Seminar: How Do I Lie With Statistics?

Judea Pearl - Book of Why (Chapters 0-5) - Review

Patrick Dammann[1]

[1]MSc. Angew. Informatik, Universität Heidelberg

February 27, 2020

Abstract

In this review/summary of Judea Pearl's and Dana Mackenzie's book "The Book of Why", the most important aspects of the first five chapters are outlined. The book deals with the problems that arose when people tried to introduce causal reasoning into the field of statistics, which grew up in anticipation to causal argumentation. It gives a brief history about how it came to this development, and introduces modern methods that make the connection between both concept really easy. In the fifth chapter, the problem is demonstrated using the smoking-lung-cancer-debate from the middle of the 20th century.

Contents

1 Basics **2**
- 1.1 Ladder of Causation 2
- 1.2 Causal Diagrams 3

2 History of Statistics **4**
- 2.1 Francis Galton 4
- 2.2 Karl Pearson 5
- 2.3 Sewall Wright 5

3 Bayes and Junctions **5**
- 3.1 Rule of Bayes 6
- 3.2 Junctions in Causal Diagrams 6

4 Confounders **7**
- 4.1 RCTs 7
- 4.2 Back-Door Criterion 8

5 The Smoke-Cancer-Debate **8**
- 5.1 Does smoking cause cancer? . 8
- 5.2 Smoking and newborns 9

6 Conclusion **10**

Introduction

"The Book of Why"[1] by Judea Pearl and Dana Mackenzie is a popular science book about the introduction of real causal methods to the sciences and how they started to help liberating the statistics from mistakes of their early days. The main tool used here are *causal diagrams*, graph-based models for causal relationships that Pearl (re-)discovered and improved over the past years.

The general question about the importance of causation as opposed to the use of data on its own can be shown via a little example: Presume having data about the number of fire fighters at an operation, compared to the damage the fire caused. Usually, a positive trend should be noticeable: The more fire fighters, the more damage. But without knowing the causal connections in the background, one cannot derive from the data whether lowering the number of fire fighters sent to an operation might lower the damage done by the fires, or not.

In this summary/review, in each chapter I will focus on the two to three main points[1] and summarize them. In most (but not all) cases, I will abstain from adding my personal thought. While the numbering of my sections 1-5 fits to the book, the titles of my summaries are chosen on my own. The sixth section then contains my personal conclusion.

1 Basics

1.1 Ladder of Causation

The guiding thread through the whole book is a metaphor called the *ladder of causation*, which describes three different problem classes seated on the three rungs of the ladder, each presenting new issues that cannot be solved with methods only that would suffice for overcoming obstacles from the lower rungs. Those rungs are named *"Association"*, *"Intervention"* and *"Counterfactuals"* and bear the following structure:

The lowest rung, *"Association"*, supports all questions whose answers can be found by looking at data alone. These include questions like "If a patient has a certain symptom X, how likely is it that he suffers from disease Y?" and "If a customer bought cookies, what are the odds the he also buys milk?". These are classic statistical queries that lead to results that can only describe the relation of different observations without intervening in the observed process.

Addressing questions of this rung is natural to all higher animals, since it is needed to interact with the outside world at all.

The second rung, *"Intervention"*, contains those problems that additionally imply some external control over the situation, like doing or preventing a certain action. This can, for example, be something like "Will my headache be cured, if I take this medicine?" or "How will poverty rates behave, when we introduce this new law?". This introduces some kind of causal thinking, since the main difference to rung one is that the queries cannot be answered with observational data alone, because it shows only in which relations the variables might occur "in the wild", but not which variable influences which and therefore what would change if someone changed something.

The fire fighter example from above belongs into rung two. We can answer rung two questions like "How many fire fighter might be there, when the fire cost X$?", but to know the difference between seeing less fire fighters (maybe because of a small fire) and making less fire fighters handling a fire (which might result in bad consequences), more knowledge is needed than the data itself supplies. This could for example be the fact that both variables might have the fire's magnitude as a common cause (which obviously yields the difference be-

[1] in my eyes

tween seeing and enforcing less fire fighters) and the circumstance that fire fighters fight fires (which leads to opposing results when reducing their numbers).

Engaging in these questions requires more brainpower than the previous ones, so it is assumed that early humans were the first animals to develop the skill to grasp the consequences of their manipulation of the world, going on until today, where toddlers learn this in their early years.

The highest rung, *"Counterfactuals"*, deals with imagining worlds where things would have been different than in the current situation, hence the name COUNTERFACT(-UAL). Instead of predictions about a general, statistical population and their behaviour under "normal" or tweaked conditions, one observes a special situation (e.g. that of a special individual) and than wants to know how the outcome of that exact scene would have changed, if some details would have differed. For example, possible queries could be "Was it the medicine, that cured my headache?" (since it is equivalent to "Would my headache also have stopped, if I hadn't taken the medicine?") or "Would Kennedy still be alive, if he hadn't been shot?".

Questions from this class of problems still seem very natural to us humans, even though they involve worlds that do not exist. This is assumed to be a skill unique to humans[2], and laying the foundation of not only all fictional story-telling, but also of human inventions at all. This is due to the fact that these questions are needed for understanding the world, since it is essential for comprehension of e.g. a method, to know what would have happened if the method wasn't applied or applied differently.

Since the book's author, Judea Pearl, looks at and partially illuminates the issue from a computer scientist's point of view, he compares these different levels to the capabilities of modern, artificial intelligence and finds two things:

On one hand, a strong AI that learn like and interact with humans on a natural level needs to be able to answer questions from all three rungs, since they are necessary for an understanding of the world and so deeply woven into the human mind that a robot's lack of understanding would limit the possibilities of easy communication greatly. On the other hand, state-of-the-art machine learning models still reside on rung one, in some cases maybe scratching on the bottom of rung two, because most training algorithms do not involve the modeling of the world but only associating from what is seen. Therefore, a strong AI seems further away than the current, sci-fi-esque era might suggest and will need more research involving causal reasoning in machine learning.

1.2 Causal Diagrams

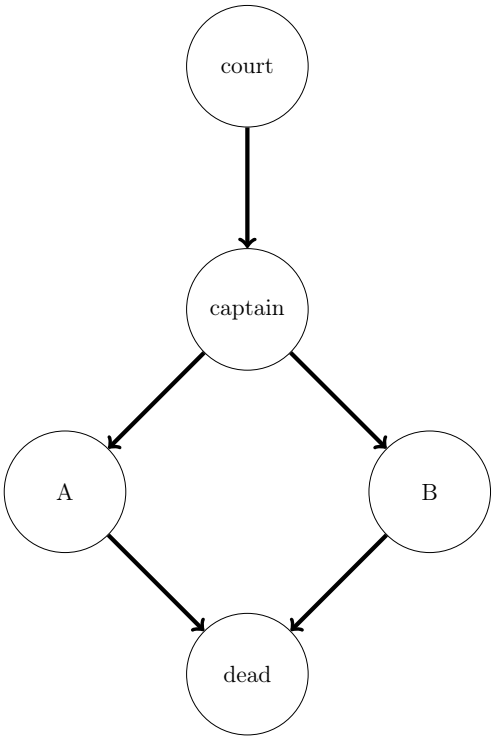
To either work with and analyze causal models or implement them in learning machines, they need a representation. And since causality consists of asymmetric relations between different variables, we can illustrate a model as a directed graph where the nodes symbolize random variables (measurable and not) and every direct causal effect from one variable to another is shown through a directed edge. A graph with these features is called *"causal diagram"*, a method highly utilized in this book.

In the following example, which is highly simplified to explain the nature of those causal diagrams, all variables are boolean (true/false) and if there is a causal effect between two variables, it is of the type "if A happens, B happens".

The idea is the modelling of a shooting squad and the causal relationships are easy: Iff the court orders the prisoner to die, the captain commands the shooting. Iff the captain commands a shooting, A shoots. Iff the captain commands a shooting, B shoots. Iff either A or B shot, the prisoner dies. There are no jammed guns, pacifistic or rampaging

[2] At least on earth..!

soldiers, missed bullets, suicidal prisoners or anything else in this scenario.



Even with a simple situation like this, we can pose many different questions from the different rungs. For rung one questions, we could for example ask "Is the prisoner dead when the court gave the order?", and following the diagram and applying the rules yields the answer. We could also ask "Did A shoot if B shot?", and even though there is no directed path between the two, we can see that there is no other possibility than yes. Both queries can be noted in statistical terms, like $P(dead|court)$ or $P(A|B)$.

Queries from rung two could include questions like "If A decides to shoot on his own, is the prisoner dead?". Even though the question defies the rules by letting a soldier shoot without the captains command, we can give an answer by manipulating the scenario in a minimal way that fits or situation while keeping everything else unchanged, than simulating everything like in the first examples. That manipulation can be implemented by removing any incoming arrows into the variable that "decides" or "is made to" and setting it to a fixed value, thus erasing all external influences on it. Since classic statistics does not work on causal relationships but data alone, this process cannot be expressed by classic notation, so a new operator is introduced, the do-operator: $P(dead|do(A = true))$. The answer to this query is obviously yes, since the still applied rules make prisoner die even on a single bullet.

To generate queries from rung three, a given situation must be described, followed by a question that involves fact that contradict the situation. For example, it could be asked: "The prisoner is dead. Would he also be, if soldier A had not shot?" This kind of questions also needs a new notation, for we cannot express this through the do-operator. Instead, we use: $dead_{A=false}(dead = true)$. With human logic, one can clearly see that the prisoner would of course be dead, too, but an algorithm to tackle problems like this with causal diagrams is also explained later in the book.

2 History of Statistics

To illustrate why the methods utilizing causal diagrams are so important and to explain how the science of statistics could maneuver itself into the situation that causality was mainly ignored, causal reasoning sometimes even considered unscientific, Judea Pearl tells the stories of three scientists that left a significant mark on the history of causality and statistics.

2.1 Francis Galton

During a lecture in 1877 for the *"Friday Evening Discourse"* at the Royal Institution of Great Britain in London, a regular and exquisite lecturing event since the 19th century, Francis Galton revealed the first great insight needed for statistics. Particularly, he stuck metal balls into a self-designed board with regularly spaced nails. The great ob-

servation here is that, while all individual balls seem to behave very chaotically, the whole population of all balls acts in a very predictable way, a bell shaped curve.

Since human body sizes are distributed in the same way, he assumed that they both behave the same way. While investigating this incident, he accidentally stumbled upon the discovery of the regression to the mean by realizing that tall fathers tend to have tall sons, but not as tall as their fathers. First, he thought it to be a phenomenon of human sizes and genetics, but when he realized that it also concerned tall sons and their tall (but not as tall) fathers, and finally found the same anomaly in all kinds of statistical data.

During this, he also came upon correlation of random variables and regression, which not only enabled the newly born statistics to thrive, but also to to gain wonderful results without the need of causal reasoning.

2.2 Karl Pearson

While Galton discovered correlation accidentally on his search for causation, Pearson intepreted his work differently. For him, correlation was the bigger and mightier concept and correlation only one of its special cases; an interpretation which was deeply influenced by his personal philosophical position.

He saw the great opportunity in statistics to finally integrate mathematical foundation into other sciences and after Galtons death, Pearson's new Biometrics Lab became the center of statistics. But according to Judea Pearl, Pearson's whole movement (even accompanied by a successful journal he published) felt more and more like a sect over the time, that listed causation as one of their deadly sins.

This refusal of all causal thinking even developed ridiculous situations, for example when Pearson and others discovered "spurious" correlations, like between chocolate consumption and number of Nobel Prize winners per country, where they could not just blame this correlation on a common cause (since causes are irrelevant to science), or the correlation that occurs when two different populations are mixed. But still, the adherents of this cause-free movement managed to shape the following time strongly.

2.3 Sewall Wright

Being a fresh scientist in the field of genetics in the beginning of the 20^{th} century, Wright found himself with the task of figuring out what factors influence the development of specific fur colors in guinea pigs. With the information present to him, he was able to draw a diagram that showed which factors have an impact on which other factors, and thus created the first known causal diagram.

After his paper about his findings, he also published another paper that elucidated the usage of his methods to solve other problems. Immediately supporters of Pearson lectured him about the meaninglessness of causation and how his methods do not work properly, without having really understood them. Luckily, Wright was not demotivated by this hard reaction, but continued improving and advertising his method that allowed to use assumptions about causal relationships to either move a rung 2 two problem down to rung 1 or prove the assumptions wrong.

3 Bayes and Junctions

This chapter focuses mainly on explaining two issues very vividly, at first the rule of Bayes, a highly important theorem in statistics, then the implication of certain junctions in causal networks, a viewpoint, that helps to identify most of its properties by looking only at three-node sub-graphs. He also gives small insight in the mechanisms of Bayesian Networks, a model that i.a. allows to feed a causal network with data, and how they could be used to identify victims after a tragical plane crash, but in my point of

view, this is not the spotlight of this chapter

3.1 Rule of Bayes

To understand the importance of the rule, it is crucial to grasp the idea of forward and backward probability. Here, this is explained with a pool table example: Imagine a pool table of length L. When hitting a billiard ball so hard that its final position can be assumed to be random, the odds of it stopping in the first 1m of the table is calculated by $P(x \leq 1|L = l)$.

But if the question posed now is instead "Imagine the ball stopping in the first 1m. How likely is is, that the length L of the table is l?", the probability $P(L = l|x \leq 1)$ is much harder to calculate. The first one is called *"forward probability"*, and in most cases those probabilities feel more natural to us humans an describe the causal direction of a process, the other one *"backward probability"*.

The Rule of Bayes tries to bring the two probabilities into context, by stating the following formula:

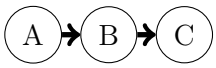
$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

When rearranging the formula to $P(Y)P(X|Y) = P(X)P(Y|X)$, one can clearly see the sense this equation makes, since the probability of e.g. finding a certain pair of attributes in a population should stay the same, no matter weather we filter for one attribute first, than for the in the remaining population, or the other way round.

3.2 Junctions in Causal Diagrams

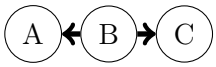
A lot of the analysis of causal networks deals with the investigation of paths through the graph. On these paths, every inner node can be seen as the middle node of a three-node sub-graph, called a *junction*. These junctions can be classified into three groups that behave very differently. To understand the difference, the concept of "controlling" for variables must be known. Ultimately, this means nothing more than clustering the data into sub-divisions in which all data-points share the same value (or lay in the same, small interval) regarding the variable controlled for.

Chain



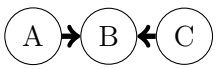
A chain has on directed path through all three nodes. The middle node here is called *"mediator"*, since it is transferring causal information from A to C. As an example one could imagine A to be a fire having started, B the smoke coming up, and C the smoke detector starting an alarm. Since the detector does not detect fire on its own, there is no direct link between A and C. The interesting thing here is that information can flow from A to C, but stops as soon as B is controlled for, meaning that e.g. only scenarios are considered where there has been smoke. In this case, A and C loose their correlation and therefore look like uncorrelated variables.

Fork



In so called forks, both A and C have a common cause B, which is called a *"confounder"*. Confounders are dangerous, because they introduce spurious correlations between variables, that are not directly causally connected. Luckily, controlling for confounders eradicates this correlation. For example, if one looks for children's shoe size (A), and their reading ability (C), they seem strongly, positively related. But when controlling for their confounder, age (B), the correlation vanishes, since older children tend to read better and have bigger feet, while for children in the same age, bigger feet do not automatically make them read better (or vice versa).

Collider



The last class of junctions are colliders, where both A and C have a causal impact on B. This is the most interesting junction, since here A and C are (usually) uncorrelated, but controlling for B makes them (seem) correlated. This effect, also called the "explain away"-effect, can be viewed from the following perspective: Imagine being famous (B) can be only achieved by being beautiful (A) and talented (C). When now looking only at the N most famous people, it seems like talented people were less attractive and vice versa. This artifact emerges from the fact, that if a person is already really good-looking, he or she does not have to be that talented to reach the Top-N of famous people. His or her talentlessness is "explained away".

4 Confounders

The forth chapter of the book explains the connection between randomized controlled trails and causal analysis and finally states a simple algorithm for annihilating confounders in data.

4.1 RCTs

A great problem that occurs when all sciences handle their data using mathematical tools that do not allow any causal statements, is that you cannot make causal statements. However, those statements are needed in practically all sciences. Imagine medicine, where one needs to know whether a certain drug causes a disease to vanish or if a special symptom is caused by a certain disease.

To tackle this issue, randomized controlled trials (short: RCTs) have been developed. Here, a large population is randomly split into groups, with each group getting a different treatment, including a so-called control group, that receives no treatment at all. Analogical to the do-operator, in a perfect setting the randomization of the groups eliminates all influences on the chosen treatment and thus possible confounders, like a certain condition making a subject both more attracted by the treatment and more likely to develop the desired result on its own.

The example given in the book is that of R. A. Fisher, who had the task of helping farmers by developing a method to test which fertilizer produces the most yield. While the yield is influenced by many factors, the choice of the fertilizer is, too. Instead of just dividing the field into $(N + 1)$ subdivisions to test the N fertilizers, the field is split into many more parts and instead of letting someone choose which fertilizer to use where, the positions are randomly chosen by a shuffled deck of playing cards. This way, even if there are clusters of the field where certain properties influence the yield in a special kind, it is highly likely that this cluster is partly occupied by all fertilizers, compared to a negative or positive effect on only one group, when the cluster would lie completely in this groups $\frac{1}{N+1}$-th of the field. The shuffled deck assures a good distribution of the groups across the field while preventing everyone from tainting the experiment with unconscious introduction of personal experience.

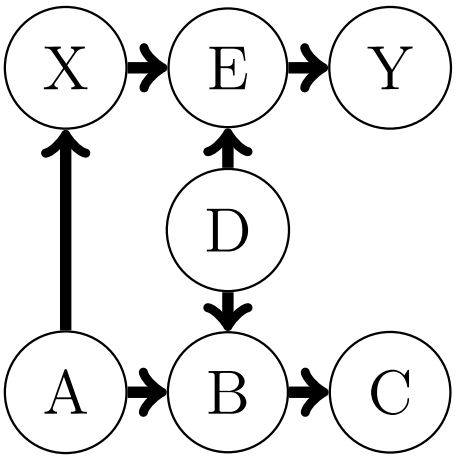
A long time, randomized controlled trails were considered the only tool that allowed to draw a causal conclusion from statistical data. But RCTs also have downsides: there are many situation, where the implementation of an RCT is not applicable, for example for ethical or financial reasons. Also, the randomization does not completely eradicate confounders. This can, for example, occur through the choice of the subpopulation to run the RCT on. This introduces a whole new problem: Even in RCTs, scientists still need to control for possible confounders. This can lead to a lot of mistakes, not only through confounders that are not controlled for, but also via accidentally controlling for a mediator or a collider, and thereby erasing important correlations or introducing spurious correlations that should not be there.

4.2 Back-Door Criterion

One great thing about causal diagrams is that (if based on correct assumptions), one can easily use it to determine confounding inside a system of random variables. This is done by using the junctions from Section 3 to block paths in the causal diagram, e.g. by controlling for confounders in forks and mediators in chains or not controlling for colliders. Which paths are to be blocks follows simple rules:

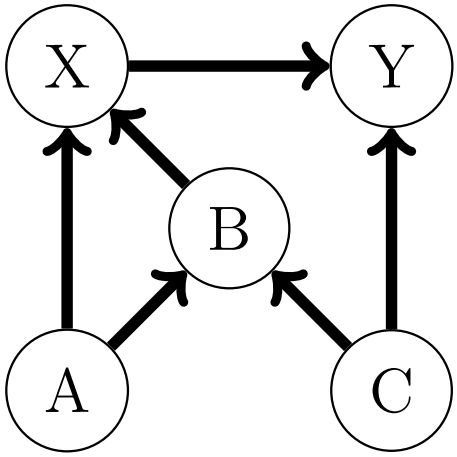
Assume we have a causal diagram and want to examine the impact of X on Y. A *"back-door path"* is an undirected[3] path from X to Y, where the first arrow points into X. The idea of this algorithm is to close all back-door paths, while assuring that the variables that get controlled for are not reachable via a directed[4] path from X. When this is achieved, X and Y do not share a common cause anymore, while the information flow between the two stays intact. If some of the variables to control for are unobservable, and there is no other set of control variables without this flaw, the system of variables cannot be deconfounded.

Here is a small example of a causal diagram from the book:



We can see the back-door path X←A→B←D→E→Y, but luckily it is already blocked by the collider around B. Since there are no other back-door paths, the situation is unconfounded, as long as B is not controlled for.

Another example from the book loos like this:



In this case, we can discern two different back-door paths: X←A→B←C→Y and X←B←C→Y. While the first path is blocked by the collider around B, the second path needs to be closed. We could achieve this by controlling for B without breaking any rule, but that would open up the first path. Therefore is would be better to close it by controlling for C instead.

5 The Smoke-Cancer-Debate

5.1 Does smoking cause cancer?

As stated before, except for randomized controlled trials, statistics did not grant any causal conclusions. This posed a big, unanswerable question in the end of the 1950s: *Does smoking cause lung cancer?* There was time-series data showing a rise in lung cancer cases right after the rise in cigarette a clearly noticeable correlation between the heaviness of smoking and the development of lung cancer, so the connection between the two could not be unseen, but the causation behind the correlation was important!

If there was a causal path from smoking to cancer, not smoking would lower the

[3] the direction of the arrows does not matter
[4] only following direction of the arrows

risk of cancer and therefore some legal steps could be initiated, like banning cigarettes or at least controlling the situation through taxes and education.

If not, say because there existed a gene that makes you both love cigarettes prone to lung cancer, and the raising case counts emerged only from other temporal factors like the newly started tarring of roads, there would not only be no need for any new, restricting laws, but also there was no reason for people to stop smoking at all, since it would not change a thing. Obviously, the tobacco industry took this side in the discussion.

Adding up to a huge concern propagating the harmlessness of cigarettes, the discussion was also very emotional for many of the involved, since nicotine is a highly addictive drug and addicts tend to defend their drug very impassioned. Therefore, clear facts without room for interpretation needed to be stated to convince on side of the other, and the best fact-producing mechanism known to mankind, mathematics, was paralyzed by its responsible field, statistics, anticipating the posed question itself.

The only way known so far to gain causal conclusions on a mathematical foundation were RCTs. Unfortunately, this was one of the cases where RCTs were not only ethically not applicable, but also not fast enough. To gain reliable results, the supervisors of the RCT had to split a large population randomly into people that are forced to smoke regularly in the next 40 years, and people who are completely forbidden to smoke in that time at all. Not only that making people potentially damage their bodies for half a life-time feels just wrong from a humane perspective, also could many people still die from lung cancer due to false information during the long time a trial like this takes.

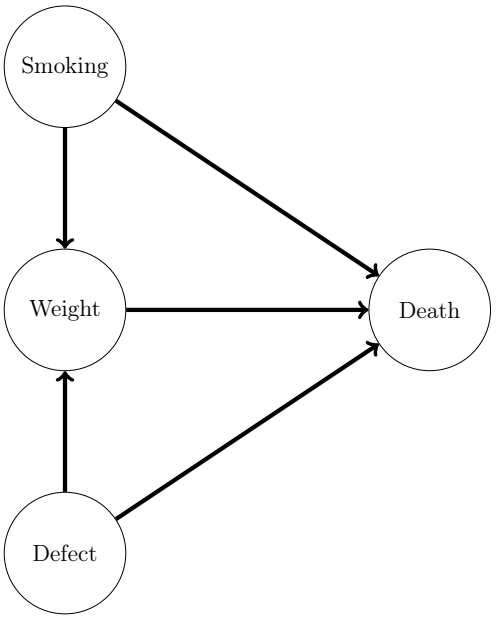
Luckily, after some time, Jerome Cornfield found a mathematical way of making the explanations and the point of view of the tobacco industry too highly unlikely to believe in, thus closing the debate. His explanation was as easy as this: Assume there was a gene that fully accounts for the lung cancer found in smokers. Since lung cancer is found nine times more often in smokers than in non-smokers, this gene must occur nine times more often in smokers than in non-smokers. That means that when 11% of all non-smokers have that gene, 99% of all smokers had it. With more than $11.\overline{1}\%$, the full accusation of the gene was mathematically impossible. Since such a tight, nearly deterministic link between a gene and a persons decision to smoke was seen as highly unlikely, the theory of the full-accountable gene was refuted, only leaving theories behind, where smoking has a causal effect on lung cancer.

After the discussion was brought to a conclusion, a committee with members of all different fields was founded to elaborate methods on how to tackle situations like this in the future. Their main outcome was the fact, that statistical methods are not sufficient for spotting causal relationships, but also a list of criteria that might give a good hint on a causal link. This list, containing criteria like consistency over different populations, specificity of the cause to a certain effect and temporal order (effect after cause), was later revised and expanded by Austin Bradford Hill and therefore became known as *"Hill's Criteria"*. The list is far from perfect, and the presence or absence of individual criteria must not mean anything regarding the causal relation, but the presence of most aspects gives a strong sign in the causal direction and without causal diagram, this was a long time the best measure present.

5.2 Smoking and newborns

Another controversy regarding smoking was detected only a little later, but then not solved for over 40 years. The scenario is the following: In a study, for many thousands of children, information was gathered about their condition after birth and whether they

died soon after. Logically, the children of smoking mothers had lower birth weight and higher child mortality. The paradox finding that was made here was the fact that, if only looking at underweight babies, the babies of smoking mothers had a much lower mortality that those of non-smoking mothers. When taking into account that there are various genetic defects, that decline birth weight and cause a child to die early, looking at a causal diagram describes and solves the problem easily:



The birth weight here serves as a collider. By controlling for it, we allow information to flow in the direction Smoking→Weight←Defect→Death. Thus, the explain-away effect holds. When an underweight baby has a smoking mother that lowers the chance of the baby having a severe birth defect compared to the other underweight babies where there is no know reason that "explains away" the low birth weight. That means that the baby has a lower chance to die that the babies with birth defects that will kill them for sure.

6 Conclusion

To state my first impression of the book is pretty simple: *I was shocked*, especially from the historic parts. Never had I thought that there has been so emotion, personal preference and struggle for power in a field as pure and logic-driven like mathematics. Maybe I was just to naive up top this point, but this shows once again that even becoming a professor in mathematics does nothing near guarding you from your human flaws.

But then again, reading the book also made me happy, because this is a bad situation, that science and humanity are about to overcome, and as a sideline producing methods that are easy to use, especially for a computer scientist[5].

Pearls argumentation that a strong AI needs to be build upon statistical methods that include causal inference makes perfect sense to me. Because nearly every thought, every plan, every wish that passes our brain contains some aspect that differs from how the world really is, I clearly see the necessity of robots being able to grasp rung three problems for being able to communicate with humans on equal terms, not even talking abut the ability to understand the world, which is incredibly important since humans can't model the world perfectly for a robot.

Finally, causal diagrams as a model for causal relations seem like a jack of all trades to me. It is simple, yet compiling the whole needed information into a little, compressed graph, on which then thousands of already existing graph algorithms can be used. Also, they motivate scientific exploration, compared to the classic statistical cookie-cutter-methods used in other sciences. They invite to play around with them to find out how the world works, using data mainly for verification (and later results), which sounds kind of romantic to me and definitely worth looking into them again some time.

[5] we all love graphs, I'm always told

References

[1] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. 1st ed. USA: Basic Books, Inc.; 2018.