit to the existing samples, while it generalizes poorly

to unseen data points. Admittedly, techniques such as k-fold cross validation [MT88] and regularization [Tik63] can reduce overfitting, and they seem to provide evidence that machines are not merely optimizing with the current instances. However, these learning systems fail to rise up to the next level of intervention, as they make no inference for the unseen data. Second, many well developed ML models remain black boxes [RSG16][SWM17] and lack interpretability [Rud19]. This impediment has caused problems in criminal justice [ALMK16], healthcare systems [VBC18] and many other fields, where potentially biased results are hard to troubleshoot due to people's lack of understanding of the decision-making process of machine algorithms. Lastly, machines are generally designed to understand correlational relationships but not causal ones. In his overview of concerns for deep learning, Marcus [Mar18] mentions an example regarding the possible association between height and vocabulary across a general children's population. ML models can easily learn a positive correlation between these two variables, since a child normally grows in height while also acquires new vocabulary. However, if a person is to investigate into the causal relationships between height and vocabulary, it is obvious that neither of them acts as a cause of the other. Instead, a latent variable, which is also a confounding factor in this example, can be concluded. A straightforward assignment of this latent variable is a child's physical and mental growth, as it affects both height and vocabulary. Most machines, nevertheless, cannot distinguish cause from effect, let alone conceive a latent variable that acts as a common cause of the two observables. This limitation of identifying cause and effect is believed to be a major barrier for machines to achieve human-level intelligence [PM18][LST15]. In Pearl's exact words [Pea18], "human-level AI cannot emerge solely from model-blind learning machines; it requires the symbiotic collaboration of data and models". Within the next decade, theories of causal inference are destined to become an important component in machine learning studies.

In the next chapter, we present a comprehensive literature review of causal inference and its applications that are pertinent to machine learning problems. Chapter 3 examines a particular example of selection bias in medical studies and discusses the limitation of a typical causal discovery model. In Chapter 4, we first look at machine learning bias and review some well-known adversarial attacks on deep neural networks. This is followed by a review of some most recent causal approaches to improving the robustness of machine learning algorithms as well as a discussion on the possible connections between machine and human intelligence.

Chapter 2

Selective Literature Review of Structural Causal Modelling Methodologies

It is generally accepted that the gold standard for inferring cause and effect relationships is Randomized Controlled Trials (RCT). In an RCT, subjects are divided into different groups of an experiment at random, and the effect of the experimental intervention is assessed in the end. The randomization process aims to reduce many types of bias that are usually not controllable in other study designs, and more generally, to reduce confounding. Yet, it is often not feasible or ethical to conduct certain types of experiments. For example, recall the debate between Fisher and Hill et al. (Section 1.2), if a research group hypothesizes that smoking causes lung cancer, they cannot simply conduct an RCT and assign a group of subjects to smoke for several years. On the other hand, it is more viable to collect observational data on subjects with past or present smoking history, and the subjects' medical records can be abstracted to analyze the effect of their smoking behaviour. As Cochran [Coc72] points out in his commentary, causal inference based on observational data aims to approximate the process of randomized trials. In

scientific researches nowadays, inferring causation from purely observational data has become a central task, and many recent efforts have been devoted to tackle the corresponding issues, such as confounding.

The main objective for this chapter is to present a comprehensive review of causal inference methodologies under the structural causal model (SCM) framework.

2.1 Introduction

Suppose for an observational study involving n variables $X_1, ..., X_n$, the joint distribution $P_{\mathbf{X}} = P_{X_1,...,X_n}$ is observed. An SCM is a nonparametric model that summarizes causal relationships of these n variables into a graphical model. Some preliminary work [SGS93] on SCMs was carried out in the mid 1990s, and Pearl [Pea00] was among the first to formalize causality using the causal Bayesian network [Pea95] he proposed.

Definition 2.1 (Structural causal models)**.** *A structural causal model (SCM)* $M := (\mathbf{X}, \mathbf{N}, \mathbf{F})$ *contains a set* $\mathbf{F}$ *of d structural functions*

$$X_j = f_j(\mathbf{PA}_j, N_j), \qquad j = 1, ..., d, \tag{2.1}$$

where $\mathbf{X}$ *is the set of observable variables in the model,* $\mathbf{PA}_j \subseteq \{X_1, ..., X_d\} \backslash \{X_j\}$ *is the set of parents of* X_j*, and* $\mathbf{N} = N_1, ..., N_d$ *is the set of the hidden noise terms that accounts for any unexplained factors influencing the observable variables.*

The set $\mathbf{N}$ contains exogenous variables, which means they are not caused by any observable variables in the model. The variables in the set $\mathbf{X}$ are endogenous, meaning they must be a descendant of at least one exogenous variable. For any endogenous variable X_j in an SCM, it is possible that there are multiple contributing exogenous variables. Thus, a noise variable N_j can be a vector of variables

$N_j = \langle Y_1, ..., Y_k \rangle$, given that $Y_1, ..., Y_k$ are all the unobserved causes of X_j.

For every SCM, a graphical causal model can be constructed by translating the functions in $\mathbf{F}$ into nodes and edges. To discuss the properties of graphical causal models in detail, it is imperative to first introduce some basic terminologies associated with graphs.

A *graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ contains an index set $\mathbf{V} \coloneqq \{1, ..., d\}$ denoting the d nodes that correspond to random variables $\mathbf{X} = (X_1, ..., X_d)$, and $\mathcal{E} \subseteq \mathbf{V}^2$ is a set of edges between the nodes, where $(v, v) \notin \mathcal{E}$ for any $v \in \mathbf{V}$. Let i and j be two nodes in a graph $\mathcal{G}$, i.e. $i \in \mathbf{V}$, $j \in \mathbf{V}$, then i and j are *adjacent nodes* if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$. The edge between two adjacent nodes i and j is called an *undirected edge* if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$. Accordingly, a *directed edge* is an edge between two adjacent nodes that is not undirected. A *directed path* is a path between two nodes i_1 and i_m if $i_k \to i_{k+1}$ for all $k < m$. If a graph $\mathcal{G}$ has no directed path from any node $i \in \mathbf{V}$ to the node i itself, then $\mathcal{G}$ is called an *acyclic graph*. An acyclic graph $\mathcal{G}$ is called a *directed acyclic graph (DAG)* if all its edges are directed.

For the graphical causal model $\mathcal{G}$ of any SCM, if we assume that there is no bidirectional causal relationships (e.g. poverty and lack of education can potentially cause each other), then it is straightforward to verify that $\mathcal{G}$ is a DAG. First, $\mathcal{G}$ is directed since the edges denote causal relationships, which by definition are directional from cause to effect. Second, since a variable cannot be the cause of itself, $\mathcal{G}$ cannot contain any cycles, then $\mathcal{G}$ must be acyclic.

Example 2.1. *Consider an SCM M with $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$ and $\mathbf{F}$:*

$$X_1 \coloneqq f_1(X_3, N_1), \qquad X_2 \coloneqq f_2(N_2)$$
$$X_3 \coloneqq f_3(X_2, N_3), \qquad X_4 \coloneqq f_4(X_3, X_6, N_4)$$
$$X_5 \coloneqq f_5(X_4, X_6, N_5), \qquad X_6 \coloneqq f_6(N_6)$$

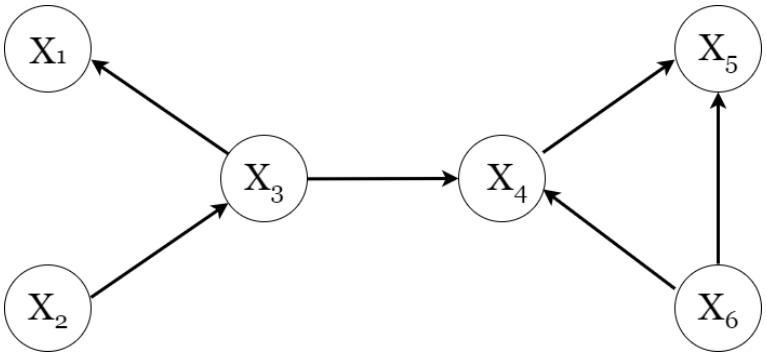

Figure 2.1: Corresponding causal diagram $\mathcal{G}$ of Example 2.1.

The associated graphical model $\mathcal{G}$ is shown in figure 2.1. The noise variables $N_1, ..., N_6$ are not included in the model, as they remain unmeasurable factors, and we simply accept their existence without investigating their further causes. An arrow pointing from X_i to X_j indicates that X_i is a direct cause of X_j. The assignments in $\mathbf{F}$ are also called structural equations.

SCMs are especially important in causal reasoning, as they can be utilized in all three levels of the causal hierarchy (Figure 1.1), while common probabilistic models only entail observational distributions from the level of Association.

Furthermore, an observed joint distribution is not sufficient to identify an SCM, since an SCM contains interventional and counterfactual information, while a joint distribution only contains associational information (Section 1.3). For observational studies involving multivariate data, a fundamental problem of causal discovery is to infer causal relationships among the variables based on purely observational data. Using the representation of SCM, this inference task is equivalent to identifying the underlying causal DAG $\mathcal{G}$ based on the observed joint distribution $P_{\mathbf{X}}$, and it relies on a fundamental assumption called causal Markov condition.

Postulate 2.1 (Local Markov condition)**.** *A DAG $\mathcal{G}$ with n variables $X_1, ..., X_n$ as nodes is considered as a possible causal structure only if every variable X_j is statistically independent of its non-descendants when conditioned on the parent variable(s) of X_j in $\mathcal{G}$, i.e. for every node X_j, we have $X_j \perp\!\!\!\perp X_k \mid \mathbf{PA}_j$ for every*

X_k that is not a descendant of X_j.

In [Lau96] the author shows that this assumption is equivalent to the factorization of the joint distribution $P_{\mathbf{X}}$ into n conditionals:

$$P_{\mathbf{X}}(X_1, ..., X_n) = \prod_{j=1}^{n} P(X_j \mid \mathbf{PA}_j) \tag{2.2}$$

Consider a DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ with n variables $X_1, ..., X_n \in \mathbf{V}$ as nodes. A node X_i is defined as a *collider* when there exist $X_j, X_k \in \mathbf{V}$ such that $X_j \rightarrow X_i \leftarrow X_k$; otherwise X_i is called a *non-collider*. A path is *blocked* by a set of variables $\mathbf{Z}$ if there exists a variable X_i on this path such that either X_i is a collider with no descendants in $\mathbf{Z}$, or X_i is a non-collider in $\mathbf{Z}$.

Definition 2.2 (d-separation)**.** *Given that $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are three disjoint sets of variables, $\mathbf{X}$ and $\mathbf{Y}$ are said to be d-separated by $\mathbf{Z}$ (denoted as $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$) if the paths between any $X_i \in \mathbf{X}$ and any $X_j \in \mathbf{Y}$ are all blocked by $\mathbf{Z}$.*

Based on the notion of *d*-separation, Lauritzen [Lau96] shows that the causal Markov condition is equivalent to the *global Markov condition*:

Postulate 2.2 (Global Markov condition)**.** *Given that $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are three disjoint sets of variables, if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$, then $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, i.e. $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$.*

Clearly, the causal Markov condition determines whether a DAG $\mathcal{G}$ can be admitted as a plausible causal hypothesis for the observed joint distribution, but it does not guarantee the uniqueness of the DAG. Therefore, more assumptions are needed to limit the number of DAGs that are acceptable as possible causal hypotheses.

A joint distribution and a DAG are said to be *faithful* to one another if the collection of conditional independences in the joint distribution correspond exactly

to the set of d-separation properties in the DAG, i.e. for any three disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we have that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \Leftrightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$. Accordingly, Spirtes et al. [SGS93] formalize another common assumption called *causal faithfulness*.

Postulate 2.3 (Causal faithfulness condition)**.** *The true causal DAG for a joint distribution $P_{\mathbf{X}}$ is faithful to $P_{\mathbf{X}}$.*

The DAGs that are faithful to the observed joint distribution are said to be *Markov equivalent*, and they form a *Markov equivalence class*. The causal Markov condition, together with the causal faithfulness condition, facilitate the inference from the conditional independences obtained from observational data to the true causal relationships of the variables. Many widely used causal discovery algorithms based on the notion of conditional independence haven been developed, and they rely on the two aforementioned assumptions to identify the Markov equivalence class as plausible hypotheses of the true causal structure. We refer readers to Section 5.4 of [SGS+01] for details on popular approaches such as the *PC algorithm* and the *Fast Causal Inference (FCI) algorithm*.

2.2 Applications of Structural Causal Modelling Pertinent to Machine Learning

In this section, we demonstrate the role of causality in facilitating statistical studies that are pertinent to machine learning. We present a comprehensive review of how causal inference and the SCM framework help address the limitations of current machine learning approaches. It is believed that [Sch19][Pea18] the transition from model-blind to model-based algorithms is the key to achieve human-level performance in machines.

2.2.1 Simpson's Paradox

This section reviews a peculiar phenomenon called Simpson's Paradox and how it is later considered resolved. Basically, it has been discovered that the association between two variables X and Y that appears in a total population can be inverted when the data points are segregated into subpopulations based on a third variable Z. Similarly, an association between X and Y that is consistent in the segregated groups can be reversed in the combined data. This phenomenon is named after the statistician E. H. Simpson, who was the first to articulate this issue in his paper [Sim51] in 1951. This was before the major advancements in causality over the past few decades, and Simpson, along with many other statisticians [Bly72][BFH75][LN$^+$81] of the 20th century, concluded that it is not possible to determine the true association in the paradox based on statistical data alone, and they suggested the use of non-statistical information as well as the notion of "exchangeability" [*ibid.*, Section 3] to help make the correct judgment.

In an attempt to investigate sex discrimination in graduate admissions at the University of California, Berkeley, Bickel et al. [BHO75] illustrates a typical example of Simpson's paradox. Table 2.1 shows the admissions data we construct loosely based on their study. Suppose that there are two departments in a university, one is Mathematics and the other is History. Upon looking at the total counts for men and women, one may conclude that the university is indeed discriminating in favour of men. However, the acceptance rates for women are higher than men in each department, and the data table containing these reversals appears to be "paradoxical".

In the context of human psychology, this paradoxical surprise partly arises from people's intuitive belief that associations between the ratios in the divided groups should be close to the associations between the ratios of sums in the total population. To describe it explicitly in arithmetic form, suppose that $a_1, a_2, b_1, b_2, c_1, c_2, d_1$

Applicants	Gender	Admit	Deny	Acceptance Rate
Department of Mathematics	*Men*	100	120	0.455
	Women	20	20	**0.5**
Department of History	*Men*	20	80	0.2
	Women	100	300	**0.25**
Total	*Men*	120	200	**0.375**
	Women	120	320	0.273

Table 2.1: Synthetic admissions data by gender and field of study.

and d_2 are natural numbers. Upon observing the two associations $\frac{a_1}{b_1} < \frac{c_1}{d_1}$ and $\frac{a_2}{b_2} < \frac{c_2}{d_2}$, one would intuitively expect that $\frac{a_1+a_2}{b_1+b_2} < \frac{c_1+c_2}{d_1+d_2}$. This is in fact an invalid conclusion, because the associations between the ratios in the partitioned data do not guarantee the same regularities in the aggregated population. Specifically, the ratios are compared equivalently in each subgroup, but they contribute different weights when the data are pooled. Yet, more explanations are required for people's strong intuition that such sign reversals are impossible, and philosophers postulate that human's intrinsic causal reasoning should be considered apart from the statistical interpretations. The sure-thing principle in decision theory proposed by Savage [Sav72] is an example that demonstrates people's causal reasoning, and Pearl develops a corrected version of the principle [Pea00] based on the mathematical expression called the $do(\cdot)$ operator [Pea95]. In terms of definition, $P(Y \mid do(x))$ differs from the observational $P(Y \mid x)$ in that the former is the distribution of Y given that we artificially set the value of X to x, while the latter is for when we observe that X takes value x.

Theorem 2.2.1 (Causal Sure-Thing Principle [PM18])**.** *If both* $P(Y \mid do(X), Z) > P(Y \mid do(\neg X), Z)$ *and* $P(Y \mid do(X), \neg Z) > P(Y \mid do(\neg X), \neg Z)$ *hold, then it must be true that* $P(Y \mid do(X)) > P(Y \mid do(\neg(X))$*, given that the action on* X *does not affect the values of* Z*.*

In other words, if an action $do(X)$ does not affect the distribution of the partitions, our intuitive conclusion is that it is impossible to observe Simpson's reversal

in the data. If the action $do(X)$ does influence the distribution of Z, then one should admit the possibility of Simpson's reversal. More formally, using the notions in graphical causal models, Pearl [Pea00] presents the back-door criterion to determine in which scenarios we should accept that a Simpson's reversal is possible. Suppose that we construct a DAG $\mathcal{G}$ that represents our belief of the causal structure.

Definition 2.3 (Back-door criterion (*ibid*,. p 79))**.** *Given a DAG $\mathcal{G}$, a set of variables* $\mathbf{Z}$ *is said to satisfy the back-door criterion with respect to a pair of variables (X, Y) if (i):* $\mathbf{Z}$ *does not contain any variable that is a descendant of X; and (ii): for any path p that ends with an arrow into X, p is blocked by* $\mathbf{Z}$*.*

If there exists a set of variables $\mathbf{Z}$ that satisfies the back-door criterion with respect to (X, Y), Pearl [*ibid.*, Theorem 3.3.2] shows that X has an identifiable causal effect on Y.

For the admissions data example in Table 2.1, we can denote the gender of an applicant as X, the application status as Y, and the choice of department as Z. The key to accepting the reversal in the data is the fact that women tend to apply to departments that are generally difficult to get in, regardless of one's gender (i.e. the department of history with relatively low acceptance rates for both men and women). This creates an arrow from X to Z in the DAG. The causal diagram can be represented as follows.

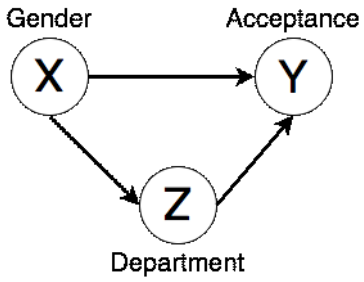

Figure 2.2: Example causal DAG of Table 2.1.

Using the back-door criterion, it is clear that neither of the two paths between

X and Y in the DAG is a back-door path that needs to be blocked. In other words, there is no spurious paths that requires Z to be conditioned on, and we should conclude that the true association lies in the combined data. Given a causal DAG, the back-door criterion indicates whether it is possible to observe a Simpson's reversal and identifies the true association from the aggregated and partitioned data, while traditional statistical methods are insufficient to articulate the reasoning behind such reversal.

2.2.2 Confounding and Deconfounding

One of the main challenges concerning observational studies is confounding, where a latent variable Z is the common cause of both treatment X and outcome Y. This hidden Z can lead to spurious associations between X and Y, and therefore it must be conditioned on when estimating the effect of X on Y. By definition, for a data-generating causal model M in which the distribution Y is dependent on X, the effect of X on Y is not confounded if and only if $P(y \mid x) = P(y \mid do(x))$, where $P(y \mid do(x))$ denotes the causal effect of X on Y, expressed by the probability of $Y = y$ when we manually manipulate X to have the value x. In other words, the two variables are not confounded if their observed association is equivalent to the association obtained from the intervention $do(x)$. To calculate $P(y \mid do(x))$, one can synthesize the intervention that sets the value of X to x. This is natural for experimental studies but difficult for observational studies. Under the SCM framework, however, the causal effect can be estimated using a graphical method derived by Pearl [Pea95] based on the back-door criterion (Section2.2.1). Essentially, the author shows [*ibid.*, Theorem 3.3] that for any set of variables $\mathbf{Z}$ in a DAG $\mathcal{G}$, if $\mathbf{Z}$ satisfies the back-door criterion with respect to (X, Y), then the identification of the causal effect of X on Y is possible, and the following adjustment formula can be applied to estimate the conditional

interventional distribution.

$$P(y \mid do(x)) = \sum_{z \in \mathbf{Z}} P(y \mid x, z) P(z) \tag{2.3}$$

On the other hand, if $\mathcal{G}$ does not contain any set of variables that satisfies the back-door criterion for (X, Y), Pearl [*ibid.*, Section 3.2] shows that the causal effect of X on Y can still be computed based on the following front-door criterion.

Definition 2.4 (Front-door criterion)**.** *Given a DAG $\mathcal{G}$, a set of variables $\mathbf{Z}$ is said to satisfy the front-door criterion with respect to a pair of variables (X, Y) if (i): for any directed path p from X to Y, there exists a variable in $\mathbf{Z}$ that lies in p; (ii): $\mathcal{G}$ does not contain any back-door path from X to $\mathbf{Z}$; (iii): for any back-door path p from $\mathbf{Z}$ to Y, p is blocked by X.*

If there exists a set of variables $\mathbf{Z}$ in the DAG $\mathcal{G}$ that satisfies the front-door criterion with respect to (X, Y), and $P(x, z) \neq 0$ for any $z \in \mathbf{Z}$, then the author shows [*ibid.*, Theorem 3.5] that there is an identifiable causal effect of X on Y, and the front-door adjustment formula can estimate $P(y \mid do(x))$:

$$P(y \mid do(x)) = \sum_{z \in \mathbf{Z}} P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{2.4}$$

Both the back-door and front-door conditions make the estimation of causal effects possible using purely observational data, yet they require relatively restrictive assumptions for the structure of a causal model. A more powerful tool for computing the causal effect $P(y \mid do(x))$ has been proposed by Pearl [Pea00] called do-calculus. It is a collection of inference rules that derives from the interventional conditionals $P(y \mid do(x))$ and produces purely associational probabilities. Consider a DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ in which $X, Y, Z, W \in \mathbf{V}$ are any four disjoint sets of observable variables. The author defines a subgraph $\mathcal{G}_{\underline{X}}$ [*ibid.*, Section 3.4.1] as the resulting graph by removing from $\mathcal{G}$ any directed edges from X_i to V_j, where

$X_i \in X$ and $V_j \in \mathbf{V} \setminus X$. Similarly, $\mathcal{G}_{\overline{X}}$ refers to the subgraph that removes all the incoming arrows toward the nodes in X. The inference rules of do-calculus proposed by Pearl [*ibid.*, Theorem 3.5] are as follows.

Rule 1 *Insertion/deletion of observations*

Consider the subgraph $\mathcal{G}_{\overline{X}}$*, if* Y *and* Z *are conditionally independent given* X *and* W*, then*

$P(y \mid do(x), z, w) = P(y \mid do(x), w)$.

Rule 2 *Action/observation exchange*

Consider the subgraph $\mathcal{G}_{\overline{X}\underline{Z}}$*, if* Y *and* Z *are conditionally independent given* X *and* W*, then*

$P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$.

Rule 3 *Insertion/deletion of actions*

Consider the subgraph $\mathcal{G}_{\overline{X},\overline{Z(W)}}$*, where* $Z(W)$ *denotes the set of variables in* Z *that are not ancestors of any* $W_i \in W$ *in* $\mathcal{G}_{\overline{X}}$*, if* Y *and* Z *are conditionally independent given* X *and* W*, then*

$P(y \mid do(x), do(z), w) = P(y \mid do(x), w)$.

For any causal effect that is identifiable, Shpitser et al. [SP12] and Huang et al. [HV12] have proved that the rules of do-calculus are sufficient to produce the equivalent probabilistic expressions without the interventional conditions. Nevertheless, it has not been determined whether any arbitrary conditional interventional distribution has an equivalent representation based on probabilities at the level of Association (Section 1.3). Therefore, the rules of do-calculus are not necessarily overshadowing the aforementioned graphical methods based on the back-door and front-door criteria.

2.2.3 Counterfactual Reasoning

At the top level of the causal hierarchy (Figure 1.1), counterfactual inference requires a cognitive system to imagine a parallel world in which an event A happened, while in reality it was $\neg A$ that happened. This type of retrospective thinking is natural for human, but it cannot be interpreted by machines from the Association level, nor can it be clearly expressed mathematically. Recent advances in causal inference have provided a potential solution to address this limitation in machines by the regimentation of counterfactual reasoning in the SCM paradigm.

Based on an SCM, a counterfactual query normally involves two endogenous variables X and Y, with the objective of computing $P(Y_x(u) = y \mid z)$, which denotes the probability of the hypothetical event that given that we have observed the evidence $Z = z$, Y would have been y (in situation u) if X were set to x by the action $do(x)$. The situation u refers to a particular realization of the exogenous variables U (i.e., the noise variables $\mathbf{N}$ in Section 2.1). In addition to the basic notations of SCM, Pearl [Pea00] introduces several terms under the same framework to help describe the effect of a local action.

First, note that the potential response $Y_x(u)$ [*ibid.*, Definition 7.1.4] considers the local action $do(X = x)$. To express the hypothetical effect of the action $do(x)$, the author defines [*ibid.*, Definition 7.1.2] a sub-model $M_x = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}_x \rangle$ based on the originally given SCM M, where $\mathbf{U}$ is the set of exogenous variables, $\mathbf{V}$ is the set of endogenous variables that contains X, and $\mathbf{F}_x = \{f_i : \mathbf{V}_i \notin X\} \cup \{X = x\}$. Second, the tuple $\langle M, P(u) \rangle$ is defined [*ibid.*, Definition 7.1.6] as the probabilistic causal model, where $P(u)$ is the distribution of the background variable U.

With these terminologies in mind, the author proposes [*ibid.*, Theorem 7.1.7] a three-step procedure to evaluate the counterfactual query $P(Y_x(u) = y \mid z)$ of a model $\langle M, P(u) \rangle$:

1. Compute $P(u \mid z)$ to replace the original distribution $P(u)$.

2. Substitute the model M with the sub-model M_x with the action $do(X = x)$ imposed.

3. Compute $P(y)$ using the new sub-model $\langle M_x, P(u \mid z)\rangle$.

However, a practical challenge arises from the first step, where the exogenous variables in $\mathbf{U}$ are no longer mutually independent when conditioned on the observed evidence z, and these dependencies need to be explicitly described in the joint distribution of $\mathbf{U}$ given z. Balke et al. [BP11] propose the so-called twin network algorithm that overcomes this challenge.

Consider an SCM with n endogenous variables $X_1, ..., X_n$, let $\mathbf{X}^* = \{X_1^*, ..., X_d^*\}$ denote the corresponding variables in the counterfactual world. The algorithm [*ibid.*, Secion 5] constructs a Bayesian network that combines both the real world and the counterfactual one. Both worlds contain the same set of exogenous variables $\mathbf{U}$, since any $U \in \mathbf{U}$ exists prior to the forced action $do(X = x)$ and is not affected by the hypothetical change. The algorithm [*ibid.*, Secion 5] can be summarized as follows:

1. Construct a Bayesian network $\langle \mathcal{G}, \mathcal{P}\rangle$, where $\mathcal{G}$ is a DAG with $3n$ nodes $\mathbf{V} = \{X_1, ..., X_n\} \cup \{X_1^*, ..., X_n^*\} \cup \{U_1, ..., U_n\}$. $\mathcal{G}$ contains all the original edges of the real world variables $\{X_1, ..., X_n\}$ as well as $\{U_1, ..., U_n\}$, and an arrow is created for every $X_i^* \to X_j^*$ such that X_i is a parent of X_j. Additionally, create an arrow for every $U_i \to X_i^*$. The set of conditional probability distributions $\mathcal{P}$ is defined based on the structural functions from the original SCM. In particular, $P(x_i \mid \mathbf{PA}_X(x_i), U_i) = 1$ if $x_i = f_i(\mathbf{PA}_X(x_i), U_i)$ and 0 otherwise. For the counterfactual part, let $P(x_i^* \mid \mathbf{PA}_{X^*}(x_i^*), U_i) = P(x_i \mid \mathbf{PA}_X(x_i), U_i)$ if $x_i = x_i^*$ and $\mathbf{PA}_X(x_i) = \mathbf{PA}_{X^*}(x_i^*)$. The distributions of the background variables $\{U_1, ..., U_n\}$ are the same as in the original SCM.

2. Assign values to the respective real-world variables for the observed evidence x_{obs}.

3. Now, to impose a local action $do(X_k^* = \hat{x_k}^*)$, simply modify the DAG structure of the counterfactual part by deleting all the incoming arrows to X_k^* and assigning X_k^* with the value $\hat{x_k}^*$.

4. The final output of the counterfactual query $P(X_{i,\hat{x_k}^*}(u) = x_i \mid x_{obs})$ is produced by performing the standard belief propagation algorithm [Pea82] on the twin network.

Balke et al. [BP11] show that although the distributions of the noise variables remain undetermined, the twin network algorithm can produce a unique output if a proper distribution can be assumed for the uncertain noise terms; if not, there exists a bounded solution based on convex optimization [BP93].

Chapter 3

Selection Bias and Causation

3.1 Introduction

This chapter examines the information-geometric causal inference (IGCI) method proposed by Janzing et al. [JMZ+12] in the presence of selection bias. The IGCI algorithm is designed to infer causal directions from bivariate observational data. That is, suppose we have the observational distribution $P_{X,Y}$, where X and Y are presumably correlated, we would like to infer whether X causes Y or Y causes X. Suppose the ground truth is that X causes Y, IGCI identifies asymmetric patterns between cause and effect by examining independence. In particular, it is assumed [*ibid.*, Section 1] in IGCI that the marginal distribution of cause and the conditional distribution of effect given cause are independent of each other. In this case, the marginal distribution P_X should not give any information about the relationship between Y and X and vice versa. However, when certain types of selection bias exist in the observational data, the aforementioned independence assumption may no longer hold, and the interpretation of causation using IGCI can thus be misleading.

In the next section we present a type of selection bias that is observed in medical studies. In particular, we focus on a nationwide epidemiological study of

dementia conducted as part of the Canadian Study of Health and Aging (CSHA) and discuss the bias that arose in the prevalent cohort survival data collected in the study. Section 3.3 gives a review of the information-geometric approach to inferring causation. Based on the case study of CSHA, we highlight that IGCI can be misleading in situations where the observational data are not representative of the population of interest.

3.2 Sampling Bias and Induced Covariate Bias: A Case Study of Prevalent Cohort Survival Data

In an effort to investigate the prevalence of dementia in the elderly, the CSHA was initiated in 1991. The researchers of the CSHA had two main objectives: first, they would like to estimate the survivor function of people with dementia measured from the onset of the disease; second, they aimed to identify the variables that are correlated with the survival time. The first phase of the study (CSHA-1) began in 1991, during which 821 subjects with existing dementia were recruited across Canada into the study and underwent a detailed clinical examination. These subjects each was diagnosed with exactly one of the three types of dementia, namely, probable Alzheimer's disease, possible Alzheimer's disease and vascular dementia. After the initial cross-sectional study, a follow-up had been conducted for 5 years. In 1996, the second phase of the CSHA (CSHA-2) started, and the remaining 21.9% of the original 821 subjects who were still alive were re-evaluated in a similar way as the subjects of CSHA-1. For each of the 821 subjects, a survival time can be calculated by the difference between the date of onset and the date of death or right censoring. Additional data collected from the subjects include age at onset of dementia, sex, education level and classification of dementia.

However, an issue of selection bias has been noticed by Wolfson et al. when

examining the data collected in the CSHA. In their seminal article, Wolfson et al. [WWA+01] point out that the median survival time following the onset of dementia is overestimated if not adjusted for the bias of the sample. Essentially, the observational data of survival times are left-truncated in the CSHA, meaning the dates at onset of the recruited subjects are all prior to the cross-sectional study in 1991. The prevalent cohort survival data of the CSHA are thus not from a representative sample of the elderly Canadian population, among which the dementia patients with longer survival times have a higher chance of being recruited into the study. If we invoke the stationary assumption that the incidence rate of dementia has not changed over time, the left-truncated survival data are defined to be "length-biased" [Wan91].

Apart from the issue of length-bias on the response variable, Bergeron et al. [BAW08] highlight another issue concerning the covariate bias that is induced by the length-biased survival data in the CSHA. The authors point out that the sampling distributions of predictor variables such as age at onset and sex are influenced by the length-biased survival times. Suppose that Y is a random variable that denotes the true survival time, with mean $\mu(\boldsymbol{\theta})$ and probability density function $f_Y(y)$. Let X be a covariate, then Bergeron et al. [*ibid.*, Section 3.1] show that for a subject from the observed sample,

$$f_B(x;\boldsymbol{\theta}) = \frac{\mu(x;\boldsymbol{\theta}) f_X(x)}{\mu(\boldsymbol{\theta})} \tag{3.1}$$

is the biased density of X, where $\mu(\boldsymbol{\theta}) = E(E(Y \mid X)) = E(Y)$, $\mu(x;\boldsymbol{\theta}) = E(Y \mid x)$, and $f_X(x)$ is the true density of X from the population of interest. Therefore, the sampling distribution of the covariate X carries information about the relationship between X and the response variable Y. In particular, since the dementia patients with longer survival times have a higher chance of being recruited into the prevalent cohort study, the observed values of the covariates are

biased towards these long-term survivors as well.

3.3 Information-Geometric Causal Inference (IGCI) Approach and Its Practical Issues

It is clear that the conventional causal inference methods developed by Spirtes et al. [SGS93] and Pearl [Pea00] are based on the notion of conditional independence and thus work with at least three observables. These methods also require certain assumptions of temporal precedence and Gaussianity to construct the graphical causal model. In this section, we focus on causal discovery involving only two variables and discuss a fundamental problem, which is inferring causal directions between two purely observational variables.

Let X and Y be two i.i.d. random variables with a joint distribution $P_{X,Y}$, and the observational data contain n pairs of values $(x_1, y_1), ..., (x_n, y_n)$. The problem boils down to inferring whether X causes Y ($X \longrightarrow Y$) or vice versa. It should be noted that this problem is challenging yet rather simplified, as it assumes [MPJ+16] that $X \not\perp\!\!\!\perp Y$, and that there is no hidden confounding variable, no selection bias, and no bidirectional causation between X and Y.

Supposing that the prevalent cohort survival data of the CSHA are used for the study of causation, one possible objective would be to infer whether a covariate X causes the response Y, for example, whether age at onset of dementia causes the survival time. The functional causal models for bivariate data can be applied naturally to solve this problem. The Information-Geometric Causal Inference (IGCI) developed by Janzing et al. [JMZ+12] aims to address a rather restrictive case in which $Y = f(X)$ with f being bijective. Here we review the IGCI method and discuss its practical issues when applied to biased data.

The main idea of IGCI is based on the following postulate proposed by Schölkopf

et al. [SJP+12].

Postulate 3.1 (Independent causal mechanisms principle)**.** *Within the same system of variables, the conditional distribution of each variable given its causes is generated by an autonomous module that is independent of every other module of the system.*

In the bivariate case, this means that if X causes Y, then the marginal distribution P_X and the conditional distribution $P_{Y|X}$ contain no information about each other. In [JMZ+12] this notion of independence is formalized in terms of algorithmic information (i.e. Kolmogorov complexity). Recall in Section 2.1 the set of d conditional distributions $P_{X_j|\mathbf{PA_j}}$ form the free parameters of the corresponding DAG $\mathcal{G}$. In [LJ13] the authors postulate that if $\mathcal{G}$ and these conditional densities represent the true causal structure, then the shortest description of the observed joint distribution $P_{X_1,\ldots,X_d}$ can be expressed by distinct descriptions of the d conditional distributions $P_{X_j|\mathbf{PA_j}}$. In other words, the observed joint distribution $P_{X1,\ldots,X_d}$ satisfies the so-called *Algorithmic Independence of Conditionals* [JS10]:

$$K(P_{X_1,\ldots,X_d}) \stackrel{+}{=} \sum_{j=1}^{d} K(P_{X_j|\mathbf{PA_j}}), \tag{3.2}$$

where $K(\cdot)$ is the Kolmogorov complexity, and $K(P_{X_j|\mathbf{PA_j}})$ denotes the shortest length of the program that computes $P_{X_j|\mathbf{PA_j}}$ using the values x_j and $\mathbf{pa_j}$. For the bivariate case, given that X causes Y, then 3.2 can be interpreted as that $P_{Y|X}$ and P_X are independent causal mechanisms, and that they generate values of y and x separately. Subsequently, the asymmetric pattern between X and Y becomes obvious from the dependence between P_Y and $P_{X|Y}$.

In the deterministic case, Mooij et al. [MPJ+16] point out that $P_{Y|X}$ contains the same information as f, since $P(Y = y \mid X = x)$ can be expressed as the indicator function $\mathbb{1}_{y=f(x)}$. Therefore, the major assumption of IGCI is equivalent to saying that P_X has no information about f and vice versa given that X causes

Y.

For any specific inference problem, IGCI requires a reference probability distribution to be specified to restrict the range of X and Y, so that IGCI can be applied to a well-defined problem. The authors [*ibid.*, Section 3.1] suggest two choices, one is the uniform distribution on an interval $[a, b]$ if X and Y are theoretically within this interval; the other is the Gaussian distribution if the ranges are unbounded. Let u_X and u_Y be such reference densities for X and Y respectively, and let $h(x)$ be any function that expresses certain properties of the conditional $P_{Y|X}$ at $X = x$. Given that X causes Y, then the functions h and p_X/u_X are uncorrelated, and the authors show [JMZ+12] that

$$\int h(x)p(x)dx \approx \int h(x)u_X(x)dx. \tag{3.3}$$

Definition 3.1 (Kullback-Leibler divergence [KL51])**.** *If p and q are two densities and p is absolutely continuous with respect to q, the Kullback-Leibler divergence (or relative entropy) from q to p is defined by*

$$D(p \parallel q) := \int log\frac{p(x)}{q(x)}p(x)dx \geq 0. \tag{3.4}$$

Let u_f be the image of u_X under f, and let $u_{f^{-1}}$ denote the image of u_Y under f^{-1}. Regarding the choice of h in 3.3, the authors [MPJ+16] consider the function $log(u_{f^{-1}}/u_X)$, which is only related to f and the reference densities but not p_X, and it is therefore postulated [*ibid.*, Principle 2] that

$$\int log\frac{u_{f^{-1}}(x)}{u_X(x)}p(x)dx \approx \int log\frac{u_{f^{-1}}(x)}{u_X(x)}u_X(x)dx. \tag{3.5}$$

Janzing et al. [JMZ+12] then formalize the IGCI method by defining

$$C_{X\longrightarrow Y} := \int log\frac{u_{f^{-1}}(x)}{u_X(x)}p(x)dx. \tag{3.6}$$

When X causes Y, by Definition 3.1 and 3.5 we have that

$$\begin{aligned} C_{X \longrightarrow Y} &\approx \int log \frac{u_{f^{-1}}(x)}{u_X(x)} u_X(x) dx \\ &= -\int log \frac{u_X(x)}{u_{f^{-1}}(x)} u_X(x) dx \\ &=: -D(u_X \parallel u_{f^{-1}}) \leq 0. \end{aligned}$$

Therefore, the IGCI approach by Janzing et al. [JMZ$^+$12] infers that X causes Y whenever $C_{X \longrightarrow Y} < 0$ and that Y causes X if $C_{X \longrightarrow Y} > 0$.

Consider the prevalent cohort survival data from the CSHA, in an ideal case, we can examine if a covariate is a cause of the survival time by estimating $C_{X \longrightarrow Y}$ from IGCI. However, this approach requires that the marginal distribution of cause gives no information about the functional relationship f between cause and effect, and in Section 3.2 we have discussed that the covariates are correlated with the mean survival time conditioned on the values of covariates. Therefore, IGCI can be misleading if it is applied to a dataset that is not representative of the population.

Chapter 4

Machine Intelligence and Bias

In spite of numerous studies led by researchers in AI [PVH01][LH07], cognitive science [BT96], neuroscience [CD14] and philosophy [Val20], it has not yet been established whether machines will be able to achieve human-level intelligence in the foreseeable future. One of the most accepted theories on this widely-discussed question is the Turing Test proposed by Alan Turing [Tur09] in 1950. It states that a computer should be deemed intelligent if the human interrogator cannot distinguish it from the other human respondent in online chat. After a few decades following Turing's claim, there has been remarkable progress in the field of natural language processing, and the validity of the Turing Test has been challenged by a very powerful counter-example called the Chinese Room argument. In his argument, the UC Berkeley philosopher John Searle [J$^+$80] raises doubts about whether machines that pass the Turing Test can truly understand language and think like human. He describes a hypothetical room in which there is a person who knows no Chinese. This person receives questions in Chinese from the outside, and he needs to output answers to the questions using a complete set of Chinese characters as well as a manual for handling the characters. Based on the standard of the Turing Test, a human interrogator will not be able to tell the difference between this Chinese room and another human respondent who actually knows Chinese,

while the person in the room is not performing any reasoning or thinking tasks, as the manual of characters is all he needs to pass the test. To this day, Searle's argument has not yet been fully justified [Har01][Hau97][PB02], but it highlights a major problem facing the machine learning community nowadays: even the most advanced machines exhibiting human-like intelligence are merely finding patterns based on certain instructions. These instructions can either be explicitly provided by human or learnt from data, but just like the person in the Chinese Room who knows no Chinese, there is no evidence that machines understand the underlying causal mechanisms that generate the observed data.

Fortunately, more and more researchers have recognized that causality is set to become a vital factor in the future of machine learning, especially of deep learning. What is particularly worth noting is the increasing interest of adversarial vulnerability of deep learning models. By tweaking the feature space of deep neural networks (DNNs), researchers are able to fool state-of-the-art machines into producing erroneous predictions with high confidence. This chapter begins by a brief review of existing adversarial examples, which provide a strong implication that machines can be fooled just like humans do. In Section 4.2 we provide a selective review of causal approaches to improving the robustness of DNNs. In the third section we draw connections between machine bias and human cognitive bias by comparing the types of intelligence they conform to.

4.1 Adversarial Examples

With the increasing data availability that is unprecedented in history, the past decade has seen a renewed importance in neural networks, which take advantage of large amounts of data more than traditional ML algorithms do. In particular, DNNs are neural networks with more than one hidden layer, and they are known to have achieved state-of-the-art performance in human-like perception tasks, such

as image recognition [KSH12], speech recognition [HDY+12] and natural language processing [BCB14]. However, concerns have arisen that call into question the robustness of many DNN models towards adversarial attacks. Szegedy et al. in 2013 [SZS+13] are among the first to investigate the adversarial vulnerability of DNNs. In their study focusing on image classification tasks, an adversary constructed slightly modified examples while maximizing the error function, and the difference between the modified and original examples cannot be detected by human observers. These mildly perturbed examples were able to fool various ML models into making false predictions, even when the models were trained on disjoint subsets of the training data. Hence, these adversarial attacks are not random results of model overfitting. Instead, they have been proved to be generalizable, revealing the blind spots of many ML models.

For the following algorithms, we denote X as an input to DNN and y as the true label of X. Let $J(\boldsymbol{\theta}, X, y)$ be the cost function that is used to train the model, where $\boldsymbol{\theta}$ denotes the parameters of the network. Let ϵ be a hyperparameter that denotes the size of the adversarial perturbation, which is usually a small value that acts as a constraint on the max-norm of the perturbation η (i.e. $\|\eta\|_\infty < \epsilon$).

In their seminal article on adversarial examples, Goodfellow et al. [GSS14] develop a straightforward linear perturbation method for generating adversarial examples. They postulate that [*ibid.*, Section 3] many DNNs show linear behaviours and that an optimized linear perturbation to the input is sufficient to impair network performance. The *fast gradient sign method* [*ibid.*, Section 4] they propose is summarized in Algorithm 1.

Algorithm 1: Fast Gradient Sign Method

Input: X, y, hyperparameter ϵ, cross-entropy cost function $J(\cdot)$, network parameters $\boldsymbol{\theta}$

Output: adversarial example X^*

1 compute $\nabla_X J(\boldsymbol{\theta}, X, y)$ using back-propagation

2 Optimize the perturbation η based on the L_∞ norm constraint and obtain $\eta = \epsilon \cdot sign(\nabla_X J(\boldsymbol{\theta}, X, y))$

3 Generate adversarial example of X as $X^* = X + \eta$

The method in Algorithm 1 is a direct and efficient method to create an adversarial perturbation that is applied to the entire feature space of X. The authors have shown in their experiments [*ibid.*, Section 4] that Algorithm 1 can cause various types of neural networks trained on different image datasets to produce misclassifications with high confidence rates, and their results support the hypothesis on the linearity of DNNs.

As an extension to the fast gradient sign method, Kurakin et al. [KGB16] propose the *iterative least-likely class method* (summarized in Algorithm 2), which achieves a much higher success rate than Algorithm 1. Unlike the non-targeted approach in Algorithm 1, the method by Kurakin et al. [*ibid.*, Section 2.3] is a targeted attack, in other words, it is designed to create examples that lead to a specific misclassification y^*. The authors [*ibid.*, Section 2.3] choose the target label y^* to be the least likely output when feeding X into the trained model, so that the misclassified label is as distinct as possible from the true label y.

Algorithm 2: Iterative Least-Likely Class Method [KGB16]

Input: X, y, hyperparameters ϵ and α, cross-entropy cost function $J(\cdot)$, network parameters $\boldsymbol{\theta}$

Output: adversarial example X^*

Obtain the least likely class from the network prediction of X as the target class $y^* = \arg\min_{y'} p(y' \mid X)$

Initialize α and the number of iterations m heuristically

Define $C_{X,\epsilon}(X')$ as the element-wise pixel clipping function [*ibid.*, Section 2] of the image X' with respect to the L_∞ ϵ-neighbourhood of the original image X

Initialize $X_0^* \leftarrow X$

for $i \in \{1, ..., m\}$ **do**

  $X_i^* \leftarrow C_{X,\epsilon}\{X_{i-1}^* - \alpha \cdot \text{sign}(\nabla_X J(\boldsymbol{\theta}, X_{i-1}^*, y))\}$

Generate adversarial example of X as $X^* = X_m^*$

It is believed [*ibid.*, Section 2.3] that Algorithm 2 leads to more interesting misclassifications than Algorithm 1, since for large datasets with more classes, Algorithm 1 might output a misclassified class that is still similar to the original one.

Another work on targeted perturbation is the Jacobian-based method developed by Papernot et al. [PMJ+16], and their approach is summarized in Algorithm 3. The authors [*ibid.*, Section 3] make the same assumption as in Algorithm 1 and 2 that information on the network architecture and parameters are known by the adversary, and they also assume that the attacked DNN is acyclic. Unlike the perturbations in Algorithm 1 and 2 that are applied to all the features of X, Algorithm 3 modifies a much smaller subset (4.02% on average) of the input features while still achieving a high success rate (97% on average) [*ibid.*, Section 8]. It also relies on a generated adversarial saliency map [*ibid.*, Section 3.2.2] to enable

the adversary to identify the set of input features that are most significant in producing a specific misclassified output.

Algorithm 3: Jacobian-based Saliency Map Method[PMJ+16]

Input: X, y^*, trained network expressed by function f, feature variation parameter η, maximum distortion parameter Υ

Output: adversarial example X^*

Initialize the adversarial example as $X^* = X$

Initialize $\Gamma = \{1, ..., |X|\}$

while $f(X^*) \neq y^*$ *and* $\|\delta_X\| < \Upsilon$ **do**

- Compute the forward derivative of X^* as $\nabla f(X^*)$ [*ibid.*, Section 3.2.1]
- Obtain saliency map $S = \text{saliency\_map}(\nabla f(X^*), \Gamma, y^*)$ [*ibid.*, Section 3.2.2]
- Identify the most significant feature $i_{max} = \arg\max_i S(X, y^*)[i]$, $i \in \Gamma$
- Add a pre-defined perturbation η to the feature $X^*_{i_{max}}$ that has the highest saliency score
- Update $\delta_X = X^* - X$, which controls the magnitude of perturbation

Return adversarial example X^*

In addition to the results of the DNN adversary, the authors [*ibid.*, Section 5.3] also formally evaluate human performance on the generated adversarial examples. Interestingly, the authors suggest that by tweaking the percentage of features perturbed in the adversarial examples, Algorithm 3 can fool human observers as well, although the success rate is not as high as that for DNNs.

4.2 Recent Causal Approaches to Improving DNN Robustness

As Schölkopf highlights in [Sch19], most ML algorithms are based on a strong assumption that the data are i.i.d.. In particular, models for image classification tasks are usually trained from benchmark datasets (e.g. MNIST [Den12], ImageNet [DDS+09] and CIFAR-10 [KH+09]), which provide a typical i.i.d. setting. On the other hand, adversarial attacks apply specifically constructed small perturbations to the original examples, and the machines are fooled as the i.i.d. data assumption is violated. This problem of robustness is not restricted to image recognition. Targeted adversarial examples have also been generated for NLP [JL17] and speech recognition [CW18] DNNs. Accordingly, a number of defence approaches (e.g. distillation in [PMW+16], input preprocessing in [GRCVDM17] and adversarial training in [MMS+17]) have been proposed, but more attacks have also been carried out [CW17][AC18] soon after, and the problem of generalization to non-i.i.d. settings remains unresolved to this day. In this section we discuss the potential of causality for improving the robustness of DNNs.

Causal discovery in image datasets

First, causal structures exist in image datasets and can be identified by machines. In [LPNC+17] Lopez-Paz et al. develop the *Neural Causation Coefficient (NCC)* model that identifies the direction of causation from an observed joint distribution of two variables. NCC can be viewed as a generalization of the bivariate observational causal discovery algorithms, such as the additive noise model (Section 1.2) and IGCI (Section 3.3), since NCC is a trained neural network that is not restricted to a specific class of causal mechanisms (i.e. the assumption on function f that generates the effect).

The training data of NCC consist of n samples of synthesized joint observations

of cause X and effect Y. For each sample $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$, the authors [*ibid.*, Section 3.1] sample the values x_{ij} of the cause variable X from a Gaussian mixture distribution, then the values y_{ij} are generated from x_{ij} using a heteroscedastic additive noise model. For each pair of x_{ij} and y_{ij}, the authors include both the causal (x_{ij}, y_{ij}) and anticausal (y_{ij}, x_{ij}) data points in the training set, and a binary label that indicates the causal direction is added to each data pair. In other words, NCC is trained on a dataset D of $2n$ synthesized samples [*ibid.*, Equation 4]:

$$D = \{(\{(x_{ij}, y_{ij})\}_{j=1}^{m_i}, 0)\}_{i=1}^{n} \cup (\{(y_{ij}, x_{ij})\}_{j=1}^{m_i}, 1)\}_{i=1}^{n}. \tag{4.1}$$

Clearly, the output of NCC for a joint distribution P_{X_i,Y_i} is the predicted probability that Y_i causes X_i.

After tuning the hyperparameters on a synthesized validation set, Lopez et al. [*ibid.*, Section 4] apply the NCC classifier to real-world image datasets. Instead of learning a correlation between input image pixels and class labels like most DNNs are doing, Lopez et al.'s study aims to *reason* about the causal relationships in terms of the presence of objects in images. In particular, the authors focus on a preliminary interventional query that whether the presence of an object in an image will be affected when another object is removed from the image. To this end, they [*ibid.*, Section 4.1] choose a subset from the MSCOCO image dataset [LMB+14], in which each image contains object(s) from at least one of the 20 categories defined in [EVGW+10]. The authors [LPNC+17] use two networks to generate the inputs for NCC: one is a pre-trained ResNet N_1 to perform feature extraction on images; the other is an image classifier network N_2 they developed to identify the 20 categories. For every image x_j, the authors [*ibid.*, Section 4.2] use N_1 to generate $f_j = f(x_j) \in \mathbb{R}^l$, which denotes the high-level features from the last hidden layer of N_1. They also applied N_2 to each image input x_j and obtained an activation vector of the output layer of N_2 as $c_j = c(x_j) \in \mathbb{R}^{20}$. The

NCC model then takes the joint distribution $\{(f_{jl}, c_{jk})\}_{j=1}^{m}$ of each feature score l and category k and predicts a probability. Based on these probabilities, the authors [*ibid.*, Section 4.2] were able to identify the set of features that are most likely caused by the presence of the object corresponding to the labeled category, and these features are defined to be *anticausal*. The authors also identified a set of features that most closely associate with the object labeled by N_2, and they define these features as *object features*. Their experiments [*ibid.*, Section 4.3] show that anticausal features and object features are statistically dependent. This result implies that causal information can be inferred from image datasets, and it is one step towards machine reasoning for real-world scenarios.

The NCC model by Lopez et al. outperforms most bivariate causal discovery algorithms, although Lake et al. [LUTG17] point out that many real-world datasets contain hidden causal variables, and the current NCC algorithm is clearly not sufficient to identify latent structures.

Interpretability and explainability of DNNs

Along with the vast range of studies on adversarial vulnerability of DNNs, there is a growing concern about the reliability and interpretability of the decisions made by these black-box models. Since 2015, an open letter [RDT15] has been signed by numerous experts in AI, calling for researchers' attention to more robust and secure AI systems. A European Union regulation [GF17] has taken effect in 2018 and requires user-dependent algorithms to explain their decisions. Despite the doubting attitudes taken by some DL experts [LeC17] towards the complete explainability of DNNs, many promising attempts have been made to interpret or explain DNN predictions based on the SCM (Section 2.1) paradigm.

In [NSVM18] Narendra et al. propose using an SCM to represent the convolution layers of a convolutional neural network (CNN). Given the network architecture of a particular CNN, the authors [*ibid.*, Section 4.1] first construct a DAG

(Section 2.1) based on the structure of the convolution layers. For each filter in each convolution layer, a corresponding node is created in the DAG, and a directed edge is drawn from each filter in the i^{th} convolution layer to every filter in the $(i+1)^{th}$ convolution layer.

Second, the authors [*ibid.*, Section 4.2] define a transformation function $\phi : \mathbb{R}^{p\times q} \to \mathbb{R}$, so that given a dataset D, the matrix output from each filter can be mapped to a real number. Narendra et al. [*ibid.*, Section 5.2] suggest two reasonable choices of ϕ to capture enough information of the filters: one is the binary transformation

$$\phi(M_j^{[i]}) = \begin{cases} 1 & \|M_j^{[i]}\| < \mu_j^{[i]} + \sigma_j^{[i]} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where $M_j^{[i]}$ is the matrix output from the j^{th} filter of the i^{th} layer, $\mu_j^{[i]}$ and $\sigma_j^{[i]}$ are the mean and variance of $M_j^{[i]}$, respectively; the other choice they propose is the Frobenius norm transformation

$$\phi(M_j^{[i]}) = \|M_j^{[i]}\|_F := \sum_p \sum_q m_{pq}^2, \tag{4.3}$$

which clearly captures more information than a binary output of (4.2).

Lastly, the authors [*ibid.*, Section 5.2] estimate the set of structural equations **F** (Section 2.1) of the SCM. For the i^{th} convolution layer, suppose that there are k filters, and the output matrices from the k filters are $M^{[i]} = \{M_1^{[i]}, ..., M_k^{[i]}\}$. Let $\vec{R}^{[i]}$ denote the vector of real numbers computed by applying the transformation ϕ to $M^{[i]}$, i.e.,

$$\vec{R}^{[i]} = \begin{bmatrix} r_1^{[i]} \\ \vdots \\ r_k^{[i]} \end{bmatrix} = \begin{bmatrix} \phi(M_1^{[i]}) \\ \vdots \\ \phi(M_k^{[i]}) \end{bmatrix}. \tag{4.4}$$

Let $F^{[i]} = \{f_1^{[i]}, ..., f_k^{[i]}\}$ denote the structural functions for nodes (the k filters) in the i^{th} convolution layer, then the authors suggest [*ibid.*, Section 5.3] that each function $f_j^{[i]}$ can be estimated by fitting a regression model to $r_j^{[i]} = f_j^{[i]}(\vec{R}^{[i-1]})$, because according to the DAG, the filters in the $(i-1)^{th}$ layer are the parent nodes of the filter for $r_j^{[i]}$. Therefore, an SCM can be generated by combining the DAG with the set of learnt structural equations. This approach by Narendra et al. makes interventional and counterfactual queries for the partial structures of DNN models possible. However, to provide a complete explanation of the decision made by a CNN, we believe that additional network components need to be considered apart from the convolutional filters, and the choice of transformation function ϕ can be improved to further reduce the loss of information.

Other recent DNN interpretability approaches include an attribution-based method [CMSB19] that constructs an SCM to estimate the causal effect of an input neuron on the output neuron. Additionally, Harradon et al. [HDR18] extract high-level human-interpretable salient concepts from input image data, and they build an SCM that computes the causal effect of these salient concepts on the model predictions. We refer readers to [CMSB19] and [HDR18] for further details.

Causal learning and anticausal learning

Lastly, it has been suggested [SJPZ11] that the direction of causation in the learning task is associated with the robustness of algorithms. In [SJPZ11] Schölkopf et al. present a preliminary investigation into improving machine robustness using information on causal direction. Similar to their other studies on bivariate causal discovery (Section 3.3), the authors' work in [SJPZ11] focus on the bivariate case, in which X is the input variable, and Y is the output, and it is assumed in their work that there is no hidden confounder. Based on the postulate of independent causal mechanisms (Postulate 3.1), the authors show [*ibid.*, Section 2.1.1] that robustness to adversarial attacks should be guaranteed if the machine

is solving a causal learning problem, that is, the task of predicting the effect from cause. Even though adversarial examples are perturbed test examples that are drawn outside of the original input distribution P_X, the new conditional distribution $P'_{Y|X}$ is still generated from an independent causal mechanism, and therefore the authors [*ibid.*, Section 2.1.1] conclude that the output should be unaffected by the change of P_X, i.e. $P'_{Y|X} = P_{Y|X}$.

However, many state-of-the-art DNNs are solving anticausal problems. In particular, most image classification tasks are anticausal, since obviously the input images are observations caused by the output labels describing their respective entities and not vice versa. As Kilbertus et al. point out in [KPS18], the pervasiveness of adversarial vulnerability is not surprising when the machines are learning in the anticausal direction, because when Postulate 3.1 is not applicable, the expectation of strong generalization [*ibid.*, Section 3] becomes problematic.

4.3 Connections to Human-Level Intelligence

In Kahneman's book [Kah11] on cognitive psychology, the author proposes a dual-system for explaining human's decision making process. The "fast System 1" is intuitive, associative and heuristics-based; the "slow System 2" is purposive, introspective and reasoning-oriented. According to Kahneman's theory, the judgmental biases discussed in Section 1.1 can be viewed as byproducts of the collaboration between these two systems. Kilbertus et al. [KPS18] relate this dual-system theory to the anticausal learning of machines. Motivated by the fact that humans construct causal generative models to solve anticausal problems [JG06], the authors [KPS18] suggest the implementation of the dual-system to improve robustness of machines. In particular, they believe [*ibid.*, Section 4] that an anticausal model that resembles "System 1" should be paired with a causal model as the "slow System 2", so that the issue of strong generalization (Section 4.2) in anticausal

learning is rectified by the generalizable causal model.

Apart from the psychological perspectives, recent findings regarding machine adversarial examples make connections to human intelligence by examining human performance on adversarial examples. In Section 4.1 we have discussed that adversarial attacks can be generalized to different classes of models. Elsayed et al. [ESC+18] show that adversarial examples that are generalizable to different models can also fool human subjects, given that the examples are displayed to the subjects for only a very short period of time. On the other hand, these examples failed to fool subjects with no time limit. Their experiments [*ibid.*, Section 4.2.2] indicate a certain degree of similarity between DNNs and human visual perception process, but whether this perception bias is related to human cognitive bias (Section 1.1) has yet to be determined. In our opinion, time-limited visual perception might be potentially associated with the "fast System 1" of Kahneman's theory [Kah11], whereas subjects with no time limit might make more robust judgments based on reasoning by the "slow System 2". Harding et al. [HRBG18] present another thorough evaluation of human performance on adversarial examples. By comparing the fast gradient sign approach (Algorithm 1) with the Jacobian based approach (Algorithm 3), the authors suggest [*ibid.*, Section 6] that adversarial examples are not always perceivable by humans, and that non-targeted attacks (perturbations that are designed with no particular class expected) fool human observers more than targeted attacks do. Again, the study by Harding et al. [HRBG18] shows that humans can also be fooled by adversarial examples, but it is recommended that more in-depth comparative analyses should be undertaken to investigate the connections between human cognition and DNN algorithms.

Chapter 5

Conclusion

In this thesis, we reviewed some causal inference methods and examined the role of causality in explaining systematic errors in both human cognition and machine learning. In Chapter 1 we discussed human cognitive bias (Section 1.1) and the existing explanations based on causation (Section 1.1.5), which motivates the following investigation into causal inference methodologies. Section 1.2 provides a comparative analysis on the difference of current research directions between statisticians and computer scientists, and we affirmed in Section 1.3 that causality is a revolutionary power for the future of machine learning and human-level AI. In Chapter 2, we presented a comprehensive review of traditional causal inference on multivariate data in Pearl's structural causal model framework (Section 2.1), and we discussed the applications of SCM (Section 2.2) that are pertinent to machine learning. To present an example of machine bias resulting from unrepresentative data, in Chapter 3 we reviewed a bivariate causal discovery approach (Section 3.3) and demonstrated its limitation in the presence of selection bias. In Chapter 4 we reviewed targeted and non-targeted adversarial examples that fool deep neural networks (Section 4.1) as well as recent causal approaches proposed to improve machine robustness (Section 4.2). We presented a comparison between machine intelligence and human cognition in Section 4.3 based on Kahneman's conjecture on

human cognition and existing studies on machine adversarial vulnerability. Even though these chapters do not seem to be closely related to one another, each of these aforementioned components has its distinct contribution to further studies on machine learning bias.

Overall, this thesis highlights the importance of causality for facilitating future developments in machine learning. Although there exist preliminary investigations on the connection between human cognitive bias and machine learning bias, our knowledge on both human brains and black-box neural networks is far from adequate, and we hope that further research on causality will confirm the association between machine vulnerability and human judgmental bias.

References

[AC18] Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

[AF04] Peter Ayton and Ilan Fischer. The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369–1378, 2004.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[BAW08] Pierre-Jerome Bergeron, Masoud Asgharian, and David B Wolfson. Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*, 103(482):737–742, 2008.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Bel99] Edward Beltrami. *What is random?: chance and order in mathematics and life*. Springer Science & Business Media, 1999.

[BFH75] YMM Bishop, SE Fienberg, and PW Holland. Discrete multivariate analysis, 1975.

[BHO75] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

[Bly72] Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[BP93] Alexander Balke and Judea Pearl. Nonparametric bounds on causal effects from partial compliance data. In *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*. Citeseer, 1993.

[BP11] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. 2011.

[BT96] Mark H Bickhard and Loren Terveen. *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*, volume 109. Elsevier, 1996.

[Bur01] Bruce D Burns. The hot hand in basketball: Fallacy or adaptive thinking? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23, 2001.

[CC67] Loren J Chapman and Jean P Chapman. Genesis of popular but erroneous psychodiagnostic observations. *journal of Abnormal Psychology*, 72(3):193, 1967.

[CD14] David Daniel Cox and Thomas Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):R921–R929, 2014.

[CMSB19] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. *arXiv preprint arXiv:1902.02302*, 2019.

[Coc72] William G Cochran. Observational studies. *Introduction to Observational Studies and the Reprint of Cochranâ ĂŹs paper âĂIJObservational Studiesâ ĂİandComments*, page 126, 1972.

[Cox92] David R Cox. Causality: some statistical aspects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 155(2):291–301, 1992.

[CW17] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[CW18] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[DH50] Richard Doll and A Bradford Hill. Smoking and carcinoma of the lung. *British medical journal*, 2(4682):739, 1950.

[DH52] Richard Doll and A Bradford Hill. Study of the aetiology of carcinoma of the lung. *British medical journal*, 2(4797):1271, 1952.

[ESC+18] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pages 3910–3920, 2018.

[EVGW+10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[Fis58a] Ronald Fisher. Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, 2:151–166, 1958.

[Fis58b] Ronald A Fisher. Lung cancer and cigarettes? *Nature*, 182(4628):108–108, 1958.

[GBY+18] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[GF17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[GH95] Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

[Gig00] Gerd Gigerenzer. *Adaptive thinking: Rationality in the real world*. Oxford University Press, USA, 2000.

[Gra69] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[GRCVDM17] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[GVT85] Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314, 1985.

[GZS19] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[Har01] Stevan Harnad. What's wrong and right about searle's chinese room argument? In *Essays on Searle's Chinese room argument*. Oxford University Press, 2001.

[Hau97] Larry Hauser. Searle's chinese box: Debunking the chinese room argument. *Minds and Machines*, 7(2):199–226, 1997.

[HDR18] Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[Hil55] A Bradford Hill. *Principles of medical statistics*. The Lancet, 1955.

[Hil65] Austin Bradford Hill. The environment and disease: association or causation?, 1965.

[Hit97] Christopher Hitchcock. Probabilistic causation. 1997.

[HJM+09] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[HKS10] Jürgen Huber, Michael Kirchler, and Thomas Stöckl. The hot hand belief and the gambler's fallacy in investment decisions under risk. *Theory and Decision*, 68(4):445–462, 2010.

[HLHG00] Ulrich Hoffrage, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer. Communicating statistical information, 2000.

[Hol86] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[HP99] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[HRBG18] Samuel Harding, Prashanth Rajivan, Bennett I Bertenthal, and Cleotilde Gonzalez. Human decisions on targeted and non-targeted adversarial sample. In *CogSci*, 2018.

[HS03] Jennifer L Hochschild and Nathan Scovronick. *The American dream and the public schools*. Oxford University Press, 2003.

[Hum00] David Hume. *An enquiry concerning human understanding: A critical edition*, volume 3. Oxford University Press, 2000.

[Hum03] David Hume. *A treatise of human nature*. Courier Corporation, 2003.

[HV12] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

[Hyv99] Aapo Hyvärinen. Survey on independent component analysis. 1999.

[J+80] Searle John et al. Minds, brains and programs. *Behavioral and Brain Science*, 3:417–424, 1980.

[JAR82] Dennis Jennings, Teresa M Amabile, and Lee Ross. Informal covariation assessment: Data-based vs. theory-based judgments. 1982.

[JG06] Karin H James and Isabel Gauthier. Letter processing automatically recruits a sensory–motor brain network. *Neuropsychologia*, 44(14):2937–2949, 2006.

[JL17] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[JMZ+12] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[JS10] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[KGB16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[KPS18] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

[Kri15] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KT72] Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[KT07] Tevye R Krynski and Joshua B Tenenbaum. The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3):430, 2007.

[Lau96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[LeC17] Yann LeCun. My take on ali rahimi's "test of time" award talk at nips. *Facebook, https://www. facebook. com/yann. lecun/posts/10154938130592143*, pages 06–12, 2017.

[LH07] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

[LJ13] Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 2013.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[LN+81] Dennis V Lindley, Melvin R Novick, et al. The role of exchangeability in inference. *The Annals of Statistics*, 9(1):45–58, 1981.

[LPNC+17] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[LST15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[LUTG17] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[Mar18] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[Mei02] Deborah Meier. *The power of their ideas: Lessons for America from a small school in Harlem*. Beacon Press, 2002.

[Mil19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[MPJ+16] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[MT88] Frederick Mosteller and John W Tukey. Data analysis, including statistics. *The Collected Works of John W. Tukey: Graphics 1965-1985*, 5:123, 1988.

[Nic02] Raymond S Nickerson. The production and perception of randomness. *Psychological review*, 109(2):330, 2002.

[NSVM18] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*, 2018.

[PB02] John Preston and Mark Bishop. *Views into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press on Demand, 2002.

[Pea82] Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science ..., 1982.

[Pea95] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[Pea00] Judea Pearl. *Causality*. Cambridge university press, 2000.

[Pea18] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

[Pea19] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, February 2019.

[PG06] Carl V Phillips and Karen J Goodman. Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerging Themes in Epidemiology*, 3(1):1–7, 2006.

[PJS11] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.

[PM18] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[PMJ+16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[PMW+16] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[PV95] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

[PVH01] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[RDT15] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114, 2015.

[Rei56] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[Sav72] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

[Sch19] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[SGS93] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. 1993.

[SGS+01] Peter Spirtes, Clark Glymour, Richard Scheines, et al. Causation, prediction, and search. *MIT Press Books*, 1, 2001.

[SHHK06] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[Sim51] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[SJP+12] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[SJPZ11] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, and Kun Zhang. Robust learning via cause-effect models. *arXiv preprint arXiv:1112.2738*, 2011.

[SJZ+10] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in neural information processing systems*, pages 1687–1695, 2010.

[SP12] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876*, 2012.

[SWM17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Tik63] Andrei N Tikhonov. Solution of incorrectly formulated problems and the regularization method. 1963.

[TK74] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[TK77] Amos Tversky and Daniel Kahneman. Causal thinking in judgment under uncertainty. In *Basic problems in methodology and linguistics*, pages 167–190. Springer, 1977.

[TK83] Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

[TK15] Amos Tversky and Daniel Kahneman. Causal schemas in judgments under uncertainty. 2015.

[Tur09] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[Val20] Jordi Vallverdú. Approximate and situated causality in deep learning. *Philosophies*, 5(1):2, 2020.

[VBC18] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11), 2018.

[Vre15] Jilles Vreeken. Causal inference by direction of information. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 909–917. SIAM, 2015.

[Wan91] Mei-Cheng Wang. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143, 1991.

[War09] Andrew C Ward. The role of causal criteria in causal inferences: Bradford hill's" aspects of association". *Epidemiologic Perspectives & Innovations*, 6(1):2, 2009.

[WWA+01] Christina Wolfson, David B Wolfson, Masoud Asgharian, Cyr Emile M'Lan, Truls Østbye, Kenneth Rockwood, and DB ft Hogan. A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116, 2001.

[WZ06] Howard Wainer and Harris L Zwerling. Evidence that smaller schools do not improve student achievement. *Phi Delta Kappan*, 88(4):300–303, 2006.

[ZH10] Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164. PMLR, 2010.

[ZH12] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.