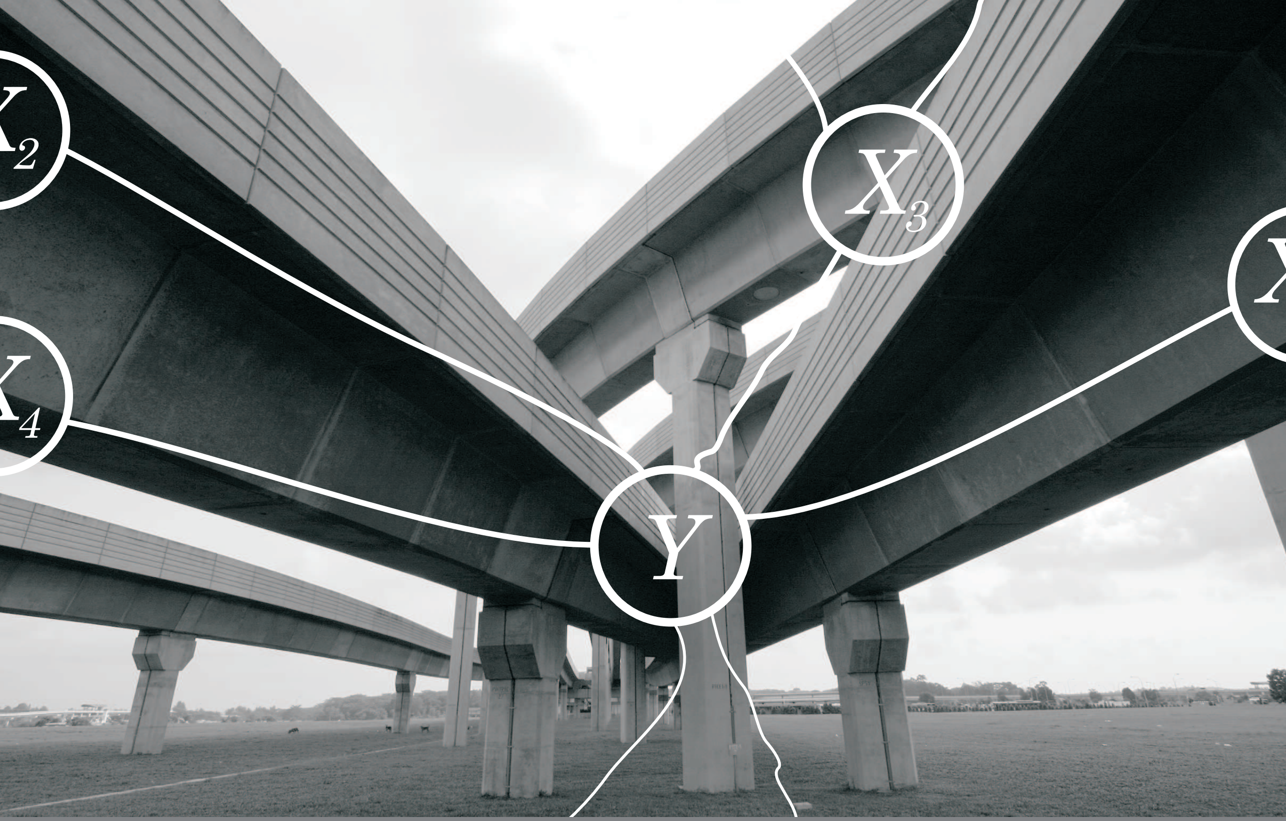

Concrete Causation

About the Structures of Causal Knowledge

Roland Poellinger

Concrete Causation

About the Structures of Causal Knowledge

Inaugural-Dissertation zur Erlangung
des Doktorgrades der Philosophie an der
Ludwig-Maximilians-Universität München

vorgelegt von

Roland Poellinger, München
`http://logic.rforge.com`

Referent: Prof. Dr. Godehard Link
(Lehrstuhl für Philosophie, Logik und
Wissenschaftstheorie, LMU München)

Korreferent: Prof. Dr. Thomas Augustin
(Institut für Statistik, LMU München)

Tag der mündlichen Prüfung: 13.02.2012

Alexander von Humboldt
Stiftung/Foundation

MindMapsCause

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

I would like to thank the Alexander von Humboldt Foundation, which partially supported my work through funding the Munich Center for Mathematical Philosophy (MCMP) at LMU Munich. I am especially thankful to the LMU working group MindMapsCause, that has shaped many of my thoughts in valuable discussions.

Title image: tommy leong/Fotolia.com, R. Poellinger

Contents

Chapter 1

Reasoning about causation

Felix qui potuit rerum
cognoscere causas

VERGIL, *Georgica* (II, 490)

Philosophers have been thinking systematically about cause and effect since the very beginnings of philosophy as a discipline. Availing itself of mathematical methods and formal semantics in the last century, epistemology at once had the means to shape prevailing problems in symbolic form, express its achievements with scientific rigor, and sort issues within formal theories from questions about intuitions and basal premisses. David LEWIS was among the first ones to utilize symbolic tools and approach causality within a framework of formal semantics.[1] After Bertrand RUSSELL had famously and brusquely turned his back on any further pursuit of establishing criteria for causal analysis in his treatise *On the Notion of Cause* (1913), David LEWIS re-thought the words of an earlier mind: In 1740 David HUME had listed causation among one of the principles that are "to us the cement of the universe" and thus "of vast consequence [...] in the science of human nature".[2] HUME gives various hints about what his account of causation might be – one way of reading suggests that he argues for an innate human causal sense by which we discover the relation of causation in our surroundings.[3] Although HUME is later sharply criticized for this empiricist account by Immanuel KANT, who in turn claims that causal principles are of *synthetic a priori*

[1] Cf. especially [Lewis 1973a].

[2] These statements are from *An Abstract of a "Treatise of Human Nature"*.

[3] Cf. [Garrett 2009].

nature,[4] David Lewis refers back to one specific counterfactual explication of the semantics of causal statements in Hume's writings, moreover bases his thoughts on Humean supervenience, and unfolds a detailed method for causal analysis in the framework of his *possible worlds semantics.* Approaching the field from a computer science perspective in the 1980s, Judea Pearl introduces networks of belief propagation as the basis for Bayesian inference engines in an AI engineering context.[5] His *interventionist* account of causation, most elaborately presented in his book *Causality* (2000/2009), draws on structural transformations of formal causal models for the identification of cause-effect relations. As a defendant of a *thicker* concept of causation, Nancy Cartwright decisively rejects Pearl's *thin*, formal approach and makes a case for a family-like understanding of causal concepts.

In the following chapters the line of thought from Lewis to Pearl shall be traced, partly by examining their replies to one another, before I want to make the attempt to locate *causation* and *causality* in the ontological landscape and try to pave the way for an epistemic understanding of the relation of causation, finally applying this conception to examples from recent and older philosophical literature. An overview on ways of implementation and applications of the suggested methods will conclude this text. Before getting into technical details, a short list of important approaches towards the analysis of causal concepts (and their most prominent advocates) shall be given – especially as a point of reference and distinction for what follows. What suggestions have fueled the philosophical discussion?

1.1 Causal powers

One metaphysical approach towards causality, which has recently gained interest again, is the ascription of essential causal powers or capacities to objects of reality.[6] As an answer to the Humean view of the world as consisting of distinct and discrete objects, causal powers theorists argue for the metaphysically real category of dispositions, which are necessarily separate from their token instantiations but at the same time linked to those instantiations of themselves through a necessary causal relation. Before this background, powers are seen like enduring states with

[4] Cf. [Watkins 2009] or also [de Pierris & Friedman 2008].

[5] Cf. e. g. [Pearl 1982].

[6] Cf. for this and the following [Mumford 2009].

the hidden disposition to objectively produce events or states by singularly contributing observable quantities to their manifestations – most times in combination with other contributing or also counteracting powers. One question that arises within this framework seems to be the question about the nature of the connection between powers and their manifestations. Can one realistically postulate a certain disposition if it, for example, never manifests itself? And if one sees causation as an asymmetric relation, is there a way to understand the *directedness of powers* as necessary *causal* directedness from cause towards effect? Controversial questions seem to remain open as yet, but if powers of this sort are understood as basic building blocks of reality, one need not stick to events as relata of causal claims – e. g., explanations of equilibria (two stellar bodies orbiting one another at a stable distance and like examples) are easily given by determining the contributions of each power to the situation under examination. And as Cartwright claims, general causal statements are best understood as compact statements about the capacities involved, as in "aspirins relieve headaches."[7] Finally, in distinction from other theories of causal relations, the main goal of the theory of causal powers is to say what and where causality really is, and – from the point of view of causal powers theorists – thus distances itself as a metaphysical enterprise from other theories that only settle for a description of the symptoms of (supposedly existing) actual causation.

Another contribution to this line of reasoning was made by Karl Popper in 1959. Popper argues against the nowadays so popular subjective interpretation of probability in favor of an objective yet not frequentist interpretation of probability with dispositional character as "a property of the generating conditions".[8] He compares these *propensities* to physical forces:

> *I am inclined to accept the suggestion that there is an analogy between the idea of propensities and that of forces – especially fields of forces. But I should point out that although the labels 'force' or 'propensity' may both be psychological or anthropomorphic metaphors, the important analogy between the two ideas does not lie here; it lies, rather, in the fact that both ideas draw attention to* unobservable dispositional properties of the physical world*, and thus help in the interpretation of physical theory.*[9]

[7] [Mumford 2009, p. 272] refers with this example to Cartwright's *Nature's Capacities and Their Measurement* (1989, Oxford: Clarendon).

[8] Cf. [Popper 1959, p. 34].

[9] Cf. [Popper 1959, pp. 30–31].

This view of (conditional) probabilities as causal dispositions has been famously criticized in 1985 by Paul Humphreys, who replies to Popper with a detailed illustration of an argument that shows how the determination of dependency between variables must fail for the propensity interpretation of probability – due to the fact that dependency is necessarily not symmetric for propensities unlike as for standard probabilities.[10] Still, Popper's thoughts have stirred notable dispute and provoked refinements of his deliberations up to now.[11]

1.2 Causal processes

At the core of process theories of causation lies the explication of causal processes and interactions, seen as more fundamental than the causal relation between events.[12] Initial versions of this programmatic move can be traced to Wesley Salmon, who – replying to Carl Hempel's deliberations on scientific explanation – grounds his own theory of explanation on causal relations and argues against subjective or agent-relative approaches towards causation for an objective account. Avoiding the question of what exactly it means to be an *event*, Salmon defines causal processes in a first version of his theory by introducing the principle of mark transmission:

> *MT: let P be a process that, in the absence of interactions with other processes would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the spacetime points A and B ($A \neq B$). Then, a mark (consisting of a modification of Q into Q^*), which has been introduced into process P by means of a single local interaction at a point A, is transmitted to point B if P manifests the modification Q^* at B and at all stages of the process between A and B without additional interactions.*[13]

He goes on by explicating the concept of *causal interaction*:

> *CI: Let P_1 and P_2 be two processes that intersect with one another at the spacetime point S, which belongs to the histories of both. Let Q be a characteristic that process P_1 would exhibit throughout*

[10] Cf. [Humphreys 1985].

[11] See e. g. [Albert 2007].

[12] Cf. for this and the following [Dowe 2009].

[13] Dowe refers with this quotation in [Dowe 2009, p. 217] to p. 148 of Salmon (1984): *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press).

> *an interval (which includes subintervals on both sides of S in the history of P_1) if the intersection with P_2 did not occur; let R be a characteristic that process P_2 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_2) if the intersection with P_1 did not occur. Then, the intersection of P_1 and P_2 at S constitutes a causal interaction if (1) P_1 exhibits the characteristic Q before S, but it exhibits a modified characteristic Q' throughout an interval immediately following S; and (2) P_2 exhibits R before S but it exhibits a modified characteristic R' throughout an interval immediately following S.*[14]

Now, what it means to be a causal process within this proposed framework is best understood by considering an example Dowe gives as an illustration: A billiard ball moving across a billiard table *is* a causal process, because the ball can be marked physically, e. g., by applying some chalk to it, and this mark is transmitted throughout the entire movement. On the contrary, the movement of a shadow cannot be understood as a causal process along these lines, because the shadow itself cannot be marked physically, and no persisting substantial feature can be made out as part of its appearance. Moreover, the collision of two billiard balls *is* a causal interaction – the two balls change speed and direction of movement but would have continued to move on unimpededly had the collision, i. e., the causal interaction, not taken place.

Especially due to various criticisms of the counterfactual core of both definitions above, which seems to shift the whole burden of justification to some semantics of counterfactual statements, Salmon eventually became dissatisfied with this approach towards the analysis of causal processes and set out to rebuild his theory on the basis of the concept of *conserved quantities* together with Phil Dowe. In an effort to distinguish actual causal processes from other subjectively perceived non-causal pseudo-processes, Dowe states a possible explication of the concepts above in terms of conserved quantities:

> *CQ1. A* causal interaction *is an intersection of world lines that involves exchange of a conserved quantity.*
> *CQ2. A* causal process *is a world line of an object that possesses a conserved quantity.*[15]

[14] As above, Dowe refers with this quotation in [Dowe 2009, p. 217] to p. 171 of Salmon (1984): *Scientific Explanation and the Causal Structure of the World.*

[15] Dowe refers with this quotation in [Dowe 2009, p. 219] to Dowe (1995): *Causality and Conserved Quantities: A Reply to Salmon* (*Philosophy of Science* 62: 321–33).

DOWE goes on by saying what these conserved quantities might be essentially. He states that "[...] current scientific theory is our best guide as to what these are: quantities such as mass-energy, linear momentum, and charge."[16] (CQ1) and (CQ2) might be close to a modern understanding of physical mechanisms, but they return in most cases too many cause candidates when queried in causal analysis. Refinements of the theory with definitions of actual causal connections are discussed controversially, especially since the "theory is claimed by both Salmon and Dowe to be an *empirical analysis*, by which they mean that it concerns an objective feature of the actual world, and that it draws its primary justification from our best scientific theories."[17] The question might be justified, whether such an analysis is – against its initial program – merely introducing metaphysical overhead into physical theories that seemingly do cope well without formalized causes? Common sense causal statements, as well as statements about mental or historical causation, can only be analyzed before the backdrop of an elaborate reductionist approach. The same is true for cases of causation by omission and the likes. One way to uphold this specific approach towards causal analysis through causal processes would supposedly be supported by Nancy CARTWRIGHT, who would see this theory as contributing to an holistic understanding of a multifarious entity, "because causation is not a single, monolithic concept. There are different kinds of causal relations imbedded in different kinds of systems [...]. Our causal theories pick out important and useful structures that fit some familiar cases – cases we discover and ones we devise to fit."[18] Causation in the natural sciences, she claims, is best traced in laboratory-like settings and under specifically described conditions.

1.3 Natural experiments

Nancy CARTWRIGHT refers back to Herbert SIMON in making her point for an approach towards causal understanding that is aware of the methodology we employ to assess settings on the search for causal connections:

> *If we want to tie method – really reliable method – and "analysis" as close as possible, probably the most natural thing would be to reconstruct our account of causality from the experimental methods we use to find out about causes [...].*[19]

[16] Cf. [Dowe 2009, p. 219].

[17] Cf. [Dowe 2009, p. 223].

[18] Cf. [Cartwright 2004, p. 805].

[19] Cf. [Cartwright 2004, sect. 2.4, p. 812].

In this short quotation the general tendency Cartwright argues for becomes obvious. In her eyes, transferring causal knowledge from narrowly defined lab conditions to situations of larger scale or everyday experience cannot follow one single principle. On the contrary, she emphasizes the fruitfulness of binding our causal knowledge to our knowledge about the methodology used for providing us with initial data about causal dependencies – *naturally* as diverse in character as the methodology applied itself.

1.4 Logical reconstruction

In reply to purely physical, "Humean" views of causal analysis and – at the same time – to naive regularity accounts that try to identify causes as events which are necessary and sufficient for the occurrence of later events, J. L. Mackie develops a structured logical form to define causal efficacy. He does so by introducing "the so-called cause [as] an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result."[20] Mackie's inus condition is defined as follows:

> *A is an* inus *condition of a result P if and only if, for some X and for some Y, $(AX \text{ or } Y)$ is a necessary and sufficient condition of P, but A is not a sufficient condition of P and X is not a sufficient condition of P.*[21]

Moreover, he identifies a set of criteria as supposed truth conditions of singular causal claims such as "A caused P":[22]

> *(i)* *A is at least an* inus *condition of P – that is, there is a necessary and sufficient condition of P which has one of these forms: $(AX \text{ or } Y)$, $(A \text{ or } Y)$, AX, A.*
>
> *(ii)* *A was present on the occasion in question.*
>
> *(iii)* *The factors represented by the 'X', if any, in the formula for the necessary and sufficient condition were present on the occasion in question.*
>
> *(iv)* *Every disjunct in 'Y' which does not contain 'A' as a conjunct was absent on the occasion in question.*

As a refinement, clause (i) is later enhanced by relativizing it to a so-called *causal field*, which sets the background of discourse and indicates, in relation to which setting the cause candidate does make a difference:

[20] Cf. for this and the following [Mackie 1965].

[21] Cf. [Mackie 1965, p. 246].

[22] Cf. [Mackie 1965, p. 247].

> *(ia) A is at least an* INUS *condition of P in the field F – that is, there is a condition which, given the presence of whatever features characterize F throughout, is necessary and sufficient for P, and which is of one of these forms: $(AX$ or $Y)$, $(A$ or $Y)$, AX, A.*[23]

An example from Mackie's text shall serve as an illustration of the concepts involved: Consider the causal statement "the short-circuit caused the house to burn down." In this statement, the short-circuit A might be considered to be an INUS condition for the result P (the burning of the house) because it can be analyzed as an insufficient but necessary part of the expression $(AB\overline{C})$, where B is the conjunction of possible other contributing factors (the presence of inflammable material, oxygen, etc.) and $\overline{C}$ stands for the absence of other impeding factors (a broken sprinkler, the fire alarm being defect, etc.). $(AB\overline{C})$ in turn can then be understood as one of the disjuncts that are individually unnecessary but jointly sufficient and necessary for the occurrence of the result P – with Y consisting of further possible circumstances $[(A'B'\overline{C'}) \vee \ldots]$ that might cause the house to burn down in other ways (stroke of lightning, arson, etc.). A causal field F indicates the context in which such a causal claim is uttered. In this example, the history of the house F serves as the background before which the short-circuit A does make a difference and does trigger a change of state. A different context F' would maybe physically partition the house, thereby emphasizing that the whole house burned down as opposed to only parts of it. Mackie is subsequently forced to base his explication of cause on basal universal propositions for both generic and singular causal claims (which in turn can only be understood in terms of counterfactual dependence).[24] Since his formulation does not hinge on the full declaration of Y (oftentimes not even of X), the proposed account somehow mirrors everyday causal talk more than fine-grained physical explanation. He emphasizes that

[23] Cf. [Mackie 1965, p. 249].

[24] J. L. Mackie illustrates the transition from generic causal claims, based on universal propositions that contain information about the necessary and sufficient conditions for the situation under examination, to singular causal claims by rephrasing the short-circuit example in [Mackie 1965, p. 254]:

> *Thus if we said that a short-circuit here was a necessary condition for a fire in this house, we should be saying that there are true universal propositions from which, together with true statements about the characteristics of this house, and together with the supposition that a short-circuit did not occur here, it would follow that the house did not catch fire.*

"much of our ordinary causal knowledge is knowledge [of] incomplete universals, of what we call elliptical or *gappy* causal laws."[25] Mackie's causal principles could thus – as principles of information transfer – be carried over to reasoning about mental causation or human action, were it not for various criticisms, especially about the purely logical program pursued with the inus condition.

Judea Pearl discusses the inus condition approach in detail when reflecting on *the insufficiency of necessary causation* in his book *Causality* (2000/2009).[26] Pearl makes out two main flaws of the logical account. The first surfaces when at-least-inus propositions such as $A \rightarrow P$ are reformulated via contraposition, thereby conserving their truth value: $\neg P \rightarrow \neg A$, where '→' is to be read as 'results in.' In this case it turns out that the negation of the effect results in the negation of the initial inus condition. $\neg P$ becomes an at-least-inus condition of $\neg A$. Or in Pearl's words: "This is counterintuitive; from 'disease causes symptoms' we cannot infer that eliminating a symptom will cause the disappearance of the disease."[27] Another problem Pearl addresses is implicitly entailed knowledge which is not explicated in the logical expression. We might reasonably consider the following chain inference:

$$\begin{array}{r} AX \vee Y \longrightarrow P \\ AX \longleftrightarrow Z \\ \hline \therefore Z \vee Y \longrightarrow P \end{array}$$

where the conclusion is licensed through Leibniz' law and A is supposed to represent an inus condition for P. Now, the inferred expression in the last line does not show A anymore – is it also justified to analogously conclude that A is not really a cause of P anymore?

Although logically structured inus conditions seem to provide deeper insight into causal reasoning (than flat statements about necessity and sufficiency) and to make patterns of causal claims more transparent, obviously even more structure is necessary.

[25] Cf. [Mackie 1965, p. 255].

[26] This is the title of the introduction to [Pearl 2009], chapter 10, *The actual cause.*

[27] Cf. [Pearl 2009, p. 315].

1.5 Correlation and probabilistic causation

Nancy Cartwright opens her critical discussion of probabilistic accounts of causation in *What Is Wrong With Bayes Nets?* (2001) by stating that "[p]robability is a guide to life partly because it is a guide to causality."[28] Although she goes on to argue against a purely correlation-based concept of causality, various philosophers have approached causal reasoning from a probabilistic perspective in two respects: For some (like Suppes) the probabilistic analysis of causation means that causal relations can be *characterized* in terms of (or even reduced to) probabilistic relations,[29] for others (like Salmon) causality being probabilistic simply means that it is *not deterministic*.[30] A probabilistic approach towards causal analysis tries to overcome those difficulties a follower of Humean regularity faces – the central claim is that the influence of causes on their effects shows in the fact that the occurrence of the cause changes the probability of its effects. This does not exclude cases where the effect occurs despite the absence of the event initially ascertained as its cause, which might be due to initially unforeseen, now efficacious additional influences. Nor are cases excluded in which the potential cause does not trigger the predicted effect. Some counteracting influences with low probability might have changed the *normal course of events*. Thus, "smoking causes lung cancer" is typically rather understood as a statement about smokers to be more likely to suffer from lung cancer (than non-smokers) than about certain and unalterable regularities. As Pearl postulates, "[a]ny theory of causality that aims at accommodating such utterances must therefore be cast in a language that distinguishes various shades of likelihood – namely, the language of probabilities."[31]

Hans Reichenbach (in his later deliberations about causation in 1956) grounds his analysis of the direction of time on the analysis of directed causation by formulating his *Principle of Common Cause* in terms of probabilistic inequalities, namely expressions of conditional independency.[32] At the core of this characterization lies the twofold probabilistic claim that (i) a cause raises the probability of its direct effects and (ii) no other event renders the cause and a direct effect probabilis-

[28] Cf. [Cartwright 2001, sect. 1].

[29] Cf. e. g. [Hitchcock 2010, sect. 3.7].

[30] Cf. for this and the following [Williamson 2009].

[31] Cf. [Pearl 2009, p. 1].

[32] Cf. for a contemporary reformulation of the original notation [Williamson 2009, pp. 188 f.].

tically independent.[33] A few years later, in 1961, I. J. Good suggests an alternative to what Reichenbach had presented, because he objects one would always be able to conceive of an event that renders two other variables probabilistically independent – thus reducing Reichenbach's principle to a vacuous analysis because it does not yield any causes anymore. Explicitely incorporating the direction of time into his account, Good provides an expression to quantitatively measure potential and actual causation.[34] With E and F being distinct events ($\overline{E}$ and $\overline{F}$ their non-occurrences, respectively) and H consisting of all background conditions including prevailing laws of nature the *tendency of F to cause E* is expressed by

$$\log \frac{P(\overline{E}\,|\,\overline{F}H)}{P(\overline{E}\,|\,FH)}.$$

Also building on the direction of time, Patrick Suppes develops the definition of a genuine cause as a *prima facie* cause which is not *spurious* resting his explication on the following definitions:[35]

Definition 1.5.1 (Suppes' Prima Facie Cause)
The event $B_{t'}$ is a prima facie *cause of the event A_t if and only if*

(i) $t' < t$,

(ii) $P(B_{t'}) > 0$,

(iii) $P(A_t|B_{t'}) > P(A_t)$.

Definition 1.5.2 (Suppes' Spurious Cause)
A prima facie cause $B_{t'}$ is a spurious *cause of the event A_t if there is a prior partition $\pi_{t''}$ of events (with $t'' < t'$) that screens off $B_{t'}$ from A_t, i. e., for all elements $C_{t''}$ of $\pi_{t''}$*

(i) $P(B_{t'}C_{t''}) > 0$,

(ii) $P(A_{t'}|B_{t'}C_{t''}) = P(A_{t'}|C_{t''})$.

In other words, an event *genuinely* causes some subsequent second event if it raises the probability of this second event and if there is no prior third event that would render the first two independent if conditioned on. This independence test excludes earlier side effects from being analyzed

[33] Cf. [Williamson 2009, p. 189].
[34] Cf. [Williamson 2009, p. 191].
[35] Cf. [Williamson 2009, pp. 191 f.].

as true causes. The same idea underlies the analysis of causation put forward by econometrician and Nobel Prize winner Clive Granger in 1969, who argues for defining causes as events that are correlated with later effect events only when the entire past history of the putative cause up to its very occurrence is held fixed, i. e., when all variables prior to the cause candidate are conditioned on.[36]

All these considerations open into the development of the concept of Bayesian networks formulated by Judea Pearl in the 1980s as the basis for automated inference.[37] One of the protagonists in the field of probabilistic accounts of causation is Wolfgang Spohn who like Suppes also emphasizes the direction of time as a prerequisite crucial to his account. Where Suppes in his reductionist approach can be seen as representing causal pluralism, since he does not stipulate which interpretation of probability is to be preferred above others, Spohn makes the case for the subjective interpretation of probability as personal degrees of belief. Thus following the original intentions of Thomas Bayes he goes one step further and characterizes the relation of causation in its core by stating that "Bayesian nets are all there is to causal dependence"[38] – in other words, sufficiently rich Bayesian nets in causal interpretation together with the causal Markov condition yield just the dependencies and independencies we also expect in scientific or everyday causal reasoning and when interacting with our environment (see chapter 2 for a detailed presentation of Bayes nets, the Markov condition, and their causal interpretation).

All arguments for probabilistic accounts of causation face substantial points of criticism. Nancy Cartwright makes out quite a list of critical observations about (more or less refined) purely correlation-based causal analysis.[39] Trying to get from probabilistic dependence to causal dependence one should be wary:

> *What kinds of circumstances can be responsible for a probabilistic dependence between A and B? Lots of things. The fact that A causes B is among them: Causes produce their effects; they make them happen. So, in the right kind of population we can expect that there will be a higher frequency of the effect (E) when the cause (C)*

[36] Cf. [Cartwright 2001, sect. 3].

[37] Cf. e. g. [Pearl 1982].

[38] This quotation refers to the title of [Spohn 2000].

[39] Cf. for this and the following [Cartwright 2004] and the detailed discussion in [Cartwright 2001].

> *is present than when it is absent; and conversely for preventatives. With caveats.*[40]

Among the issues Cartwright addresses is the fact that correlation might be induced by common or correlated causes (or preventatives, respectively). Moreover, two causes might also – due to the fact that they jointly produce an effect – be correlated in populations where the effect is strongly present (or absent, respectively), maybe because populations are overstratified in the respective set-up of a study – i. e., in Bayes net terminology, two otherwise causally unrelated variables are dependent conditional on a common successor in collider structures. Thirdly, certain variables may show the same time trend without being causally related, at all. The prototypical example: The Venetian sea level rises with the same tendency as does the bread price in London, although neither actually causes the other nor would we try to attribute the correlation to some latent common cause. Fourthly, one remark about the assumption of stability (sometimes also 'faithfulness'): The information conveyed by Bayes nets is actually encoded in the absences of directed edges through which pairwise (conditional) independence between two variables is indicated. Especially when we try to build Bayes nets from raw data, the assumption of stability tells us that if data does not signal dependency between two variables, we have no reason to nevertheless insert an edge between the two corresponding nodes in our Bayes net. The underlying assumption is that it takes very precise values to cancel correlation where there actually are (physical etc.) mechanisms at work, and that such preciseness is rarely if ever found in imprecise disciplines or in oftentimes necessarily inexact measurements. Still, the theoretical possibility exists that, e. g., positive and negative effects of a single factor neutralize, thereby obscuring causal influence. Cartwright illustrates this point with Germund Hesslow's canonical birth-control pill example: "The pills are a positive cause of thrombosis. On the other hand, they prevent pregnancy, which is itself a cause of thrombosis. Given the right weights for the three processes, the net effect of the pills on the frequency of thrombosis can be zero."[41] In this example it might not only be the case that data does not show dependency where we would normally suspect causal mechanisms at work, but it might moreover be an important goal of medical research to achieve this very independence and to delete dependency from data – still acknowledging the physical, physiological etc. processes in nature.

[40] Cf. sect. 4, *From probabilistic dependence to causality*, in [Cartwright 2001].

[41] Cf. [Cartwright 2001, sect. 3].

Jon Williamson extends Cartwright's list in his critique of oversimplistic applications of the *Principle of Common Cause* and its implications. He points out once more that two positively or negatively correlated events do not have to be related causally but may instead be "related logically (e. g. where an assignment to A is logically complex and logically implies an assignment to B), mathematically (e. g. mean and variance variables for the same quantity are connected by a mathematical equation), or semantically (e. g. A and B are synonymous or overlap in meaning), or are related by non-causal physical laws or by domain constraints. In such cases there may be no common cause to accompany the dependence, or if there is, the common cause may fail fully to screen off A from B."[42] Nancy Cartwright sums up these critical points in pragmatic fashion – rejecting the philosophical effort to unify the representation of causal relations she says:

> *The advice from my course on methods in the social sciences is better: "If you see a probabilistic dependence and are inclined to infer a causal connection from it, think hard. Consider the other possible reasons that that dependence might occur and eliminate them one by one. And when you are all done, remember – your conclusion is no more certain than your confidence that you have eliminated all the possible alternatives."*[43]

Supposed opponent Judea Pearl agrees with Cartwright on the fact that the shortcomings of a purely probabilistic reductionist approach towards causal analysis prohibit at least *direct* application. He marks the distinction between mere observation (or *acts*) represented in data and knowledge about the impact of (hypothetical) intervention (or *action*) which is not part of statistical models. Conditioning on certain variables just switches the subpopulation and does not yield information about the causal machinery at work.[44] Obviously, a modification of the method is necessary to also encode counterfactual knowledge.

1.6 Counterfactual analysis

One of the first references to the idea of characterizing causation in terms of counterfactual conditionals dates back to as early as 1748, when David Hume compactly analyzed a cause to be "an object followed by another,

[42] Cf. [Williamson 2009, p. 200].

[43] Cf. [Cartwright 2001, sect. 5].

[44] Cf. e. g. the section *Actions, Acts, and Probabilities* in [Pearl 2009, pp. 108 ff.].

[...] where, if the first object had not been, the second never had existed."[45] Suzy throws a stone and shatters a window with it. Had she not thrown the stone, the window would not have broken to pieces. Obviously, this counterfactual analysis seems to capture much of our intuition about causation. It ties the observed course of events – when considering causal relations at token level – to the mechanisms that govern our world underneath the surface and are of more use to us than mere listings of successive happenings because they contain hints at how to manipulate the respective setting to achieve different outcomes. L. A. Paul notices that "[i]n everyday life as well as in the empirical and social sciences, causes are identified by the determination of manipulation: Cs are causes of Es if changing Cs changes the Es, that is, if we can manipulate Es by manipulating Cs. In this way, experimental settings are designed to test for the presence of causation by testing for the presence of counterfactual dependence."[46] In his seminal article *Causation* (1973) David Lewis offers a detailed presentation of causal analysis on the basis of counterfactual dependence together with a full-blown semantics for evaluation.[47] For him, counterfactual dependence between two successive and suitably distinct events is sufficient for causation. But his possible worlds semantics of counterfactuals does not yield transitivity of counterfactual statements in contrast to our intuition that causation should be characterized as transitive. Thus, causation cannot be simply reduced to counterfactual dependence. In the following, three notorious prima facie problematic cases shall be considered and possible fixes thereof sketched in brief – namely the cases of *side effects*, *pre-empted* potential causes, and *overdetermined* events.[48]

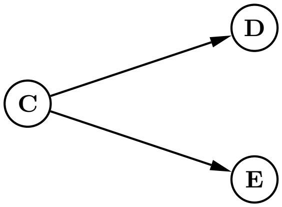

Fig. 1.1: D counts as side effect of E in this common cause fork.

[45] This is actually the second part of his famous twofold explication – see below, chapter 2, and cf. [Hume 1748, Section VII].

[46] Cf. [Paul 2009, p. 166].

[47] Cf. [Lewis 1973b].

[48] Cf. for this and the following the extensive discussion in [Paul 2009, sects. 2–3].

1. **Side effects of common causes.** The fork structure in fig. 1.1 represents the case where C simultaneously causes D and E. Now, one would assume that D always occurs when E does. The reverse does not hold, whatsoever. The counterfactual statement 'if D had not occurred, E would not have happened, either' does not hold if *backtracking counterfactuals* are forbidden, as is the case with Lewis' possible worlds semantics. Consequently, if D does not occur one is not licensed to infer that C has not taken place either, since C might have happened but at the same time failed to cause E due to extraneous preventatives. So, basing the counterfactual analysis of causation on non-backtracking conditionals yields the expected results for common cause fork structures.

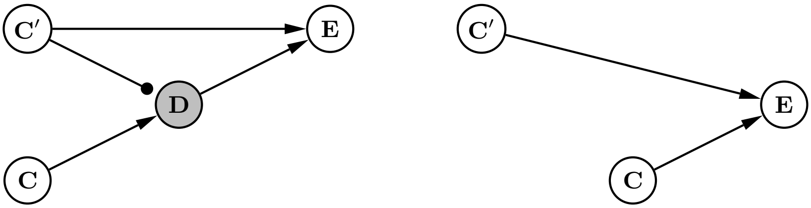

Fig. 1.2: C' pre-empts the potential cause C in the case of early pre-emption (left) and late pre-emption (right).

2. **Early and late pre-emption.** David Lewis also offers a solution for the problem of pre-empted potential causes. The left "neuron diagram" in fig. 1.2 depicts a situation where C potentially causes E. But the causal chain from C to E is disrupted by the influence of C', the occurrence of which prevents C from being causally relevant to E through the deactivation of intermediate event D (indicated by the round arrowhead pointing from C' to D). Simple counterfactual analysis yields that C' is no cause of E since if C' had not occurred, C would have triggered E – E does not counterfactually depend on C' anymore. To deal with this problem, Lewis extends causal dependence to a transitive relation (as the ancestral of causal dependence) by defining the concept of a causal chain.[49] In *Causation* he sums up his counterfactual analysis of the subject:

> *Let c, d, e, . . . be a finite sequence of actual particular events such that d depends causally on c, e on d, and so on throughout. Then this sequence is a causal chain. Finally, one event*

[49] Cf. e. g. [Menzies 2009a, sect. 2.3].

> *is a cause of another iff there exists a causal chain leading from the first to the second.*[50]

The so-called case of *late pre-emption* faces a different kind of problem. If the right diagram of fig. 1.2 is interpreted as a series of events succeeding one another in time from left to right then event C' prevents C from becoming causally effective by causing event E earlier. To give an illustrative example: Suzy and Billy throw stones to shatter a glass bottle. Suzy's stone hits the bottle earlier than Billy's and thus can be counted as the cause of the breaking of the bottle, whereas Billy's stone must fail to break the bottle because it is already broken to pieces. In this case (potential) causal efficacy of either event C or C' cannot be accounted for in terms of counterfactual dependence or causal chains. One way to side-step this problem is to introduce a fine-grained concept of events and to make events *fragile* with respect to time.[51] This makes counterfactual dependence applicable again: Suzy's throw caused the bottle to break at time t_1 – call this event E_{t_1}. Had she not thrown the stone, the bottle would have broken later, at time t_2, as event E_{t_2} brought about by Billy's throw. E_{t_1} would not have occurred either in this case, thus validating the counterfactual $\neg C' \;\Box\!\!\rightarrow\; \neg E_{t_1}$.[52] The question remains, if this kind of fine-graining still captures everyday causal talk, not to mention type-case utterances.

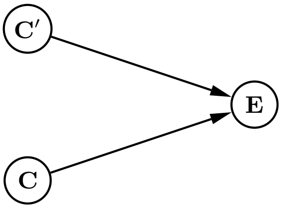

Fig. 1.3: C and C' overdetermine the event E in this collider structure.

3. **Cases of overdetermination.** A slightly modified variant of the bottle shattering example can be used to illustrate the case in which two events jointly overdetermine a later third event, as sketched in the neuron diagram of fig. 1.3. Assume that Suzy (C')

[50] Cf. [Lewis 1973a, p. 563].

[51] Cf. [Menzies 2009a, sect. 4].

[52] The counterfactual formula in words: 'If C' did not occur, E_{t_1} would not occur, either' (or the respective past tense form); see chapter 2 for an explication of the truth conditions of counterfactuals.

> and Billy (C) throw their stones and hit the bottle at the exact same time, causing it to break (E). Again, this situation cannot be analyzed in accordance with our intuition if one relies on plain counterfactual dependence. If either C or C' had not occurred, the respective remaining event would have caused the bottle to break. Resorting to temporally fragile interpretations of the situation does not remedy things either, because – as the example goes – both stones simultaneously hit the bottle. Of course it might be argued that there are no genuine simultaneous events and that a sufficient fine-graining of the physical description of the situation will always ultimately yield a solution when drawing on the temporal fragility of events or – going one step further – on an extension of fragility to properties in general. So the interesting questions arise when we allow for *fine-grained overdetermination*[53] and ask ourselves what the actual cause of a truly overdetermined event (also in the fine-grained sense) really is. This sounds like a tough question in such an abstract formulation, and, e. g., L. A. Paul finds it noteworthy "how differently we feel about the clarity of cases of fine-grained overdetermination versus that of cases of early and late pre-emption. [...] it just isn't clear how each cause is bringing about the effect all on its own, given that another cause is also bringing about the effect all on its own and the causation is not joint causation."[54]

One further general problem counterfactual theories of causation face is the charge of circularity. If the definition of causal dependence rests on counterfactual dependence, the semantics of counterfactuals must avoid relying on causal relations. If this is not possible, more has to be said about the grounding of higher level on lower level causal claims or basic causal assumptions (as do Woodward[55] and Pearl[56]).

Although the counterfactual analysis seems to truly capture essential features of our understanding of causal relations, refined approaches are needed, obviously. E. g., Judea Pearl proposes directed *structural equations* and claims that these are actually expressions of counterfactual

[53] For a discussion of variants of fine-graining and overdetermination cf. [Paul 2009, pp. 178 ff.].

[54] Cf. [Paul 2009, p. 180]; *joint* causation means that both causes are needed to bring about the effect in the precise way it actually occurred.

[55] Cf. [Paul 2009, p. 172].

[56] Cf. [Halpern & Pearl 2005a, p. 849].

knowledge.[57] On a higher level he even defines how to interpret the probability that an event $X = x$ "was the cause" of an event $Y = y$ in terms of counterfactuals: $P(Y_{x'} = y'|X = x, Y = y)$ can be understood in his framework as the probability of Y not being equal to y had X not been x, given that $X = x$ and $Y = y$ *are* (observed) facts in the respective situation. Relative to his definition of probabilistic causal models, Pearl lists the three steps necessary for counterfactual evaluation (in corresponding twin networks): abduction, action (i. e., intervention), and prediction.[58]

Another contribution to the ongoing discussion – especially for the solution of cases of overdetermination – has been brought forward by Christopher Hitchcock, who enhances structural representations of test cases by introducing *default* and *deviant* values, thus emphasizing our intuition that an event is more likely attributed causal efficacy if it *deviates* from the *normal course of events* in a sufficiently significant way.[59] The semantics of normality nevertheless remains a point of controversy, as does the semantics of counterfactuals in the possible worlds presentation due to troublesome transfer into application, as, e. g., Judea Pearl points out insistently.[60]

1.7 Ranking theory

Another critic of the counterfactual account of causation (especially as presented by Lewis) calls its self-imposed claim of objectivity in question. Wolfgang Spohn derives in his own approach causes from reasons as *subjective degrees of belief*, thereby relativizing causation to an observer or epistemic individual. He criticizes Lewis:

> *[T]he stance of the counterfactual theory towards the objectivity issue is wanting, I find. The official doctrine is, of course, that the counterfactual theory offers an objective account of causation; this definitely counts in its favor. However, this objectivity gets absorbed in the notion of similarity on which* Lewis*' semantics for counterfactuals is based. [. . .] I wonder whether such similarity judgments are significantly better off in the end than, say, judgments about beauty, and hence whether the semantics for counterfactuals should not rather take an expressivist form like the*

[57] Cf. [Pearl 2009] and a presentation of Pearl's framework in chapter 2.

[58] Cf. [Pearl 2009, p. 206], Theorem 7.1.7.

[59] Cf. e. g. [Hitchcock 2009a] and [Hitchcock 2007].

[60] Cf. Pearl's reply to Lewis' article *Causation* in [Pearl 2009, pp. 238 ff.].

> *semantics of "beautiful". The question is difficult to decide, and I do not want to decide it here. The only point I want to make is that the whole issue is clouded behind the objectivistic veil of the counterfactual theory. It is clearer, I find, to jump right into subjectivity [. . .]*[61]

Consequently, SPOHN bases his account of causal reasoning on weightings of personal *reasons* in the form of *ranking functions*. Other than purely qualitative representations of epistemic entities (or changes in such entities, respectively) in the area of knowledge representation or reasoning with uncertainty, ranking functions quantitatively represent the epistemic state of a subject in terms of degrees of belief, but at the same time induce a notion of yes-or-no belief complying with the constraints of rational reasoning (i. e., consistency and deductive closure).[62]

Only a very brief sketch of ranking theory shall be given here to support the following points. A belief function β measures the strength of an agents subjective belief in proposition A, depending on whether $\beta(A) > 0$ (belief in the truth of A), $\beta(A) < 0$ (belief in the falsity of A), or $\beta(A) = 0$ (indifference as to whether A is true or false), with $\beta(A) \in \mathbb{Z} \cup \{-\infty, +\infty\}$.[63] Now, in this framework A is defined as a *reason for* B, iff $\beta(B|A) > \beta(B|\overline{A})$, i. e., (the occurrence or the perception of) A strengthens the belief in B.[64] Moreover, reasons are systematically classified as *additional*, *sufficient*, *necessary*, or *weak* reasons, depending on whether $\beta(B|\overline{A})$ yields a value below or equal to 0 (in the cases of A being a sufficient, necessary, or weak reason for B), whether $\beta(B|A)$ yields a value above or equal to 0 (in the case of A being an additional, a sufficient, or necessary reason for B), and how these conditions are combined.[65] In a next step, causes are quite simply derived from the definition of reasons: A is an (additional, sufficient, necessary, or weak) direct (*token* or *singular*) cause of B iff (i) A is an (additional, sufficient, necessary, or weak) reason for B, (ii) A and B actually obtain, and (iii) A temporally strictly precedes B.[66] Quite like LEWIS, Wolfgang SPOHN goes on by defining a cause to be a proposition A that is connected to another proposition B by a *chain* of direct causes. This is – in a nutshell

[61] Cf. [Spohn 2001, sect. 8].

[62] Cf. [Huber 2009, p. 1351].

[63] Cf. for this and the following e. g. [Spohn 2001].

[64] This is of course formulated relative to (i. e., conditional on) an agent's given doxastic state C, which is omitted here for purposes of compactness.

[65] Cf. [Spohn 2001, sect. 4], definition 5b.

[66] Cf. [Spohn 2001, sect. 5], definition 6.

– all that is needed to follow Spohn's comparison of how cases of overdetermination can be treated in Lewis' purely counterfactual account and in his own ranking-theoretic analysis.

If numbers are distributed in accordance with our understanding of the situation depicted in fig. 1.3, we could possibly come up with the following tables showing concrete values for $\beta(C|X \cap Y)$:[67]

$\beta(C\|\cdot)$	B	$\overline{B}$
A	1	-1
$\overline{A}$	-1	-1

(a) A and B are joint sufficient and joint necessary causes

$\beta(C\|\cdot)$	B	$\overline{B}$
A	1	0
$\overline{A}$	0	-1

(b) A and B are joint sufficient but not necessary causes

$\beta(C\|\cdot)$	B	$\overline{B}$
A	2	1
$\overline{A}$	1	-1

(c) A and B are overdetermining causes of C

Each table specifies the degree of belief $\beta(C|X \cap Y)$ in C conditional on $X \cap Y$, where $X \in \{A, \overline{A}\}$ and $Y \in \{B, \overline{B}\}$. Numerical values possible change from one epistemic subject to another (under preservation of necessity, sufficiency, or overdetermination if constrained correctly).

Case (a) represents the standard understanding of joint causes – both A and B have to occur in order to bring about C, neither A nor B alone suffice for that, given through $\beta(C|A \cap B) > 0 > \beta(C|A \cap \overline{B}) = \beta(C|\overline{A} \cap B) = \beta(C|\overline{A} \cap \overline{B})$. Only the joint occurrence of A and B raises the belief in C from a negative number (disbelief) to a positive (belief).
Spohn also considers case (b) an example of joint causation, obviously not as definite as case (a) since, e. g., in the presence of A the occurrence of B is (per definitionem) a sufficient contribution to C, but a necessary one in the absence of A: $\beta(C|A \cap B) > 0 = \beta(C|A \cap \overline{B}) > \beta(C|\overline{A} \cap \overline{B})$. The occurrence of either A or B raises the belief in C from disbelief (< 0) to indifference ($= 0$), but only the joint occurrence of A and B lifts the degree of belief in C to a positive value.
Scheme (c) finally exhibits the case of overdetermining causes. Each of A and B already suffices to produce C by raising the belief in C from a negative to a positive degree and can be understood as an additional contribution to C in the presence of the other, raising the degree of

[67] This example is taken from Spohn's illustration of the problem of overdetermination in [Spohn 2001, sect. 7].

belief in C even further, as specified in the ranking function β and given through $\beta(C|A \cap B) > \beta(C|A \cap \overline{B}) = \beta(C|\overline{A} \cap B) > 0$. The high degree of belief in C can be interpreted as strong doubt in the fact that C would obtain if neither of A or B actually occurred.

Wolfgang Spohn summarizes why ranking theory copes with the fine-grained representation of causal intuitions much better than the counterfactual approach:

> *[R]anking functions specify varying degrees of disbelief and thus also of positive belief, whereas it does not make sense at all, in counterfactual theories or elsewhere, to speak of varying degrees of positive truth; nothing can be truer than true. Hence, nothing corresponding to scheme (c) is available to counterfactual theories.*[68]

Nevertheless, justified questions appear on the scene as soon as one tries to tie ranking theory to application, e. g., in implementations of belief revision or automated reasoning. How does an epistemic agent obtain those specific numerical values as initial degrees of belief? And if ranking theory starts constructing *singular causes* from *subjective reasons* – how, if at all, could any notion of objectivity be established?[69] And the computer science engineer might add: Isn't there any more compact way of representing and implementing degrees of belief and changes of epistemic states than as plain listings of each and every ratio?

1.8 Agency, manipulation, intervention

Manipulationist theories of causation build upon our very basic intuition that an effect can be brought about by an apt manipulation of the putative cause. In other words and cum grano salis, if an event C causes some distinct event E, then a modification of C in some way will change the outcome E correspondingly. And conversely, if by (if only hypothetical) manipulation of an event C some subsequent event E were to present itself differently (relative to the expected normal and unmanipulated course of events), C causes E (even if in this counterfactual formulation the manipulation were not actually to be performed). This idea has received considerable attention in the recent literature on causal inference, even in non-philosophical publications of medical research, econometrics,

[68] Cf. [Spohn 2001, sect. 7].

[69] Of course, Spohn does say more about the task of objectivizing ranking functions e. g. in [Spohn 2001] and [Spohn 2009].

sociology, even psychology and molecular-biology etc. because it maps the quest for causes onto the practice of experimentation. Manipulationist theories in this way go beyond the determination of mere regularities in observed processes or the plain investigation of correlation, and introduce the (virtual) capability of interaction into the test setting. This is done differently in different flavors of manipulationist theories.[70] *Agency* theories in anthropomorphic fashion emphasize an agent's freedom of action involved in performing a manipulation of the respective situation. The gap between (human) agency and causation is then bridged by the notion of *agent probability.* The causal efficacy of an event C on E is linked to C's raising the agent probability of E, when this agent probability is interpreted as the probability that E would obtain if an agent were to *choose to realize* C. Since in this formulation causation is broken down into atomic building blocks of *free acts*, agency theories do avoid circularity – they face a different problem, though, namely their very limited scope of application. E. g., how is causal efficacy attributed to friction of continental plates resulting in an earthquake? Surely we utter causal claims about such geo-physical happenings with the same confidence as we talk about someone's throwing a stone as being the cause of some bottle's shattering.[71]

Judea Pearl[72] or James Woodward[73] draw on a different and more general kind of manipulation. In their congeneric accounts of causation the capability of interaction is given through *hypothetical interventions* on variables in fixed causal models. Those variables are basically suitably distinct events of interest that the designer of a causal model deems to be worth considering and contributory to the understanding of the respective situation. Causal models moreover – in essence – merely list for each variable in the model its immediate predecessors, i. e., causally interpreted, its direct causes (thus obeying the so-called *causal Markov condition* if the resulting structure does not contain cycles of parenthood, see below, definition 2.6.1). The underlying idea of both Pearl's and Woodward's approach is *modularity* as a requirement for causal models to be reliable sources of information and thus useful for explanation.

[70] Cf. for this and the following [Woodward 2009].

[71] See Woodward's discussion of the earthquake example due to Menzies and Price together with the potential but controversial solution via *projection* in [Woodward 2009, pp. 238 ff.].

[72] Cf. [Pearl 1995] and for a more elaborate presentation [Pearl 2009] (the second edition of his 2000 book).

[73] Cf. [Woodward 2003].

Modularity accounts rest on the postulate that each link between two variables represents a *mechanism* for the effect, which can vary modularly and independently of mechanisms for any other variables in the causal model.[74] If those mechanisms are represented as individual equations, the researcher can mathematically utilize them to learn about the effects of interventions – as Pearl puts it:

> *In summary, intervention amounts to a surgery on equations [. . .] and causation means predicting the consequences of such a surgery.*[75]

Nevertheless, how such interventions are precisely implemented in Pearl's and Woodward's account slightly varies in detail. Pearl compactly defines atomic interventions as external deactivations of some variable's links to its causal parents (i. e., its direct causes) or analogously as the deletion of the respective functional connection in the corresponding structural model:

> *The simplest type of external intervention is one in which a single variable, say [X], is forced to take on some fixed value [x]. Such an intervention, which we call "atomic", amounts to lifting [X] from the influence of the old functional mechanism [linking the value assignment of X to the values of its parents] and placing it under the influence of a new mechanism that sets the value [x] while keeping all other mechanisms unperturbed.*[76]

As Woodward notes critically, this explication induces a definition of cause that relies on certain mechanisms to remain unperturbed. If these mechanisms are of causal character themselves, then Pearl has to defend his definition against the charge of circularity. He indeed does suggest a possible reading of interventions that avoids circularity in [Halpern & Pearl 2005a]. (For a detailed presentation of Pearl's account of causation see chapter 2.) Woodward follows a slightly different route by introducing specific *intervention variables* into his framework and by constraining those variables in a suitable way. An intervention I on some variable X is then defined relative to the putative effect Y in order to characterize what it means for X to cause Y:[77]

[74] See [Cartwright 2004, pp. 807 ff.] for a critical discussion of the modularity requirement.

[75] Cf. [Pearl 2009, p. 417] – highlighting modified.

[76] Cf. [Pearl 2009, p. 70].

[77] See Woodward's presentation of the requirements of such Woodward-Hitchcock interventions in [Woodward 2009, p. 247].

1. I must be the only cause of X – i. e., the intervention must completely disrupt the causal relationship between X and its preceding causes so that the value of X is entirely controlled by I (in other words, the set of parents of X only contains I);

2. I must not directly causally influence Y via a route that does not go through X;

3. I should not itself be caused by any cause that influences Y via a route that does not go through X;

4. I must be probabilistically independent of any cause of Y that does not lie on the causal route connecting X to Y.

In contrast to Pearl's explication, it is not excluded that such an intervention variable I is causally related to or probabilistically dependent on other variables in the causal model, but it is specified exactly, which variables I is required to be (causally and probabilistically) independent of. And in contrast to agency theories, as Woodward emphasizes, a "purely natural process, not involving human activity at any point, will count as an intervention as long as it has the right causal and correlational characteristics."[78]

The pivotal idea of modular manipulationist accounts is the exploitation of causal diagrams for reliable causal inference. The Bayes net methodology provides the desired framework and readily extends causal yes-or-no reasoning to an analysis of causal influence in terms of degrees of belief. Mapping this approach onto the standard proceeding of a randomized controlled clinical trial exemplifies the general applicability of interventionist theories.

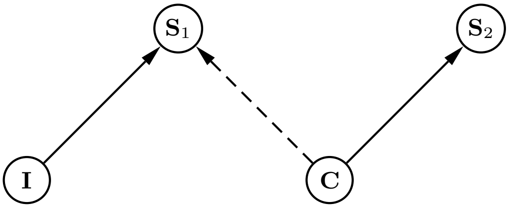

Fig. 1.4: Symptom S_1 is lifted from the causal influence of cause C by means of intervention I in a randomized controlled trial.

[78] Cf. [Woodward 2009, p. 247].

Consider the situation depicted in figure 1.4 where some cause C (perhaps some disease or some behavior detrimental to health) results in the simultaneous occurrence of symptoms S_1 and S_2.[79] Arrows mark direct causal influence. In order to discover the causal relationship between C, S_1, and S_2, the test candidates (exhibiting characteristic c or $\neg c$ in the dichotomous case) are divided randomly into test groups (subpopulations) where in one group symptom S_1 is induced somehow and in the other group prevented – according to the decision taken by setting I. Now, if inducing or preventing symptom S_1 were to bring about a significant change in the measurement of symptom S_2, we would be entitled to postulate some causal connection between both variables due to the above explication of intervention. Randomization of test groups nevertheless precisely amounts to lifting the variable S_1 from the influence of variable C and cutting the connection (thereby cancelling the correlation) between S_1 and C (as indicated by the dashed arrow pointing from C to S_1 in figure 1.4) and consequently between S_1 and S_2. S_1 is therefore analyzed as *not* directly causally influencing S_2.

The example makes obvious how tightly the principle of modularity is connected with causal reasoning – without the assumption of modular (and modularly separable) causal links the whole enterprise of randomization in our controlled clinical trial would have failed. Another investigation WOODWARD undertakes in the explication of his interventionist account is centered about the question, what the nature of mechanisms is in essence. He reminds the reader of his *Making Things Happen* (2003) of "the absence of any consensus about the criteria that distinguish laws from nonlaws and the difficulties this pose[s] for nomothetic accounts of explanation"[80] and discusses *invariance* as an intrinsic feature of the generalizations required for causal inference, thereby strengthening the modularity requirement:

> *The guiding idea is that invariance is the key feature a relationship must possess if it is to count as causal or explanatory. Intuitively, an invariant relationship remains stable or unchanged as various other changes occur. Invariance, as I understand it, does not require exact or literal truth; I count a generalization as invariant or stable across certain changes if it holds up to some appropriate level of approximation across those changes. By contrast, a generalization will "break down" or fail to be invariant across certain changes if it fails to hold, even approximately, under those changes.*[81]

[79] This illustration follows the motivational example in [Woodward 2009, sects. 2/5].

[80] Cf. [Woodward 2003, p. 239].

[81] Cf. [Woodward 2003, p. 239].

Woodward goes on by bridging the gap between his theoretical claims and (the practice of) explanation in the special sciences:

> *In contrast to the standard notion of lawfulness, invariance is well-suited to capturing the distinctive characteristics of explanatory generalizations in the special sciences. [...]*
>
> *[A] generalization can be stable under a much narrower range of changes and interventions than paradigmatic laws and yet still count as invariant in a way that enables it to figure in explanations.*[82]

In her critical discussion, Nancy Cartwright acknowledges the advantages of invariance methods but also points out that these methods require a great deal of antecedent causal assumptions or knowledge about the causal influences at work, because what it means for some generalization to be invariant or stable across certain changes *of the right sort* must carefully be explicated when applying the method to individual causal hypotheses. Due to these demanding requirements, Cartwright argues, invariance methods "are frequently of little use to us."[83] Another reason for her to object to invariance methods is the fact that the situation under consideration must be of modular nature, which is – in Cartwright's eyes – only the case for a limited set of situations and does not carry over to general application.[84] Pearl refutes this argument sharply in [Pearl 2010] by claiming that formal structural systems (e. g., as used in econometrics) are usually established on the basis of the modularity assumption.[85] Woodward (necessarily) bases his interventionist notion of causation precisely on the modularity principle and defines what it means for an event to be a direct cause of some other subsequent event through combination of *multiple* interventions as follows:

Definition 1.8.1 (Woodward's Direct Cause)[86]
A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set V is that there be a possible intervention I on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by additional interventions that are independent of I.

[82] Cf. [Woodward 2003, p. 240].

[83] See Cartwright's critical discussion in [Cartwright 2004, pp. 811 f.].

[84] Cf. [Cartwright 2004, pp. 807 ff.].

[85] Cf. [Pearl 2010, pp. 73 ff.].

[86] See definition (DC) in [Woodward 2009, p. 250].

Using this definition as a starting point for further considerations, conditions for X to be a *contributing* cause of Y are consequently built on the notion of chains of direct causes. How Judea Pearl uses the notion of intervention in his framework and how he ultimately arrives at the concept of *actual cause* is presented in more detail in chapter 2.

1.9 Decisions to take

The brief presentation of manipulationist strategies concludes the systematic overview of the most prominent approaches towards the analysis of causal claims. Obviously, the different approaches more or less differ in their attempt to provide answers to a set of questions that are central to the analysis of causation. The most important decisions one has to take when setting out to trace the notion of causality in some way shall be collected in the following catalogue.[87]

1. What are the relata of causal relations? Can objects or regions of time-space be the cause of entities of the same kind? Or does any causal talk of objects always have to be understood in terms of the more fundamental concept of events, i. e., either as instantiations of properties in objects at a given time (in the Kimian sense) or as plain random variables (in a probability-theoretic sense) that signify events by assuming one of at least two different values?

2. Does the causal analysis in the framework to be developed relate single case entities (at token level) or does it ascribe meaning to claims about generic causation (at type level)? If both cases are treated in the framework – which one is prior to the other? Are cases of singular causation as "the rain yesterday at noon caused my driveway to get wet" to be understood as more fundamental than cases of generic causation as "rain causes a street to get wet" – or vice versa? And can one be derived from the other, e. g., by some kind of induction principle?

3. Do causal claims relate entities at population level or at individual level? E. g., a certain approach might be able to deal well with medical findings connecting some epidemic in a given group with the presence of some virus but might fail to account for the

[87] This compilation expands the listings in [Williamson 2009, pp. 186 f.] and [Paul 2009, pp. 160–165].

outbreak of a disease in a specific observed individual. The explanatory power of a causal theory must be balanced well if it is to deal with up-scaling or down-scaling of test settings.

4. Is the theory capable of explicating *actual* causation (maybe post factum) or *potential* (possible, but maybe, e. g., pre-empted) causation? This question is obviously closely connected to the question, if the theory can handle counterfactual intuitions and yield answers to queries of the what-would-have-happened-if-things-had-been-different kind (possibly by ascribing causal efficacy to factually void, disrupted, or prevented events).

5. What is causation grounded in *ontologically*? Is the theory to be developed talking about some objective, physical notion of causation or about subjectively perceived or reconstructed mental causation within an (idealized) epistemic agent? And if both alternatives are not exclusive – how can knowledge about one side be carried over to the other side? If the subjective notion is prior to the objective one – how, if at all, is objectivization possible? Moreover, *causal explanation* clearly seems to be mind- and description-dependent. Now, if causation itself is of epistemic nature – what then is the difference between causal explanation and causation, if there is one to be made out?

6. Does the suggested account explore causation conceptually or ontologically? In other words, does it present an analysis of how we gain epistemic access to causal relations and how we internally structure our experience of causal processes, or is the account attempting to give an insight into what really is at the core of causation in the world? And if a conceptual approach does not deny the metaphysical existence of causal goings-on – how can the interconnection between both be described?

7. Is the approach taking a descriptive or a prescriptive route, i. e., is the account offering an illuminative picture of our concept of causation (via description) or is it prescriptively focusing on the construction of an improved formulation – maybe in a technically strongly constrained framework or for a certain branch of the special sciences?

8. Is causation holistically treated as one monolithic concept such that a technical description of causal relations can be applied to any kind of causal claims, independent of research area or jargon?

Or is ‘causation’ rather understood as a sort of cover term for the description of an irreducible multifarious, yet family-like conception?

9. If causation is understood as a family of different concepts that are best not reduced to allow for a better understanding of each of the single concepts – then which of the concepts is analyzed by the theory (if an analysis is attempted)? Is it the folk concept, the scientific concept, or possibly even the concept of a special branch of science? Does the theory approach causation from the philosophical or epistemological perspective? And moreover, how (if at all) are thought experiments, personal intuitions (as about cases of causation by omission), causal talk as linguistic expression, or special physical theories addressed?

10. Are causal relations themselves reduced to other non-causal more fundamental entities or in a non-reductive approach seen as the basic building blocks of causal claims? What could be candidates for more fundamental non-causal entities – powers, processes, mechanisms? And to pose a question closely connected to the problem of reducibility: How are the laws of physics to be understood and how (if at all) are they allocated in the framework?

11. Another important question marks the distinction between, e. g., Pearl’s account of causation and Spohn’s ranking-theoretic approach: What role does time play for the identifiability of causes and effects? Is time induced by the formal representation of cause-effect relations? Is it at best compatible with the arrangement of causally connected events? Or is it even seen as a necessary prerequirement that might ultimately go into the definiens of causal relations? And one more notorious question a theory of causation should be addressing: Is *backward causation* something worth considering, is it explicitly excluded from treatment or even denied by the framework?

12. Does an account of causation refer to deterministic causal relations or to probabilistic causation? And if it does talk about probabilistic causal relations – in what sense are these causal relations probabilistic, in a genuinely ontologically aleatoric sense or (in an epistemic sense) simply as a feature of our shortcoming to model supposed deterministic processes in a deterministic way? And does the framework allow for going back and forth between the deterministic

and the probabilistic rendition if both are addressed (maybe with one prior to the other)?

13. The final question is maybe one which is to be left to the evaluation of the theory as it proves itself in practice (or does not). It will be justified to ask, how applicable the theory finally turns out to be. To what degree does the proposed definition of cause prove operationally effective? How well can the suggested account be coupled with existing frameworks – especially when trying to embed causal concepts in the special sciences? The answer to these questions will obviously be related to the choice of the formal framework with its notation and the mathematical tools therein.

In chapter 3 below these questions will be reconsidered and related to an extension of the interventionist treatment of causation. It will be argued that causation be understood as an epistemic concept in order to illuminate certain disputed examples and controversial intuitions. Reasoning about causation will lead to the conclusion that a formal understanding of causal relations tells us more about the texture of reasoning itself. Our knowledge is – as will be argued – efficiently structured by the guiding principle of *causality.* It is such structured knowledge, mapped onto patterns of unified causal and non-causal information, which ultimately permits cognition *causarum rerum.*

Chapter 2

Causation and causality: From Lewis to Pearl

> Truth, or the connection between cause and effect, alone interests us. We are persuaded that a thread runs through all things; all worlds are strung on it, as beads
>
> Ralph Waldo Emerson,
> *Montaigne; or, the Skeptic*

2.1 What is a theory of causation about?

Within the last forty years the literature about theories of causation has increased immensely: Language analysts built new alliances with computer scientists and computational linguists. From this very corner probability theory was fueled, which bestowed upon philosophers the possibility of thinking about probabilistic causality. Pearl himself is a computer scientist and as such eager to offer effective tools aiding in finding concrete solutions to concretely posed questions. He thus turns on the purely metaphysical non-treatment of the concept of causation and devises a causal-theoretic toolbox for economists, physicians, sociologists – in short: for all those on the hunt for causes. At the same time he analyzes the prevalent situation with the following words:

> *Ironically, we are witnessing one of the most bizarre circles in the history of science: causality in search of a language and, simultaneously, speakers of that language in search of its meaning.*[1]

A theory of causation – however furnished – ought to be instrumental to the user and yield answers to queries like these, at least *in agreement* with personal intuition:[2]

- Is X a cause of Y?
- Is X a direct (respectively, an indirect) cause of Y?
- Does the event $X = x$ always cause the event $Y = y$?
- Is it possible that the event $X = x$ causes $Y = y$?

LEWIS and PEARL share common grounds in acknowledging that in everyday language stating causes is our base of explanation and justification, and that prediction of future events intrinsically and inextricably rests on causal assumptions. The analytical approaches deviate from one another, nonetheless.

2.2 HUME's counterfactual dictum

David LEWIS' paper *Causation* in the *Journal of Philosophy* 1973 opens with HUME's famous twin definition from 1748:

> *We may define a cause to be an object followed by another, and where all objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed.*[3]

The first part of this quote from David HUME's *An Enquiry About Human Understanding*, Section VII, sums up, what the regularity analysis of causation rests on. The mere uniform succession of events shall license the observer to identify an event occurring (or is it only *being observed*?) before a second event as a genuine cause of this very second event. Here and in what follows I will only talk about ordinary events

[1] Cf. [Pearl 2009, p. 135] – a slight variation of his original formulation in [Pearl 2000a, p. 135].

[2] Cf. [Pearl 2009, p. 222].

[3] Cf. [Hume 1748, Section VII].

to follow Lewis' own self-restriction: Lightning in a thunderstorm, battles between nations, chats amongst friends, etc. David Lewis formulates various critical notes against this regularity analysis of causation, thereby criticizing the advocates of probabilistic causality, who base their theory on the very correlation between events or states. In particular, any scenario exhibiting the regular succession of an event c (for *cause*) and an event e (for *effect*) can be analyzed reversely, so that – following Hume's words – e counts as a genuine cause of c.[4] This might possibly be against the arrow of time if time itself is not explained "into" the theory. Simultaneousness remains a tough case. Epiphenomena as echo of the causal history of an event c cannot be distinguished from epiphenomena of genuine effects e. And inefficacious pre-empted potential causes that might well have had causal influence, if not being pre-empted, are not even touched by the regularity analysis.

As an alternative, Lewis turns to Hume's "other words": "If the cause c had not been, the effect e would not have occurred, either." This "had not – would not" analysis constitutes one piece of the jigsaw of David Lewis' grand agenda of analyzing counterfactual statements, which culminates in his opus *Counterfactuals*, published in 1973, in the same year as *Causation.* Lewis presses his point there:[5]

> *True, we do know that causation has something or other to do with counterfactuals. We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well.*

Years later Judea Pearl will argue in the same direction.

As a sole spoiler, the meaning of counterfactual statements seems to evade intuition at first sight. Lewis counters this the following way:

> *Why not take counterfactuals at face value: as statements about possible alternatives to the actual situation, somewhat vaguely specified, in which the actual laws may or may not remain intact?*

Lewis does take counterfactuals at face value and designs an apparatus for the evaluation of such statements in his book *Counterfactuals.*

[4] Cf. [Lewis 1973a, p. 557].

[5] For this and the following cf. [Lewis 1973a, p. 557].

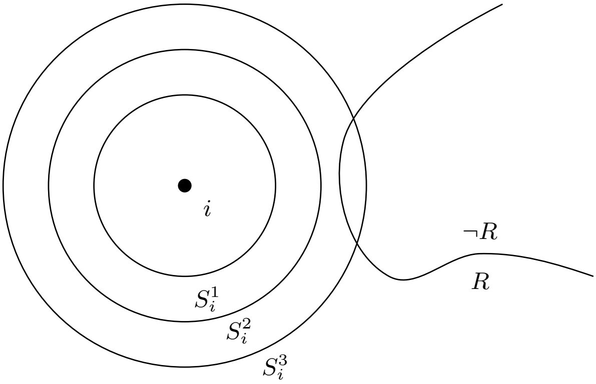

Fig. 2.1: LEWIS' possible worlds semantics, illustrated: Our actual world i, concentrical spheres S^1, S^2, S^3 of the same similarity to i with regions in which R holds or does not, respectively.

2.3 A possible worlds semantics with similarity

"Possible alternatives" to the "actual situation" are understood by LEWIS as metaphysically existing possible alternative worlds, centered around our actual world i in concentrical spheres according to their respective degree of similarity to i. As a matter of fact, such considerations can be made relatively to an *arbitrary*, distinguished world w. The spheres around our actual world explicate the structure of the similarity relation:[6]

1. It ought to be a weak ordering of worlds, within which two worlds may be on the same level being of same similarity with respect to the center; and any pair of worlds must be commensurable in that sense.

2. Our actual world ought to be the most similar to itself – more similar than any world different from it.

3. Moreover, there may not exist a unique set of worlds that are most similar but not equal to the actual world; in that sense the ordering may be dense and admit worlds which are more and more similar but never equal to i in an infinite regress. Claiming this, LEWIS disagrees with the initial ideas of Robert STALNAKER, who postu-

[6] Cf. [Lewis 1973a, p. 560] and [Weatherson 2009, section 3.2].

lates a uniquely distinguished sphere of most similar worlds with his *Limit Assumption*.

Any single of this possible worlds can be understood as an exhaustive state description: If it is raining in our actual, world we find the claim R to be true in the center i of our modeling. In some other world differing from ours with respect to the weather we would encounter $\neg R$. Typically, as the *meaning* of single claims like R or $\neg R$ their respective worlds are bundled to *propositions*, i. e., sets of possible worlds.

The counterfactual statement "If ϕ had occurred, ψ would have occurred as well" may now be evaluated in possible worlds semantics:

$$\phi \mathrel{\Box\!\!\rightarrow} \psi \tag{2.1}$$

is true exactly if and only if there is no world in which $\phi \wedge \neg\psi$ holds being closer to our actual world i than a world in which $\phi \wedge \psi$ holds. Or in David Lewis' own words:

> *[...] a counterfactual is nonvacuously true iff it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent.*[7]

And in formal fashion:

$$\begin{aligned} i \vDash \phi \mathrel{\Box\!\!\rightarrow} \psi \quad :\Longleftrightarrow \quad & \neg\exists w(w \vDash \phi) \ \ \Vee \\ & \exists u\Big(u \vDash \phi \wedge \psi \Wedge \forall v(v \vDash \phi \wedge \neg\psi \Longrightarrow v >_i u)\Big), \end{aligned} \tag{2.2}$$

for possible worlds w, u, v and the similarity relation $\leq_i$ with respect to our actual world i. The truth of the second disjunct of (2.2) depends on the existence of some possible world u satisfying $\phi \wedge \psi$ and being closer to i than any $\phi \wedge \neg\psi$-worlds. This second disjunct is not formulated relative to *the* $\phi \wedge \psi$-worlds closest to i, because Lewis strictly rejects the *Limit Assumption*, i. e., the assumption that for any specific proposition there is a unique set of worlds closest to some fixed actual i. Instead, there could be closer and closer worlds (or narrower and narrower spheres of possible worlds, respectively) infinitesimally close to but never reaching $\neg(\phi \wedge \psi)$ in a *limitless* infinite regress of refinement – in contrast to, e. g., Robert Stalnaker, who endorses the *Limit Assumption* in his

[7] Cf. [Lewis 1973a, p. 560].

formulation of the truth conditions of counterfactuals.[8] Moreover, in the non-continuous finite and discrete case the truth of counterfactuals can always be evaluated deploying the limit assumption, because for any proposition *the* i-closest worlds can be made out uniquely.

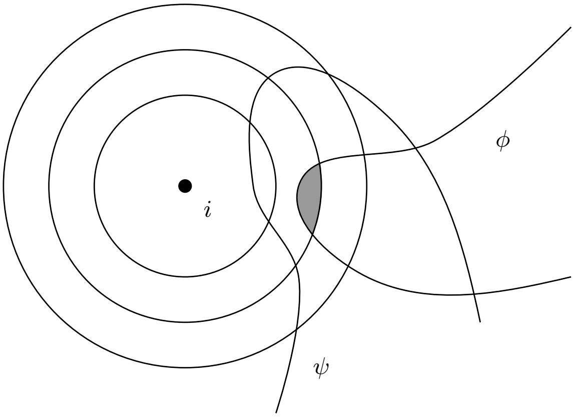

Fig. 2.2: $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is true at i if and only if there is no $\phi \wedge \neg\psi$-world that is closer to our actual world i than a $\phi \wedge \psi$-world.

Formula (2.2) can be rephrased in set-theoretic notation so that the similarity relation $\leq_i$ with respect to the actual world i is represented by the hierarchy of cumulative spheres $\$_i$, where $S_i^n = \bigcup_{x=0}^n S_i^x$:

$$\begin{aligned} i \vDash \phi \mathrel{\Box\!\!\rightarrow} \psi \quad :\Longleftrightarrow \quad & \bigcup \$_i \cap [\![\phi]\!] = \varnothing \;\; \vee \\ & \exists S \in \$_i \Big(S \cap [\![\phi]\!] \neq \varnothing \wedge S \cap [\![\phi]\!] \subseteq [\![\psi]\!] \Big), \end{aligned} \tag{2.3}$$

where $[\![\phi]\!] := \{u \,|\, u \vDash \phi\}$ represents the worlds where ϕ holds (in other words, the *proposition* ϕ). Analogously to (2.2), this formulation does not postulate that there be *the* narrowest sphere S in which ϕ-worlds can be found, but holds true if there is a sphere S, at all, so that the ϕ-worlds within S are included in the set of ψ-worlds. If this is the case, then it already holds for the sphere with the ϕ-world most similar to i, since the similarity relation $\leq_i$ is represented in the cumulative hierarchy of $\$_i$.

[8] See e. g. [Weatherson 2009, sect. 3.2] for a comparison of Lewis' and Stalnaker's differing truth conditions of counterfactuals and Lewis' discussion of the *Limit Assumption* in [Lewis 1973b, pp. 19 ff.].

Two technical remarks shall once more illuminate potential border cases:[9]

1. If there is no world where the antecedent of the counterfactual is evaluated with truth value 1, the counterfactual statement is defined to still hold *vacuously* – this is due to the first disjuncts of (2.2) or (2.3), respectively.

2. According to the formulations above, the counterfactual statement also includes the non-counterfactual case, i. e., where the antecedent ϕ already holds in our actual world. In this case the counterfactual is true if and only if the mere material implication with the same sub-statements holds true. Accordingly, ψ must also hold in our actual world.

Lewis' critics are particularly aiming at his similarity measure which he ultimately employs to objectivize counterfactual statements: What are the criteria for such a measure? Does one not need to consider, in addition to a given reference world, a certain aspect in question, with respect to which one can speak of greater or smaller similarity between worlds? What is a manageable metric on this relation supposed to look like? Is it not the case that any possible similarity assessment in the very core rests on the subjective evaluation of one's environment? And how is the following example to be analyzed?[10]

Example
If Richard Nixon had pushed the button, there would have been nuclear war.

Does a world with a sole dysfunctional button resemble our actual world to a greater degree than an alternative world in which we are facing a nuclear catastrophe? Lewis allows insight into his conception in his paper from 1979, *Counterfactual Dependence and Time's Arrow*, in which he equips the aspects of his similarity measure with priorities in imperative manner:

1. It is of the first importance to avoid big, widespread, diverse violations of law.

2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

[9] For this and the following cf. [Lewis 1973b, p. 16 ff.] and (in causal-theoretic context) also [Lewis 1973a, p. 560 f.].

[10] For the following cf. [Weatherson 2009, sect. 3.3].

3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

With hindsight to this list of priorities, a world where all machinery of modern warfare targeted at nuclear destruction failed would resemble our actual world far less than the post-nuclear apocalypse. We are certainly not having a hard time comparing worlds where logical necessities are overridden, where the laws of physics are suspended, or where extensive geographical restructuring affects the environment. The question remains, if the proposed similarity measure over possible alternative worlds is (i) natural, i. e., in accordance with our conceptualization, and (ii) to be understood as operationally effective in any way.

2.4 From counterfactual dependence to veritable causes

To be able to identify causal relationships between clusters (i. e., sets) of events, David Lewis extends the counterfactual analysis to relationships between families of propositions.[11] Now, if each two elements from these coupled sets of propositions are related counterfactually, we speak of counterfactual dependence between the respective sets of propositions. Typically, measuring processes as well as observation and control routines are characterized by counterfactual dependence between large families of alternatives: E. g., the family of alternative barometer readings counterfactually depends on the family containing alternative values of atmospheric pressure – under the assumption, the barometer works properly, is calibrated correctly, not perturbed etc.

Without much hesitation Lewis reduces causal dependence (in any case between actual events such as lightning in a thunderstorm etc.) to counterfactual dependence between the according propositions (i. e., sets of possible worlds).

Hume's counterfactual formulation – the second part of the twin definition – thus parallels Lewis' definition of *causal dependence* between two events, not of *causation* itself. This dependence holds relative to the truth of the following pair of counterfactuals:

[11] For this and the following cf. [Lewis 1973a, p. 561 f.].

(i) $\phi \;\Box\!\!\rightarrow \psi$ and

(ii) $\neg\phi \;\Box\!\!\rightarrow \neg\psi$.

Causal dependence (or its extension to *causal chains*, i. e., chains of events, where each event causally depends on the respective predecessor) between two specific events entails causation, according to David Lewis. The reverse direction must not be taken for granted: Causation does not imply causal dependence, in general. Causation should be transitive if it is to capture our intuitions, while this does not apply to causal dependence. Lewis be quoted here with his own example (illustrated in figure 2.3):

> *[...] there can be causation without causal dependence. let c, d, and e be three actual events such that d would not have occurred without c and e would not have occurred without d. Then c is a cause of e even if e would still have occurred (otherwise caused) without c.*

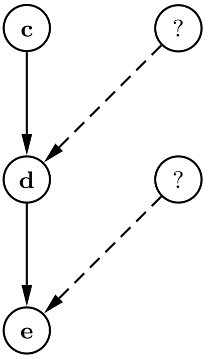

Fig. 2.3

"otherwise caused"? – The quote leaves the reader wondering, since the scenario, as stated, does not leave room for other causes, and the twofold counterfactual analysis in the above formulation expresses precisely the *necessary* causal succession $c \rightarrow d \rightarrow e$. So, how can Lewis' words be understood? Firstly, the example is not to be understood as a self-contained story. It does not depict a *closed world situation* with no potential external influence. The strand $c \rightarrow d \rightarrow e$ is to be seen as part of some metaphysically existing web of causes and effects that we have access to via reasoning and intuition. When we talk about such scenarios we might very well agree on more causally influential entities that we deem relevant, e. g., for explanation. c remains a *potential* (but void) cause of an actually occurring event e, even if c is *prevented* or *pre-empted* by other obtaining circumstances. Secondly, Lewis strictly rejects the possibility of *backtracking counterfactuals.* Knowledge about the consequent does not tell us anything about the antecedent. This precisely reflects the idea that the actual cause of an event may emerge from the abundance of potential cause candidates and override other (maybe standard) causes. c and e do stand in the relation of causation, even if c did not take part in bringing about e. Judea Pearl will make explicit what it means for an event to be caused by unforeseen circumstances – relative to a fixed set of event variables in a (probabilistic) causal model. Lewis' cause candidate *miracles* (which ultimately unfold the concentrical structure of the possible worlds semantics) are replaced in Pearl's framework by the breaking of the *closed world assumption* qua interventions in formal systems.

Lewis' *Causation* finally deals with the analysis of epiphenomena and pre-empted, potential causes to cover more intricate cases and to trace our reasoning and intuitions within the counterfactual framework.[12] Until 1986 a total of six *Postscripts* are drafted by Lewis to address various aspects that obviously seem to him to be explained in too little detail in *Causation*. *Causal dependence* is therein replaced by *quasi-dependence* to facilitate the analysis of special border cases of pre-emption. In his later 2000 paper, *Causation As Influence*, he even discards this approach altogether again – this time in favor of a completely new theory which takes causation to be some gradual influence of causes on potential effects.

2.5 Pearl's reply to Hume

When examining Hume's opening quote Judea Pearl utterly agrees with David Lewis: Regularity analysis *must* fall short.[13] In modern terminology: Correlation does not suffice to identify causes and effects. Part two of Hume's quote, his rephrasing in "other words" to clothe causal analysis in counterfactual fashion, cannot be an equivalent presentation of the problem for Pearl's taste: Correlations are based on observations, while deciphering counterfactual statements seems to be a virtual exercise of the mind. Pearl decisively refrains from such reductionist approaches and strongly campaigns for admitting natural causal assumptions in one's reasoning – he answers Hume and Lewis (and thinkers of the same direction like John S. Mill) directly:

> *[...] [D]iscerning the truth of counterfactuals requires generating and examining possible alternatives to the actual situation as well as testing whether certain propositions hold in those alternatives – a mental task of nonnegligible proportions. Nonetheless, Hume [...] and Lewis apparently believed that going through this mental exercise is simpler than intuiting directly on whether it was A that*

[12] Cf. [Lewis 1973a, pp. 565 ff.].

[13] For this and the following cf. [Pearl 2009, p. 238].

> *caused B. How can this be done? What mental representation allows humans to process counterfactuals so swiftly and reliably, and what logic governs that process so as to maintain uniform standards of coherence and plausibility?*

Moreover, Pearl makes out some inherent circularity in the similarity relation over possible worlds as proposed by Lewis: When assessing varying deviations from actuality in accordance with the above mentioned weighting, one cannot simply apply *arbitrary* principles – they must in any way at least conform with our conception of causal laws.[14] The nuclear first strike lies just on one causal line with Nixon's decision to push the fateful button. Less resistance is offered here to this causal flow than in a world with a deficient mediating button.

To evade such a circle, Judea Pearl turns to concrete identifiable, invariant single mechanisms for the comparison of two alternative situations. These single mechanisms may well be resting on causal assumptions, which nevertheless only become relevant locally within exact confines.

2.6 Pearl's agenda

In his book *Causality* (2000, and extended in the second edition 2009) Pearl explicates the philosophical and technical fundament of his approach towards modeling causal relationships. Although the notion of causality almost conveys something like lawlike necessity, as Pearl stresses on the first pages of his book, and the notion of probability rather seems to imply uncertainty and lack of regularity, various good reasons point in the direction of a fruitful exploration of the probabilistic treatment of causation.[15] With this probabilistic approach Pearl follows thinkers like Hans Reichenbach, I. J. Good, and in particular Patrick Suppes who illustrated the foundation of this agenda on probabilistic maxims exemplarily: By giving reasons like "you will fail the course because of your laziness" we know very well, that the antecedent (the laziness) makes the consequent (failing the course) more probable, but surely not absolute certain. A language of causality should capture such an intuition, in any way. Another protagonist be mentioned here: In the 1980s Wolfgang Spohn developed his ranking functions on probabilistic fundaments as well.

[14] Cf. [Pearl 2009, p. 239].

[15] For this and the following cf. [Pearl 2009, chapter 1].

Pearl avails himself of methods he finds among statisticians, who had successfully offered observations and measured data in compact form to various disciplines for years. In the following I want to present the keystones of the technical background in due brevity.

To recover information about observable dependences from raw data one may make use of *probabilistic models*, represented by *joint probability functions*. A probabilistic model is an encoding of information that permits us to compute the probability of every well-formed sentence S in accordance with the *Kolmogorov Axioms*.[16] S can here be seen as one specific event, in particular an *elementary event* to which all the random variables under consideration contribute a certain value. In the case of a dichotomous variable A, the contribution to the conjunction S will either be a or $\neg a$.

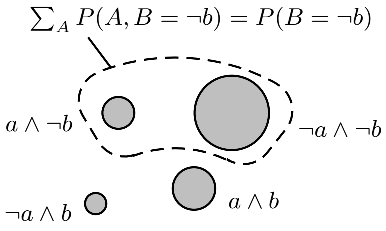

Fig. 2.4: The joint probability function over A and B assigns mass to possible worlds.

A *joint probability function* provides exactly the required assignment of probabilities to elementary events. This can be interpreted as an assignment of mass (or weight) to a universe of possible worlds representing these elementary events, i. e., a set of disjoint formulae. Figure 2.4 displays an example: A universe of four possible worlds represents a set of four disjoint formulae, namely the combinatorial permutations of possible values (a, $\neg a$, b, $\neg b$) of the dichotomous (random) variables A and B. The size of the different worlds (i. e., the diameter of the circles in this geometric interpretation) is then viewed as mass or weight attributed to those worlds by the joint probability function. Any other desired quantity may be calculated using this information: The *marginal* probability $P(B = \neg b)$ is simply the combined mass of all worlds in which $\neg b$ holds (as shown in figure 2.4) – regardless of the other circumstances in these worlds. Whenever a joint distribution function P over n random variables $X_1, \ldots, X_n$ is available, it is possible to perform a factorial decomposition of $P(x_1, \ldots, x_n)$[17] in accordance with the general form

[16] Cf. [Pearl 2009, p. 6] or below, pp. 137 ff.

[17] In this notation, $x_1, \ldots, x_n$ represent specific values of the variables $X_1, \ldots, X_n$.

$$P(A,B) = P(A \,|\, B)P(B). \tag{2.4}$$

Iterated application of (2.4) permits the decomposition of $P(\cdot)$ as a product of n conditional distributions:[18]

$$P(x_1,\ldots,x_n) = \prod_j P(x_j \,|\, x_1,\ldots,x_{j-1}). \tag{2.5}$$

In so-called Bayes nets independences between random variables may be presented clearly and compactly. In this context, random variables actually are measurable functions from a probability space in a measurable observation space, or in other words: Random variables, taken as functions, relate concrete (sets of) outcomes of a random experiment to the mathematical representation of these outcomes, e. g., on a discrete scale. The days of the week for example would be assigned the integers between 0 and 6.[19] We always expect independence between two variables – intuitively speaking – if we do not expect the value of one variable to influence the value of the second variable (to be exact, we are examining the *changing* of values). In general, we would not expect the occurrence of the event *Rain*, or *No Rain*, to have any influence on the day of the week, and vice versa.

A Bayes net can be defined as a tuple consisting of a directed acyclic graph G and a set of random variables V which are represented in the graph as nodes. The directed edges (arrows) in the graph G encode precisely the available knowledge about conditional independences between the represented variables – in accordance with their joint probability distribution. For any two independent variables in the Bayes net it can be stated that the probability of the first variable conditional on the second variable already equals the mere a priori probability of the first variable alone, formally:

$$A \perp\!\!\!\perp B \iff P(A \,|\, B) = P(A). \tag{2.6}$$

Here, limiting the sample space to certain B outcomes (e. g., Wednesday as one day of the week) has no effect on the probability of certain A outcomes (e. g., the occurrence of rain). The mathematical form of conditional dependence is given by the famous *Bayes Theorem*:

$$P(A \,|\, B) \;=\; \frac{P(B \,|\, A) \cdot P(A)}{P(B)}. \tag{2.7}$$

[18] Cf. [Pearl 2009, p. 14] – equation (1.30).

[19] For a specification of the notion of random variable see below, pp. 137 ff.

In particular, a Bayes net satisfies the Markov property: Every variable X is, conditional on its parent variables PA_X, i. e., parent nodes in the graph G, independent of all its non-descendants. Referring back to Equation (2.5) we can define the concept of *Markovian parents*:

Definition 2.6.1 (Markovian Parents)[20]
Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables, and let $P(\mathbf{v})$ be the joint probability distribution on these variables. A set of variables PA_j is said to be Markovian Parents *of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_n\}$ satisfying*

$$P(x_j \,|\, \mathbf{pa}_j) = P(x_j \,|\, x_1, \ldots, x_{j-1}) \tag{2.8}$$

and such that no proper subset of PA_j satisfies 2.8.[21]

Put in prose: Direct parent nodes screen off their child nodes. And already interpreted causally: Direct causes screen off their effects from potentially perturbing predecessors.

The concept of *Markovian Parents* can easily be understood within a graph-theoretical framework – namely by applying it to the parents of a node in a (directed) graph. PEARL concludes that a necessary condition for a DAG G to be a Bayesian network of probability distribution P is for P to admit the product decomposition dictated by G, as given in definition 2.6.1[22] (i. e., linking parent nodes in the graph to random variables which will be Markovian parents to those random variables linked to the child nodes of the aforementioned parent nodes). This leads to the definition of *Markov Compatibility*:

Definition 2.6.2 (Markov Compatibility)[23]
If a probability function P admits the factorization of definition 2.6.1 relative to a DAG G, we say that G represents P, that G and P are compatible, or that P is Markov relative to G.

As the philosophical foundation for the mapping of concrete situations to such Bayesian nets Thomas BAYES' very own interpretation of probabilistic quantities may be called on. In his works probabilities

[20] In [Pearl 2009, p. 14]: definition 1.2.1.

[21] As PEARL adds: Lowercase symbols (e. g., x_j, pa_j) denote particular realizations of the corresponding variables (e. g., X_j, PA_j).

[22] Cf. [Pearl 2009, p. 16].

[23] In [Pearl 2009, p. 16]: definition 1.2.2.

are not understood frequentistically, but rather as subjective degrees of personal convictions – degrees of belief. Developing efficient, graph based algorithms in the early 1980s Pearl coins the term 'Belief Propagation,' the transfer of more or less solid convictions with respect to certain possible facts, i. e., realization of variables, respectively, induced by correlation in the underlying probability distribution and represented by the directed edges in the graph. A short remark in parentheses: Wolfgang Spohn for example carries this line of thought further by stating that causation lies *in the eye of the beholder.*[24]

When establishing his framework, Judea Pearl emphasizes directed acyclic graphs as a powerful means for the analysis of complex causal relations. As he states, the role of graphs in probabilistic and statistical modeling is threefold:[25]

1. they provide convenient means of expressing substantive assumptions;
2. they facilitate economical representations of joint probability functions; and
3. they facilitate efficient inferences from observations.

A directed acyclic graphs answering queries about conditional independences may not necessarily be interpreted as reflecting causal relationships between variables right away. In most cases an alternative graph with a different ordering of the variables under consideration, i. e., nodes, respectively, may be constructed for the very same list of independence statements. Lightning and thunder with a reversible arrow in the graphical interpretation is a prototypical example. Nevertheless, there obviously exists a certain preference for a variable ordering when modeling even complex situations, as Pearl notes. He calls on Reichenbach's famous dictum from 1956, his *Common Cause Principle*, which may be applied to Markovian net structures if these are interpreted causally: *"No correlation without causation."*[26] – In other words: If two variables are probabilistically dependent, either one variable exerts causal influence on the second, or there exists a third variable as a common cause of the first-mentioned (thereby indirectly dependent) variables.

[24] Cf. Spohn: *Ranking Theory* (forthcoming – 2009, p. 422).

[25] Cf. [Pearl 2009, p. 13].

[26] Cf. [Pearl 2009, p. 30].

Certainly, the most salient ingredient in Pearl's analysis of causation on probabilistic foundations ultimately are deterministic functions determining the value of each variable in the modeling by taking as arguments merely the values of the parent nodes together with some potential exogenous and stochastic influence. These autonomous mechanisms are to be understood as asymmetric assignments; as a list of structural equations they describe the skeleton of a causally interpreted graph:

$$x_i = f_i(\mathbf{pa}_i, u_i), \quad \text{with} \quad i = 1, \ldots, n; \tag{2.9}$$

put in prose: The value x of the ith variable X is determined by a uniquely assigned equation f_i, that takes as arguments the set of parent variables of the ith variable together with a stochastic, uniquely assigned, non-observed disturbance quantity, which does not appear in the modeling as measured variable.

Advocating these deterministic mechanisms Pearl picks up what he grants the developers of the Structural Equation Modeling (SEM) framework: the *directed* nature of value assignment by structural equations with *causal* intension. That this intension seems to be suppressed from theory and practice and forgotten altogether Pearl does not get tired of pointing out.

To sharpen the contrast with Lewis once again, it shall be reminded that, in opposition to Pearl's structural approach with *deterministic* causal mechanisms, Lewis also speaks of *intrinsically uncertain* effects, in general, and writes in the supplemental Postscript B to *Causation* under the title "Chancy Causation" the following:[27]

> *[...] I certainly do not think that causation requires determinism. (Hence I regard "causality" as a naughty word, since it is ambiguous between "causation" and "determinism.")*

The reader wonders, on what scale "causation" and "determinism" might mark the extremes. David Lewis goes on in the same passage:

> *Events that happen by chance may nevertheless be caused. Indeed, it seems likely that most actual causation is of this sort. Whether that is or not, plenty of people do think that our world is chancy; and chancy enough so that most things that happen had some chance, immediately beforehand, of not happening.*

[27] Cf. [Lewis 1986b, p. 175].

Pearl does not have to discern uncertain causation and deterministic processes in his conceptualization: Causal mechanisms as structural equations represent just those invariant causal laws, while the causal flow may well be "diverted" in concrete cases by (i) exogenous, i. e., non-observed and unmodeled influences or by (ii) intentional calibration of these very deterministic mechanisms for the comparison of alternative settings.

To complement technical conceptualities with intuitions, I will present a compact standard example as illustration in the following.[28]

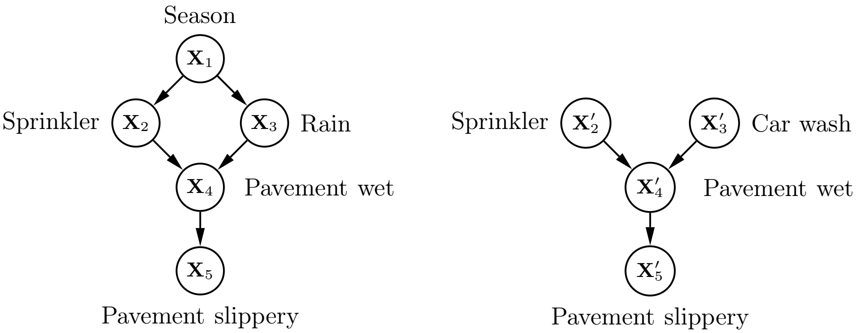

Fig. 2.5: Example of a Bayes net with five variables – on the right a modified variant without superstructure.

Figure 2.5 shows the graph of a Bayesian network representing dependencies among the variables X_1 through X_5 in compact form. X_1 stands as sole four-valued variable for the season of the year, X_2 stores whether it is raining or not, and X_3 whether the sprinkler in the front yard is switched on or off. The node X_4 represents the question, if the pavement is wet, X_5 the subsequent question, if the pavement is slippery in addition. The variables X_2 through X_5 are dichotomic variables and assume the values 'true' or 'false'.

The situation depicted by the left graph might be found in a typical Californian suburb where people switch on their sprinklers in the front yard to water the lawn during the hot and dry summer months and leave the sprinkler off to save water in the winter time when the chance of rain every once in a while is not too bad. Either way, if it rains or if the sprinkler is on, the curb separating the front yard from the street gets

[28] This example is taken from [Pearl 2009, pp. 21 ff.].

wet. As a consequence, the thin film of dust and dirt on the pavement turns into some slippery slide. The ecologically aware residents moreover base their decision (whether to switch the sprinkler on or not) on the current season of the year and their experience, that tells them if rain is rather likely to occur in that season or rather unlikely: The sprinkler will certainly be switched on during the hot and dry summer months, maybe operated by an automatic timer for convenience.

The information expressed by the left graph can be read off the edges in the graph, the *absent* (possible) edges, and the chosen arrow directions conveying knowledge or assumptions about dependencies among the variables. In accordance with the Markov property parent nodes as direct causes *screen off* their children from the influence of any *non-descendants.* E. g., whether the pavement gets slippery in the end (i. e., whether X_5 takes the value 'true') does not depend on the influence of any variable not mediated by X_4 – neither the occurrence of rain nor the running sprinkler *directly cause* the pavement to become slippery but first influence the wetness of the ground which in turn entails the slipperiness of the pavement.

Figure 2.5 does not display any exogenous, non-observed confounders influencing each variable distinctly independently. Each variable X is paired with a respective variable U (for "unobserved") with the same lower index. The invariant causal mechanisms underlying the modeling of our scenario can now be listed as functions:

$$\begin{aligned}
x_1 &= u_1 \\
x_2 &= f_2(x_1, u_2) \\
x_3 &= f_3(x_1, u_3) \\
x_4 &= f_4(x_2, x_3, u_4) \\
x_5 &= f_5(x_4, u_5)
\end{aligned} \tag{2.10}$$

x_1 is assigned a value only from outside the model. The value of x_2 is derived from evaluating x_1 *and* exogenous influences, as well as the value of x_3, and so forth.

The nonlinear functions f_2 through f_5 can now be specified as follows:

$$\begin{aligned} x_2 &= [(X_1 = \text{spring}) \vee (X_1 = \text{summer}) \vee u_2] \wedge \neg u_2', \qquad (2.11)\\ x_3 &= [(X_1 = \text{fall}) \vee (X_1 = \text{winter}) \vee u_3] \wedge \neg u_3',\\ x_4 &= (x_2 \vee x_3 \vee u_4) \wedge \neg u_4',\\ x_5 &= (x_4 \vee u_5) \wedge \neg u_5', \end{aligned}$$

with x_i on the right side representing the assignment of the truth value 1 to the respective variable X_i. The u values stand for potentially contributive or also preventative, exogenous, i. e., unmodeled influences. x_2 becomes true if it is spring or summer or if some unexpected influence u_2 contributes positively, as long as no unexpected influence $\neg u_2'$ completely prevents the assignment of truth value 1 to x_2. u_4 in the third line marks some potential additional (but unmodeled) influence causing the pavement to be wet – e. g., some burst water pipe – while $\neg u_4'$ as obstructive antagonist might stand for some plastic cover on the pavement not considered during modeling phase.

In contrast to purely probabilistic models in the form of joint probability distributions, Bayesian networks with structural equations may readily be enriched by additional influences, i. e., further variables. As when inserting a new component into a schematic circuit diagram, the effects of such a modification can be understood quite easily, precisely because it takes place locally and behaves clearly directionally. Some plastic cover on the pavement may be integrated into the model as local attachment of nodes and edges to the graph and as update of specific lines in the list of equations. Such an augmentation can be understood as "zooming in" on the scenario.

Certain settings may be analyzed in quite complex structures. To enable the analyst to read off from the graph which variable influences which other variable, Pearl offers a graphical criterion fit for this task: the so-called d-separation criterion for directed graphs. Applying this tool one may determine if the flow of information along the paths in the diagram is blocked possibly, or – interpreted a different way – if the transfer of degrees of belief along a certain path works or does not. And expressed in the terminology of Bayesian networks: Whenever a joint probability function P and a DAG G are *Markov compatible* in accordance with definition 2.6.2, one should be able to read off the graph the conditional independencies embedded in the probabilistic model represented by P.

To facilitate easy access to this information, Pearl gives a precise definition of his graphical criterion.

Definition 2.6.3 (*d*-Separation)[29]
What it means for a path (or analogously for two distinct nodes, respectively) to be d-separated by a set of nodes can be explicated on the basis of the pair of notions activated–deactivated *as follows:*

1. *A path p (i. e., a sequence of links) is said to be* deactivated *(or* blocked*) by a set of nodes Z iff p contains a chain* $i \rightarrow m \rightarrow j$ *or a fork* $i \leftarrow m \rightarrow j$ *where the middle node m is an element of Z.*
2. *If a path p is not deactivated in the first place it is said to be* activated *by a set of nodes Z iff it contains at least one* inverted fork *(also called* collider*)* $i \rightarrow m \leftarrow j$ *such that the middle node m of* each *collider in p (or a descendant of such an m) is in Z.*[30]

Consequently, a set of nodes Z is said to d-separate two nodes X and Y iff every path from X to Y is inactive*: either* deactivated *by the choice of Z or* not activated*.*

The '*d*' in '*d*-separation' denotes *directional*, which is the reason for the twofold formulation of definition 2.6.3. It makes a significant difference how the arrows along the path under consideration are directed. Choosing a set of nodes Z can be understood as fixing one's knowledge about the elements of Z or as gaining information about specific realizations of the Z variables. The d-separation test in a graph thereby tells the researcher if he ought to change his beliefs regarding the realization of a variable Y in case a third variable X has changed, given the background knowledge about the variables in Z. In general, Pearl speaks of two sets of nodes X and Y being d-separated by the set of nodes Z – since these sets of nodes can be seen as complex variables as well, it will be sufficient to consider single nodes in the following for brevity.[31] What it means for a probability distribution compatible with the DAG G that two nodes (or sets of nodes) X and Y are d-separated in G is spelled out in formal fashion in appendix B, which restates Pearl's formulation of the implications of d-separation.

[29] Pearl gives a slightly different formulation of the d-separation criterion with def. 1.2.3 in [Pearl 2009, pp. 16 f.], yet another variant in [Geiger *et al.* 1990, pp. 513 f.], and a very compact presentation in [Pearl 1995, p. 671].

[30] It is important to note here that, once a path has been deactivated by an appropriate choice of Z, it cannot be activated by any set of nodes.

[31] If 'X' and 'Y' denote sets of nodes, a third set of nodes Z is said to d-separate X from Y iff Z blocks *every path* from X to Y. Cf. [Pearl 2009, p. 17].

The Californian front yard shall be consulted for illustration once more. Looking at the left graph of figure 2.5 we can see clearly what Pearl means by *blocking paths* (cf. part 1 of definition 2.6.3): The set $\{X_1\}$ blocks the path $X_2 \leftarrow X_1 \rightarrow X_3$. X_1, which encodes the *season of the year*, acts as a balance between the *status of the sprinkler* (X_2) and the *occurrence of rain* (X_3) in our modeling intention – whenever the season provides rain, the sprinkler will remain switched off, and vice versa. If one does not know anything about the value of X_1 (season), it is still possible to know the value of X_2 (sprinkler) by learning the value of X_3 (rain), since *Rain=off* makes *Sprinkler=on* more likely, and – again – vice versa. This knowledge is simply derived from the knowledge of the causal mechanisms at work (which in this case yield some kind of negative correlation between X_2 and X_3). Now, if one gains knowledge about the actual value of X_1 (i. e., which season it is), X_2 and X_3 become independent of each other: Knowing it is hot and dry summer, learning that the sprinkler is off *would not change our belief* in the amount of rain, namely that there is no rain at all. We would rather find some other exceptional explanation for the sprinkler being off, e. g., that it is broken.[32] What has just been applied to *forks* (paths of the pattern $i \leftarrow m \rightarrow j$) is also applicable to simple *chains* (paths of the pattern $i \rightarrow m \rightarrow j$) in the graph (e. g., the path $X_1 \rightarrow X_3 \rightarrow X_4$).

In the case of *inverted forks* (or *colliders*), part 2 of definition 2.6.3 applies. The only collider in the left graph of figure 2.5 can be found in the path $X_2 \rightarrow X_4 \leftarrow X_3$. For reasons of clarity, this structure is replicated in the right graph of figure 2.5, freed from the superstruction of $X_2 \leftarrow X_1 \rightarrow X_3$. The situation in this graph is similar to the above sketched: A running sprinkler in the front yard leads to wet pavement, just as the house owner washing his car in the driveway in front of the garage. Either way, the pavement gets wet, again, and turns into a slippery slide posing danger to passer-bys.

It is important to note here that the status of the sprinkler is *completely independent* of any car wash possibly taking place in the driveway – X'_2 and X'_3 are d-separated if we do not condition on anything (i. e., know nothing about X'_4, which encodes the wetness of the pavement). In other words: If we do not know the actual value of X'_4 (*is the pavement wet*

[32] Conditional independence is symmetrical, so if we know it is hot and dry summer, learning that it is raining would not change our belief in the status of the sprinkler, which is switched on during the summer as an unalterable rule in the situation sketched above – for example by an automatic timer.

or not?), learning the value of X'_3 (e. g., *someone is washing the car in front of the garage*) does not change our belief in X'_2 (*the status of the sprinkler*) or – to be a bit more precise – does not change our degrees of belief in the different possible values X'_2 can assume (expressed by the *unconditional* probability distribution $P(X'_2)$). The missing edge between X'_2 and X'_3 is a consequence of our modeling intention: E. g., the car owner would not switch off the sprinkler *because* he is washing the car (maybe because he has put the sprinkler close to the driveway and would not want himself to get sprinkled) – resulting in the absence of the path $X'_2 \leftarrow X'_3$. Neither does our modeling intention tell us that the status of the sprinkler somehow influences the decision of the house owner to wash his car or not – resulting in the absence of the path $X'_2 \rightarrow X'_3$.
Nevertheless, once we know the value of X'_4 (i. e., *condition* on the value of X'_4), the variables X'_2 and X'_3 become conditionally dependent on each other, and the path $X'_2 \rightarrow X'_4 \leftarrow X'_3$ is not *d*-separated any more. In other words, once we know the value of X'_4, learning the value of X'_3 would change our belief in the different possible values of X'_2 – and vice versa.

With his book *Causality* Judea Pearl is pleading for the use of graphs on the hunt for causal influences and efficient causes. He promotes those graphs as a mathematical means of precise notation in which complex relationships can be presented compactly and easily accessibly to the user of any discipline. In the second, extended edition of *Causality* Pearl emphasizes his standpoint in poetic manner:

> *As X-rays are to the surgeon,*
> *graphs are for causation.*[33]

2.7 From modeling to model

A graph as in figure 2.5 together with explicit functional and deterministic mechanisms exemplifies the analysis of situations as undertaken by economists, by sociologists, by epidemiologists, or diagnostic physicians. Joint probability distributions assigning degrees of belief (in Bayesian interpretation) to any possible combination of atomic outcomes as disjoint elementary events may serve as the basis for such a kind of study. A network may then be generated algorithmically to (graphically) display all conditional dependencies, however, not necessarily uniquely in

[33] Cf. [Pearl 2009, p. 331].

most cases. A characteristic momentum in the modeling phase is the embedding of natural causal assumptions which cannot be derived from the mere collection of data. These causal assumptions originate either in the expertise of the modeler or in rather robust basic intuitions: When posing the question, in which direction to point an arrow between two nodes representing the age of a person and secondly her susceptibility to a certain disease, we would prefer the arrow to be rooted in the age node without much dispute. Obviously, this decision is grounded in quite basal assumptions about stability and continuity of certain processes in our world.[34]

To extend the rather informal notion of *modeling* to the formal concept of *model* in the model-theoretic sense, I want to reproduce Pearl's definition of a *causal model* at this point. It certainly has to go beyond the concept of model in probability theory.[35] There, a *joint probability function* over the variables under consideration yields a truth value for any proposition. E. g., the proposition "the probability for event A to occur is greater than $\frac{1}{2}$" is assigned the truth value 1 or 0 – depending on the state of facts. Now, a causal model should be able to encode the truth value of statements about causal relationships. This includes sentences like

- "B occurred because of A,"
- "A may cause B,"
- "B will occur if we bring about A,"
- and counterfactual observations like "B would have been different were it not for A."

Obviously, such statements cannot be evaluated in standard propositional logic or any probability calculus, because they talk about – as Judea Pearl puts it – changes in the *outer world* and not about changing convictions regarding a *closed static world.* Thus, causal models are meant to provide information about possible *external* changes – they do that by explicitly representing structural mechanisms, which are to be modified through *external* alterations. More about that in the next step, but firstly on to the definition of a causal model.

[34] For this cf. Pearl's analysis of *Simpson's Paradox* in [Pearl 2009, pp. 177 f.].

[35] For this and the following cf. [Pearl 2009, pp. 202 f.].

Definition 2.7.1 (Pearl's Causal Model)[36]
A causal model *is a triple*

$$M = \langle U, V, F \rangle$$

where:

(i) U *is a set of* background *variables (also called* exogenous**), that are determined by factors outside the model;*

(ii) V *is a set* $\{V_1, V_2, \ldots, V_n\}$ *of variables, called* endogenous*, that are determined by variables in the model – that is, variables in* $U \cup V$*; and*

(iii) F *is a set of functions* $\{f_1, f_2, \ldots, f_n\}$ *such that each* f_i *is a mapping from (the respective domains of)* $U_i \cup PA_i$ *to* V_i*, where* $U_i \subseteq U$ *and* $PA_i \subseteq V \backslash V_i$ *and the entire set* F *forms a mapping from* U *to* V*. In other words, each* f_i *in*

$$v_i = f_i(\mathrm{pa}_i, u_i), \quad i = 1, \ldots, n,$$

assigns a value to V_i *that depends on (the values of) a select set of variables in* $V \cup U$*, and the entire set* F *has a unique solution* $V(u)$*.***,***

Pearl's respective footnotes shall be added for completeness in the following:

* [^{2} in Pearl's def.] We will try to refrain from using the term "exogenous" in referring to background conditions, because this term has acquired more refined technical connotations [. . .]. The term "predetermined" is used in the econometric literature.

** [^{3} in Pearl's def.] The choice of PA_i (connoting *parents*) is not arbitrary, but expresses the modeller's understanding of which variables Nature must consult before deciding the value of V_i.

*** [^{4} in Pearl's def.] Uniqueness is ensured in recursive (i. e., acyclic) systems. Halpern [*Axiomatizing causal reasoning. In G. F. Cooper and S. Moral, editors, Uncertainty in Artificial Intelligence, pages 202-210. Morgan Kaufmann, San Francisco, CA, 1998*] allows multiple solutions in nonrecursive systems.

Pearl's definition of a causal model, though formal and compact, comes with some loose ends, which are tightened in Appendix A (p. 137) by giving the formal mathematical definition of a random variable as

[36] Cf. def. 7.1.1 in [Pearl 2009, p. 203], footnotes given below with changed index symbols. Also see Pearl's preliminary def. 2.2.2, [Pearl 2009, p. 44] where a causal model is defined as a pair $M = \langle D, \Theta_D \rangle$ consisting of a causal structure D (which is a directed acyclic graph connecting a set of nodes V) and a set of parameters Θ_D that specify the functional value assignment of each of the nodes in V.

function of the outcome of a stochastic experiment. With these definitions at hand, we may revisit Pearl's definition of F, the set of deterministic mechanisms in the causal model. When looking at the sprinkler example, again, we can pick out one element of F, perhaps f_4, which is associated with the variable X_4 representing the answer to the question, if the pavement is wet or not. X_4 is assigned its value x_4 by feeding the values of its parent variables (in the Markovian sense) along with the stochastic influence u_4 into the corresponding function f_4: $X_4 = f_4(x_2, x_3, u_4)$. Each variable, seen as a function, has its own domain such that X_4 is actually assigned its value by $f_4(X_2(\omega_2), X_3(\omega_3), U_4(\chi_4))$ with $\omega_2 \in \Omega_2 = \mathit{Dom}(X_2) = \{\text{rain}, \text{no rain}\}$, $\omega_3 \in \Omega_3 = \mathit{Dom}(X_3) = \{\text{sprinkler on}, \text{sprinkler off}\}$, and undescribed $\chi_3 \in \mathit{Dom}(U_3)$, which is not specified in any more detail. After all, any exogenous influence is an unmodeled influence, by choice of design – $\mathit{Dom}(U_3)$ might therefore contain plastic covers, burst water pipes, shattered coffee mugs, etc. For reasons of clarity, here and in the definition below we combine all potential exogenous influences on X_4 into one complex variable U_4 (containing both positively contributing and preventative external factors).[37] Summing things up: The deterministic mechanisms f_i take as arguments elements of the respective ranges of the parent variables (and the one complex exogenous variable), seen as functions.

Part (iii) of definition 2.7.1 can now be supported with the following formulation of causal mechanisms:

Definition 2.7.2 (Causal Mechanisms)[38]
F is a set of causal mechanisms for V, i. e., n functions $\{f_1, f_2, \ldots, f_n\}$ (determining the value of each variable V_i in V) such that

$$v_i = f_i(\mathbf{pa}_i, u_i)^{39}$$

with

$$F = \{f_i \,|\, f_i : \big(\prod_k \mathit{Ran}(\mathrm{PA}_{i_k})\big) \times \mathit{Ran}(U_i) \rightarrow \mathit{Ran}(V_i)\},$$

where $1 \leq i \leq |V|$ and for every i: $1 \leq k \leq |\mathrm{PA}_i|$, $U_i \in U$ (possibly combining multiple contributing and/or preventing disturbance factors into one complex variable), and $\mathrm{PA}_i \subseteq V \backslash \{V_i\}$. The entire set F forms a mapping from U to V with a unique solution.

[37] This is possible w. l. o. g., since all external factors satisfy the Markov property by definition, i. e., they are influencing the associated variables distinctly independently.

[38] The formulation given here is in agreement with the definition of the concept of *structural model* with deterministic functions in [Halpern & Pearl 2005a, p. 847].

[39] The boldface $\mathbf{pa}_i$ collects for each variable V_i the values of its parent variables.

Example

Consider the random variable V_4 with the associated set of parents $\mathrm{PA}_4 = \{V_1, V_2, V_3\}$ and the complex variable U_4 as an exogenous unmodeled factor. V_4 is assigned its value by the function

$$f_4 : Ran(\mathrm{PA}_{4_1}) \times Ran(\mathrm{PA}_{4_2}) \times Ran(\mathrm{PA}_{4_3}) \times Ran(\mathrm{PA}_{4_4}) \times Ran(U_4) \to \Sigma_4,$$

where Σ_4 is the set of possible realizations (possible values) of V_4.
The concrete value is now assigned nonlinearly by

$$v_4 = f_4(\mathrm{pa}_{4_1}, \mathrm{pa}_{4_2}, \mathrm{pa}_{4_3}, \mathrm{pa}_{4_4}, u_4).$$

Every causal model can be coupled with a directed graph in which each variable of the model is represented by a node. The concrete *functional* and above all *autonomous* mechanisms are abstracted from in this causal graph by means of directed edges. Arrows from the parent nodes to the child nodes mirror the set of functions F. The specification of all nonlinear assignments is stored in the structural model itself.

2.8 Triggering causes, bringing about effects

Whether an event may be called a cause of a second event, obviously depends on how the influence of the associated first variable on the second variable behaves. In particular, one node in the diagram should be seen as a cause of a second node if assigning a specific value to the first node evokes a difference in the evaluation of the second one (in the vocabulary of Bayesian networks). PEARL's approach thus centers around the notion of *causal effect.* Such a causal effect may be tested in analogy with a controlled experiment in the laboratory: The scenario is manipulated locally, certain conditions of the setting are modified and fixed in such a manner that occurring changes in the values of observed variables can be measured. Now, quite in agreement with this procedure, in the causal model the value of a specific structural function will be modified and fixed, thereby *cutting the links* between the respective variables and their parents. As a formal expression of this intervention, of this manipulation from outside, PEARL introduces a new operator which does not become effective within a model but precisely converts one causal model into a second. The so-called $do(\cdot)$-operator, which may very well be read imperatively, thus induces a *transformation* of the model under consideration, unambiguously. In doing so, it explicitly breaks the *Closed World Assumption*, on which in particular probabilistic models rest.

A causal effect can now be expressed as *probabilistic quantity* which may be calculated from a probability distribution *upon transformation*:

Definition 2.8.1 (Pearl's Causal Effect)[40]
Given two disjoint sets of variables, X and Y, the causal effect *of X on Y, denoted either as $P(y \,|\, \hat{x})$ or as $P(y \,|\, do(x))$, is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y \,|\, \hat{x})$ gives the probability of $Y = y$ induced by deleting from the structural causal model all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*

This definition precisely expresses that the variable X does not depend functionally on any other variables any more. It will be assigned its value *from outside* by an intervention *external to the model.* This process of assigning is not encoded in the model itself, but is part of just such a transformation symbolized by the $do(\cdot)$-operator.

Our Californian sprinkler example may be consulted for illustration, once more. The notion of external intervention becomes more transparent if one sets out to examine the causal influence of the sprinkler on the slipperiness of the pavement: In our list of structural equations (2.10) the value of the random variable X_2 is set to 'switched on', i. e., 'true'. The corresponding equation thus becomes inoperative, and the value x_2 in the equation for X_4 will as well be fixed to 'true'. Any possible alternative for the value of X_5 is eliminated – it shall be remarked here, that is was certainly possible for X_5 to assume alternative values before the intervention, i. e., 'true' or 'false'. The unblocked causal flow from X_2 to X_5 now ultimately brings about the actual slipperiness of the pavement, of course modulo obstructive exogenous influences as plastic covers and such.

In the corresponding graph the modification of the structural equations becomes evident if for any variable the elimination of functional dependencies, graphically interpreted, means the elimination of influent edges. In the sprinkler example this means in particular that the transfer of degrees of personal belief between X_1 and X_2 becomes blocked. *Before* the intervention the modeling traces the mere *observation* of the setting. As soon as the running sprinkler is observed, one can infer with great certainty that is is summer or spring, due to the underlying positive correlation of X_1 and X_2. The dry seasons are finally responsible for the sprinkler being switched on, as was our modeling intention.

[40] Cf. def. 3.2.1 in [Pearl 2009, p. 70].

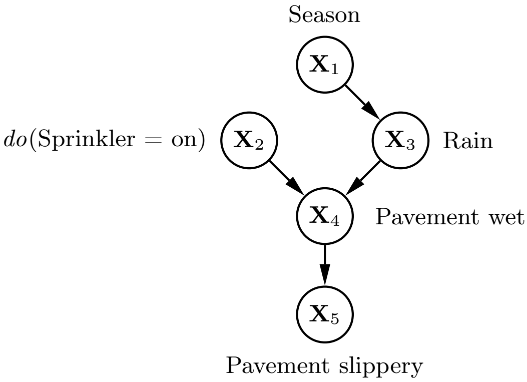

Fig. 2.6: The sprinkler is "switched on by intervention" – applying $do(\cdot)$ transforms the graph and breaks the link between X_1 and X_2.

An intervention external to the model can be understood as deliberate manipulation of the setting, not influenced by any conditions within the model. In the transformed model one cannot infer the dry season from the observation of the running sprinkler anymore. This deliberate manipulation of X_2 does not depend on the value of the variable X_1, in particular, which is marked by eliminating the connecting arrow: The sprinkler may now be switched on and off in all seasons virtually if the causal connection with the slipperiness of the pavement is to be tested.

On the basis of these structural local modifications, together with dependence tests and quantitative comparison, PEARL establishes the fine-grained formal representation of statements about causes, direct causes, indirect causes, and potential causes.[41] In short: The variable X is a cause of Y in this framework if (given the values of all background variables) there exist two possible values x and x' such that the choice of a value for X (either in favor of x or in favor of x') makes a difference in the evaluation of the variable Y.

David LEWIS shall be consulted once more for comparison: To determine in his possible worlds semantics counterfactually if an event P was causally responsible for a second event Q to occur, one had to stride through a metaphysically existent similarity space with great mental effort to test for metaphysically existent alternative worlds of various

[41] Cf. [Pearl 2009, p. 222].

similarity distance, whether the statements $\neg P$ and $\neg Q$ describe the respective settings there correctly – or not. Even when restricting ourselves to dichotomous variables, i. e., bivalent logic, respectively, depending on the number of propositional constants we only obtain a semi-decidable procedure for the identification of causes in the worst case, since we universally quantify over possible worlds.

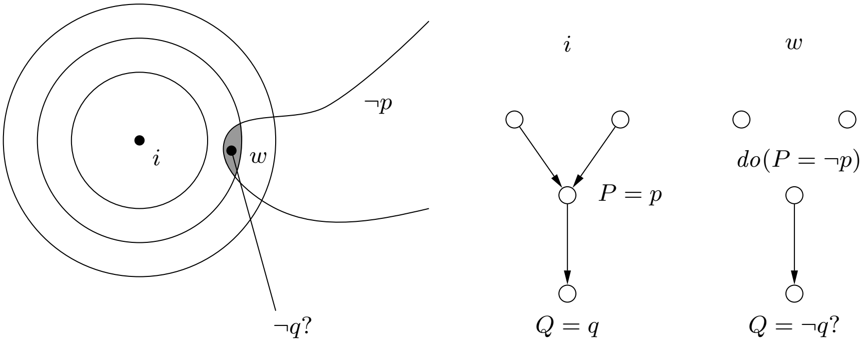

Fig. 2.7: On the search for alternative testing environments for Q moving from the *setting* i to the *setting* w – as proposed by Lewis (on the left) and Pearl (on the right).

In direct comparison: When moving away from the actual world i along the similarity relation in Lewis' framework, we have to check in the closest $\neg p$ worlds w, if we also find $\neg q$ to hold there. When modeling generic relationships in the actual world i in Pearl's formalism, we obtain a graph G mirroring the mere observation of correlations. The transition to an alternative world w where $\neg p$ is to be determined can be achieved qua intervention by means of the $do(\cdot)$-operator: The variable P is set to 'false', influent edges are eliminated. The question is now, whether the assignment of the P value also leads to a measurable difference in the evaluation of the variable Q. If the causal effect of P on Q in the model is identifiable (and Pearl gives algorithmic criteria for determining if it is or not), then it can be calculated uniquely and efficiently on the basis of the stable functional mechanisms. Identifying causes in Pearl's formalism can thus be understood both as *natural* and *operationally effective* at the same time, because the invariant mechanisms represent *intuitively obvious* basal assumptions, and because the external interventions are limited to local surgeries of the graph.

2.9 Computing observational data for causal inference

Making automated learning as efficient as possible is one of the chief goals of computer scientists and has driven research in the field of artificial intelligence. Several algorithms for inductive automated construction of Bayes nets from observational data have been developed and refined to achieve computational tractability. Typically, learning of Bayesian networks is divided into two tasks: (i) learning of the underlying (graphical) structure of the net, i. e., its topology, and (ii) determining the conditional probabilities for each node (which can be represented in CPT's – *conditional probability tables*).[42]

Causal inference is ultimately permitted by the stable skeleton of a causal model – its structure. Now, if we are given an arbitrary joint probability distribution over a fixed set of variables V and want to derive directed arrows from raw data, we need to have some benchmark to be able to compare different topologies with one another. The joint probability distribution, understood as the vector of masses that are assigned to all different possible worlds (with size $|V|^2$, in case all variables are dichotomous), will serve as benchmark which each distribution induced by a Bayes net structure candidate can consequently be measured against. Different distance measures have been suggested for this task. The difference between two joint probability vectors can be determined by using the Euclidian distance, thus summing up the squares of the distances of all vector components, i. e., the weight differences between each two corresponding possible worlds:

$$d_E(\overline{x}, \overline{y}) = \sqrt{\textstyle\sum_i (x_i - y_i)^2}, \tag{2.12}$$

where $\overline{x}, \overline{y}$ are vectors (in our case, of masses of possible worlds). Alternatively, the information-theoretic Kullback-Leibler divergence returns the weighted sum of the distances between the logarithms of each two corresponding masses:

$$d_{KL}(\overline{x}, \overline{y}) = \sum_i y_i(\log_2 y_i - \log_2 x_i), \tag{2.13}$$

again with the vectors $\overline{x}, \overline{y}$ as above. Further refinements introduce weighting factors into the distance measure to, e. g., rank thinner networks above denser ones in accordance with the demand for *minimal*

[42] Cf. for this and the following [Ertel 2009, pp. 219ff.].

structures.[43] Finally, creating an efficient algorithm for the search of the fittest Bayes net topology amounts to a minimization problem for which good heuristics are needed.

Example

Wolfgang Ertel *considers the complete probabilistic model a meteorologist might use to predict the amount of rainfall in the afternoon, solely projected from the weather conditions in the morning of the respective day.*[44]

Sky	*Bar*	*Rain*	$P(Sky, Bar, Rain)$
clear	rising	dry	0.40
clear	rising	rain	0.07
clear	falling	dry	0.08
clear	falling	rain	0.10
cloudy	rising	dry	0.09
cloudy	rising	rain	0.11
cloudy	falling	dry	0.03
cloudy	falling	rain	0.12

Table 2.1: The complete probabilistic model for the prediction of rainfall (*Rain*) in the afternoon, based on the weather conditions in the morning, if it is cloudy or clear (*Sky*), and whether the barometer rises or falls (*Bar*).

Table 2.1 displays the masses of all eight possible worlds, combinatorially listing the joint probabilities for the sky (Sky) to be clear or cloudy, the barometer (Bar) to rise or fall, and for rainfall (Rain) to occur or not. Since we want to compare vectors, the joint probability of the variables Sky, Bar, and Rain can now be presented as an 8-tuple:

$$\mathbf{P} = \langle 0.40, 0.07, 0.08, 0.10, 0.09, 0.11, 0.03, 0.12 \rangle \tag{2.14}$$

Making this vector explicit now allows us to compare the following structures with one another:

[43] Cf. e.g. [Ertel 2009, pp. 221 f.]; if N is the graphical structure of a Bayes net candidate to be measured against some joint probability distribution $\mathbf{P}$ given in the form of vector of masses, then N is ranked by $f(N) = \mathit{size}(N) + w \cdot d_{KL}(\mathbf{P}_N, \mathbf{P})$, where $\mathbf{P}_N$ is the joint probability distribution induced by the graph N, $\mathit{size}(N)$ is the number of entries in N's CPT's, and w is some additional weighting factor that needs to be adjusted manually to balance the criteria of size and vector distance.

[44] Cf. [Ertel 2009, p. 175 and pp. 220 f.].

(a) Sky → Rain ← Bar and

(b) Sky → Bar → Rain,

with the corresponding joint probability vectors $\mathbf{P}_a, \mathbf{P}_b$. *In the following, the computation of the distance between* $\mathbf{P}_a$ *and* $\mathbf{P}$ *shall be carried out in detail before* $\mathbf{P}_a$ *and* $\mathbf{P}_b$ *are compared in order to find the minimum distance vector (with respect to* $\mathbf{P}$*).*

Step 1: Calculate the marginal probabilities. *The a priori probabilities of the orphan variables Sky and Bar can be read off table 2.1 directly by summing up the probabilities of equal variable assignments:* $P(\mathit{Sky} = \mathit{clear}) = 0.65$ *and* $P(\mathit{Bar} = \mathit{rising}) = 0.67$. *Of course, in the dichotomous case* $P(X = \neg x)$ *is given by* $1 - P(X = x)$.

Step 2: Generate the CPT's for all inner variables. *This is done by determining for each non-orphan variable X the ratio of each of the probabilities for* $X = x$ *conditional on X's parent variables, i. e., the variables represented by its parent nodes in the graph. In the example this is only the variable Rain, for which the following table can be calculated:*

Sky	*Bar*	$P(\mathit{Rain} = \text{dry} \mid \mathit{Sky}, \mathit{Bar})$
clear	rising	0.85
clear	falling	0.44
cloudy	rising	0.45
cloudy	falling	0.2

Here, e. g., the first line is calculated from table 2.1 straightforwardly:

$$P(\mathit{Rain} = \text{dry} \mid \mathit{Sky}, \mathit{Bar}) \tag{2.15}$$
$$= \frac{P(\mathit{Rain} = \text{dry}, \mathit{Sky} = \text{clear}, \mathit{Bar} = \text{rising})}{P(\mathit{Sky} = \text{clear}, \mathit{Bar} = \text{rising})} = \frac{0.40}{0.40 + 0.07} = 0.85$$

Step 3: Generate the mass vector. *In accordance with the rule of iterated factorial decomposition (equation 2.5) and the independence condition stated in the definition of Markovian parents (equation 2.8 in definition 2.6.1), the mass vector induced by topology (a) can now be built up by listing the masses of all possible worlds, generated combinatorially.*

E. g., the first entry is calculated as follows:

$$\begin{aligned}
&P(\mathit{Sky} = \text{clear}, \mathit{Bar} = \text{rising}, \mathit{Rain} = \text{dry}) \\
&= P(\mathit{Rain} = \text{dry} \mid \mathit{Sky} = \text{clear}, \mathit{Bar} = \text{rising}) \cdot \\
&\quad P(\mathit{Sky} = \text{clear}) \cdot P(\mathit{Bar} = \text{rising}) \\
&= 0.85 \cdot 0.65 \cdot 0.67 = 0.37
\end{aligned} \tag{2.16}$$

This is repeated analogously for all other possible worlds, finally resulting in the vector

$$\mathbf{P}_a = \langle 0.37, 0.065, 0.095, 0.12, 0.11, 0.13, 0.023, 0.092 \rangle. \tag{2.17}$$

Step 4: Measure the vector distance. *Application of the measures above yields the distances*

$$\begin{aligned}
d_E(\mathbf{P}_a, \mathbf{P}) &= 0.0029 \quad \textit{and} \\
d_{KL}(\mathbf{P}_a, \mathbf{P}) &= 0.017.
\end{aligned}$$

Step 5: Find the minimum distance vector. *The above steps did only trace the procedure for Bayes net candidate (a). Of course, every conceivable alternative structure should be treated analogously. If steps 1–4 were to be performed for topology (b), the respective distances would be*

$$\begin{aligned}
d_E(\mathbf{P}_b, \mathbf{P}) &= 0.014 \quad \textit{and} \\
d_{KL}(\mathbf{P}_b, \mathbf{P}) &= 0.09.
\end{aligned}$$

Since both measures mark $\mathbf{P}_a$ as the vector closer to $\mathbf{P}$ than $\mathbf{P}_b$, Bayes net structure (a) is considered to fit the data better than (b).[45]

Nevertheless, the task of finding the minimum distance vector faces problems of complexity. The search space, i. e., the number of possible DAG structures for a given set of variables V, super-exponentially grows with the number of variables $|V|$. Efficient and plausible heuristics are needed to limit the search space before minimization is performed. Ertel suggests that one way of regaining computational tractability is to

[45] It is important to note that if an additional arrow *Sky* → *Rain* were to be inserted into (b), thereby rendering the graph *fully connected*, the resulting graph (c) would yield distance $d_E(\mathbf{P}_c, \mathbf{P}) = 0$, because in this uninformative presentation any two variables are dependent.

pick out those graphs that suit our basic causal assumptions, in the first place. If $V_1, \ldots, V_n$ is a *causal* ordering of the variables in V, only those graphs would be considered that contain directed edges $\langle V_i, V_j \rangle$ with $i < j$.[46] Since we precisely want to infer the set of potential causal structures merely from raw observational data, the latter suggestion seems to counteract our primary goals. Quite different from the metric approach, Thomas Verma and Judea Pearl constructively develop an algorithm for the generation of *marked patterns* from *stable probability distributions* (over observed variables), i. e., *classes of observationally equivalent latent structures* that are compatible with given data:

> *An autonomous intelligent system attempting to build a workable model of its environment cannot rely exclusively on preprogrammed causal knowledge; rather it must be able to translate direct observations to cause-and-effect relationships. However, given that statistical analysis is driven by covariation, not causation, and assuming that the bulk of human knowledge derives from passive observations, we must still identify the clues that prompt people to perceive causal relationships in the data. We must also find a computational model that emulates this perception.*[47]

In Pearl's formulation of the algorithm for induced causation, reducing computational complexity through manual selection of potential structural candidates (and perhaps erroneously dropping relevant graphs) is traded for the task of (if only partly) directing the edges in the limited set of graphs the algorithm finally returns.

Some notions must be clarified before the algorithm can be stated. What the algorithm will return is a class of observationally equivalent *latent causal structures*, i. e., a set of DAGs over observed and unobserved (*latent*) variables. Pearl adds the postulate of *structure preference*: One latent structure L is to be preferred to another one, L', if and only if the observed part of the DAG of the latter can *mimic* the observed part of the first. This amounts to saying that L should be favored if by tweaking the precise specifications of all the functional mechanisms of L' (represented by the edges in the graph of L') the joint probability distribution over the observed variables of L can be reproduced exactly.

[46] A *causal* ordering of the variables $V_1, \ldots, V_n$ can be understood as satisfying certain constraints – e. g., if time is considered constitutive of causation, there cannot be two variables V_i, V_j with $i < j$ and $V_j <_{temp} V_i$, where $<_{temp}$ indicates strict temporal precedence.

[47] Cf. [Pearl 2009, p. 42].

Succinctly, referring to the principle of Occams's razor Pearl claims that "following standard norms of scientific induction, it is reasonable to rule out any theory for which we find a simpler, less elaborate theory that is equally consistent with the data [...]. Theories that survive that process are called *minimal*."[48] Consequently, *minimality* is defined relative to a class $\mathcal{L}$ of latent structures such that a structure L is minimal with respect to $\mathcal{L}$ if and only if there is no other structure L' strictly preferred to L.[49] Then of course, as in the metric approach, the generated structure must fit the data, i. e., it must be *consistent* with the given distribution $\hat{P}$ over the observed variables. In other words, for a structure L to be consistent with observational data $\hat{P}$ there must be a specification of the functional mechanisms in L that induces a joint probability distribution (over all observed variables) equal to $\hat{P}$.[50] With these notions at hand we can explicate what *inferred causation* means:

Definition 2.9.1 (Pearl's Inferred Causation)[51]
Given $\hat{P}$, a variable C has a causal influence on variable E if and only if there exists a directed path from C to E in every minimal latent structure consistent with $\hat{P}$.

Pearl adds that he makes "no claims that this definition is guaranteed to always identify stable physical mechanisms in Nature. It identifies the mechanisms we can plausibly infer from nonexperimental data; moreover, it guarantees that any alternative mechanism will be less trustworthy than the one inferred because the alternative would require more contrived, hindsighted adjustment of parameters (i. e., functions) to fit the data."[52]

Nevertheless, theoretically we are not guaranteed that nonexperimental data will always be minimal in the sense that it has only one *unique* minimal causal structure (modulo *d*-separation equivalence). The additional *assumption of stability* tells us that given data $\hat{P}$ is usually highly unlikely to hide probabilistic dependencies by precise *cancelling*. Stability thus implies that the list of independencies embedded in $\hat{P}$ remains the same even if the specification of individual functional mechanisms

[48] Cf. [Pearl 2009, p. 45].

[49] As above, we definitely want to rule out the maximal case, i. e., the fully connected graph that could mimic the behavior of any probabilistic model if the parameters (functional mechanisms) are tweaked the right way.

[50] Cf. [Pearl 2009, pp. 45 f.] for the precise definitions of *latent structure*, *structure preference*, *minimality*, and *consistency*.

[51] Cf. [Pearl 2009, p. 46], definition 2.3.6.

[52] Cf. [Pearl 2009, p. 47].

changes.[53] Of course, when we decide to allow for latent variables in the structure, a stable input distribution $\hat{P}$ will not yield a unique minimal DAG, because – if not restricted by neighboring edges – the correlation of two variables can be due to either direct causal influence (either way) or an unknown common cause (a hidden, latent common parent variable). Accordingly, the IC* algorithm cumulatively enhances the structure to be built up by adding individual arrowheads step by step (possibly returning bidirectional edges). The output of IC* is then a *marked pattern* with four types of edges (explained below), representing the class of observationally equivalent minimal latent structures consistent with the data.

Pearl's IC* Algorithm (Inductive Causation with Latent Variables)[54]

Input: $\hat{P}$, a stable distribution (with respect to some latent structure).
Output: core($\hat{P}$), a marked pattern.

```
1. For each pair of variables $a$ and $b$, search for
   a set $S_{ab}$ such that $(a \perp\!\!\!\perp b \,|\, S_{ab})$ holds in $\hat{P}$.
   If there is no such $S_{ab}$, place an undirected
   link between the two variables $a — b$.
2. For each pair of nonadjacent variables $a$ and $b$
   with a common neighbor $c$, check if $c \in S_{ab}$.
      If it is, then continue.
      If it is not, then add arrowheads pointing
         at $c$ (i.e., $a \rightarrow c \leftarrow b$).
3. In the partially directed graph that results,
   add (recursively) as many arrowheads as
   possible, and mark as many edges as possible
   according to the following two rules:
      R1: For each pair of nonadjacent nodes $a$
         and $b$ with a common neighbor $c$, if the
         link between $a$ and $c$ has an arrowhead
         into $c$ and if the link between $c$ and $b$
         has no arrowhead into $c$, then add an
         arrowhead on the link between $c$ and $b$
         pointing at $b$ and mark that link to
         obtain $c \overset{*}{\rightarrow} b$.
```

[53] Cf. [Pearl 2009, p. 48] for the definition of stability.

[54] Cf. for this and the subsequently given characteristics of the resulting edges [Pearl 2009, pp. 52 f.].

R$_2$: if a and b are adjacent and there is a directed path (composed strictly of marked $*$-links) from a to b, then add an arrowhead pointing toward b on the link between a and b.

The resulting edges are divided into four groups:

1. a marked arrow $a \overset{*}{\rightarrow} b$, signifying a directed path from a to b in the underlying model (hinting at genuine causation);

2. an unmarked arrow $a \rightarrow b$, signifying either a directed path from a to b or a latent common cause $a \leftarrow L \rightarrow b$ in the underlying model (thereby denoting potential causation);

3. a bidirected edge $a \longleftrightarrow b$, signifying some latent common cause $a \leftarrow L \rightarrow b$ in the underlying model (spurious association); and

4. an undirected edge a — b, standing for either $a \leftarrow b$ or $a \rightarrow b$ or $a \leftarrow L \rightarrow b$ in the underlying model.

Rule R_1 basically fixes the direction of an otherwise undirected edge avoiding the introduction of an additional v-structure (which would imply an additional independence). Rule R_2 fixes the direction of an otherwise undirected edge according to the requirement of acyclicity (which would be violated if the respective edge were oriented the other way).

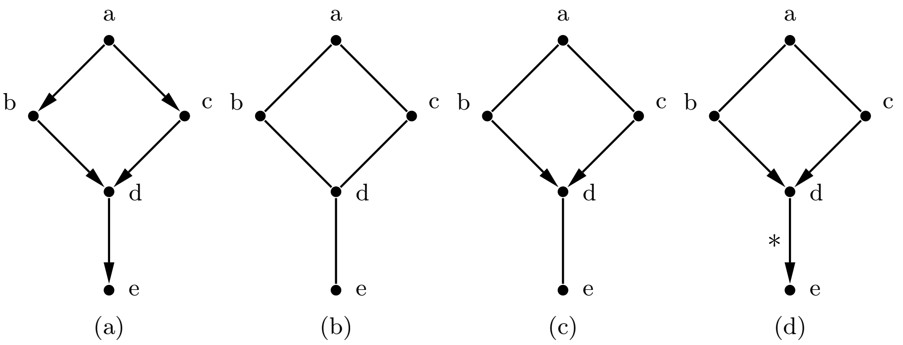

Fig. 2.8: Graph (a) displays the underlying actual structure encoded by a stable input distribution $\hat{P}$, (b) and (c) show the intermediate output of the IC* algorithm after working steps 1 and 2, finally resulting in (d) upon step 3.

Figure 2.8 illustrates the working steps of the algorithm. Consider (a) to be the underlying actual structure of our nonexperimental data $\hat{P}$. (b) displays the intermediate output of the algorithm after step 1, (c) the introduction of v-structures in step 2, and (d) the output of IC*, a marked pattern with inserted arrowheads and starred links.

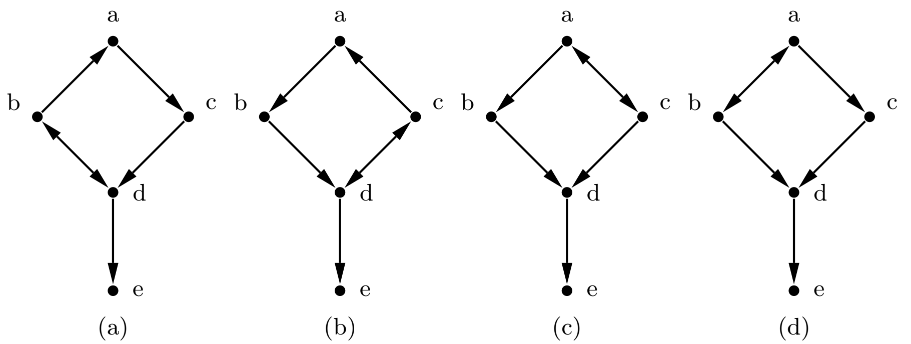

Fig. 2.9: The set of latent structures observationally equivalent to graph (2.8a), specifying the ambivalently marked edges of graph (2.8d).

Combinatorially spelling out the class of observationally equivalent latent structures amounts to drawing the four graphs depicted in figure 2.9. These graphs all have the same v-structures in common, namely the sole one collider node $b \rightarrow d \leftarrow c$. Also, the directed link $d \rightarrow e$ is present in each and every latent structure, introduced by virtue of rule R_1 which directs edges in one way if the other direction were to introduce additional v-structures into the graph (not derived from the independencies in $\hat{P}$).[55] The superstructure b — a — c is then spelled out in all four possible variants such that no new v-structure emerges.[56] Now, to be able to read the causal relation between two certain variables off the resulting graph, some suitably chosen *intervention variable Z* must serve as hypothetical control knob – just as with structural manipulations, with the only difference "that the variable Z, acting as a virtual control, must be identified within the data itself, as if Nature had performed the experiment."[57] Pearl states that in this respect the IC* algorithm even leads to the discovery of such control variables within the observational data. The notions of *potential cause* and *genuine cause* are accordingly defined with respect to such control variables and moreover depend on the avail-

[55] The introduction of additional v-structures would generate additional independencies that are not in the data but could be read off the graph.

[56] Note that the double-headed arrow combinatorially cycles the upper diamond once to produce the set of *markov-equivalent* structures in figure 2.9.

[57] Cf. [Pearl 2009, p. 54].

able *contexts* encoded in the probability distributions, i. e., on how two test variables X and Y behave if surrounding variables are conditioned on.

Definition 2.9.2 (Pearl's Potential Cause)[58]
A variable X has a potential causal influence *on another variable Y (that is inferable from $\hat{P}$) if the following conditions hold.*

1. *X and Y are dependent in every context.*
2. *There exists a variable Z and a context S such that*

   (i) X and Z are independent given S (i. e., $X \perp\!\!\!\perp Z \mid S$) and
   (ii) Z and Y are dependent given S (i. e., $Z \not\perp\!\!\!\perp Y \mid S$).

In the sense of definition 2.9.2, variable b can be identified as a potential cause of d in figure 2.8a if $Z = c$ represents the virtual control variable (with respect to d) conditional on the context $S = \{a\}$. In other words, the putative cause variable and the hypothetical control variable must be independent in some context – if a were missing altogether in the graph, it would even suffice to specify $S = \varnothing$. Nevertheless, b only qualifies as a *potential cause* of d, because – as displayed in figure 2.9 – the respective link between b and d can as well be realized as double-headed arrow, representing a possible common cause structure (e. g., as in figure 2.9a). Thus, b cannot be analyzed as being a genuine cause of d, i. e., as being linked to d by unambiguously directed edges, respectively.

Definition 2.9.3 (Pearl's Genuine Cause)[59]
A variable X has a genuine causal influence on another variable Y if there exists a variable Z such that either:

1. *X and Y are dependent in any context and there exists a context S satisfying*

   (i) Z is a potential cause of X,
   (ii) Z and Y are dependent given S (i. e., $Z \not\perp\!\!\!\perp Y \mid S$), and
   (iii) Z and Y are independent given $S \cup X$ (i. e., $Z \perp\!\!\!\perp Y \mid S \cup X$);

   or
2. *X and Y are in the transitive closure of the relation defined in criterion 1.*

[58] Cf. definition 2.7.1 in [Pearl 2009, p. 55].
[59] Cf. definition 2.7.2 in [Pearl 2009, p. 55].

In the sense of definition 2.9.3, variable d in figure 2.8a is analyzed as genuinely causally influencing e. This time, we can take as control variable either of b, c, both being potential causes of d. No variables are needed to fill the context: $S = \varnothing$. d and e are linked by a (if only one-link) chain of unambiguously directed edges. In a way, the test for the existence of some control variable Z is utilized in both definitions above to determine the direction of the "causal flow," just as what interventions as structural manipulations are devised for. In case two variables X and Y are dependent in *some context* (as a weakening of *potential cause*), but no direction of the causal relation can be made out by use of virtual controls, PEARL states one further definition: True *spurious association* can only be attributed to the existence of common causes.[60]

Now, in contrast with ERTEL's suggestion to pick out the graphs that fit our causal intuitions *before* metrically computing some input distribution $\hat{P}$, in PEARL's approach the selection of potential working candidates amongst possible structures is put off till after $\hat{P}$ has been processed by IC*. The output class of observationally equivalent latent structures might be evaluated against the backdrop of expert *causal* knowledge or basal everyday assumptions. Once the output class is restricted further and further, the marked pattern returned by IC* may be a helpful guide in designing possible test scenarios for the falsification of chosen arrow heads (through indication of potential further variables that could be taken into consideration, e. g., additional exogenous common causes).

2.10 About the identifiability of effects in causal models

If a set of observational data (e. g., in the form of a joint probability function) is available to the empirically based causal analyst who wants to evaluate the causal influence of one observed variable X on a distinct second, correlated variable Y, he is best advised to firstly exclude that the assessed correlation is actually induced by other factors than the potential causal influence to be examined. So-called *spurious correlation* between X and Y is generated by *confounding factors* Z influencing both X and Y at the same time, confounding our analysis, and ultimately biasing the estimate of the influence under consideration. This goes under 'confounding bias' in the respective literature.[61]

[60] Cf. definition 2.7.3 (Spurious Association) in [Pearl 2009, pp. 55 f.].

[61] Cf. for this and the following [Pearl 2009, pp. 182 f.].

As an example from econometrics, consider *Okun's law* which maps the relationship between unemployment and economic growth within national economy and postulates in compact manner linear dependence (which is one of the reasons for the law being so popular). Misinterpreting this dependence one might state: One necessary *requirement* for decreasing unemployment is strong economic growth. Critics of Okun's law point to the fact that long-term variances of other parameters equally important within national economy (as productivity, working time, job offers) tend to significantly confound the direct relationship between the examined quantities *unemployment* and *economic growth.* These additional parameters, however, do not occur in the formulation of the law. To speak of cause, effect, causal relation, or causal influence in either direction would certainly overstrain Okun's law.

In his book *Causality* (2000/2009) Pearl makes out why the concept of *confounding* goes largely unheeded in statistics course books:

> *As simple as this concept is, it has resisted formal treatment for decades, and for good reason: The very notions of "effect" and "influence" – relative to which "spurious association" must be defined – have resisted mathematical formulation. The empirical definition of effect as an association that* would *prevail in a controlled randomized experiment cannot easily be expressed in the standard language of probability theory, because that theory deals with static conditions and does not permit us to predict, even from a full specification of a population density function, what relationships would prevail if conditions were to change – say, from observational to controlled studies. Such predictions require extra information in the form of causal or counterfactual assumptions, which are not discernible from density functions [...]*[62]

Density functions, i. e., probabilistic descriptions, precisely talk about *closed* worlds and *fixed* environmental conditions, whereas in the framework of structural causal models the $do(\cdot)$-operator serves as an efficient tool for the virtual inspection of dependencies in *alternative* test scenarios.

Now, as soon as the researcher – merely on the basis of given non-experimental and purely observational data – sets out to model a certain situation, building a causal graph G and putting together the list of functional relationships renders it possible to examine, whether the causal

[62] Cf. [Pearl 2009, p. 183].

influence of one factor X on another one, Y, can be uniquely estimated, at all – always, of course, within the scope of his modeling. PEARL explicates the central idea behind this in his definition of *identifiability*:

Definition 2.10.1 (PEARL's Identifiability of Causal Effects)[63]
The causal effect of X on Y is said to be identifiable if the quantity $P(y \mid do(x))$ can be computed uniquely from any positive distribution of the observed variables that is compatible with G.

Especially when there exist latent variables in the model potentially causally influencing both X and Y at the same time (if only indirectly), a quantitative estimate within the model must be adjusted by means of other observed concomitant variables to exclude *confounding bias* and *spurious correlation.* How is a set of variables fit for this task to be found?

To efficiently accomplish the search for a suitable variable set, Judea PEARL formulates two criteria applicable again to the graph of a causal model. Making use of the so-called *back-door* and *front-door criterion* enables the researcher to easily identify from the diagram the set of nodes Z (of course representing the corresponding variables in the probability distribution compatible with G) with which confounding influences can be subtracted out. This *adjustment* is achieved by suitably summing up the potential values of all variables in Z. The two criteria shall be presented in due brevity in the following:

Definition 2.10.2 (PEARL's Back-Door Criterion)[64]
A set of nodes Z satisfies the back-door criterion relative to an ordered pair of nodes $\langle X_i, X_j \rangle$ in a DAG G if:

(i) no node in Z is a descendant of X_i; and

(ii) Z blocks every path between X_i and X_j that contains an arrow pointing towards X_i.

Analogously, Z satisfies the back-door criterion relative to two disjoint sets of variables $\langle X, Y \rangle$ if Z satisfies the back-door criterion relative to any pair $\langle X_i, X_j \rangle$ with $X_i \in X$ and $X_j \in Y$.

[63] Cf. definition 4 in [Pearl 1995, p. 674], slightly adjusted here to maintain consistent notation.

[64] Definition 3.3.1 in [Pearl 2009, p. 79].

Such a set Z accordingly d-separates all paths that would leave open a *back-door* into X_i for some possible confounding factor – hence the name of the criterion. In the miniature example given in the left graph of figure 2.10 the direct influence of variable X_i on variable X_j shall be assessed – obviously along the path $X_i \rightarrow X_6 \rightarrow X_j$. Employing the d-separation criterion, potential confounders outside the path $X_i \rightarrow X_6 \rightarrow X_j$ can be made out, i. e., variables that – within the causal diagram – influence both X_i and X_j simultaneously when wiggled qua modification. The back-door criterion now identifies the minimal sets $\{X_3, X_4\}$, or $\{X_4, X_5\}$ alternatively, as sufficient for screening off spurious influences. X_4 alone would not do the job, because although – according to the definition of the d-separation criterion – the path $X_i \leftarrow X_4 \rightarrow X_j$ would become blocked by conditioning on X_4, quite on the contrary X_4 opens the *flow of information* along the outer path via X_1 and X_2 as the collider node in this v-structure.

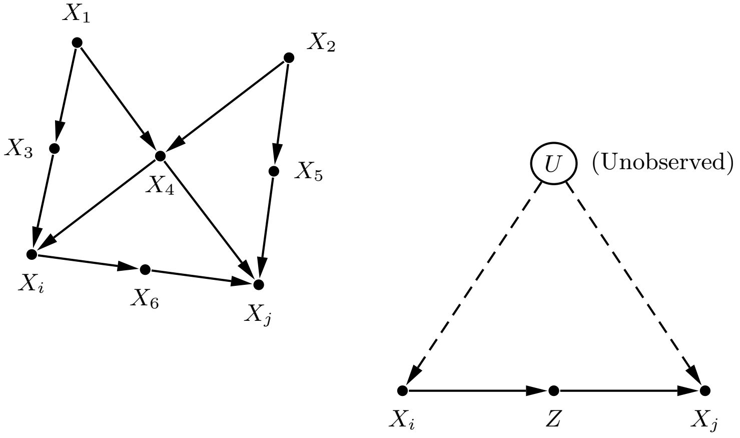

Fig. 2.10: In the left diagram the effect of X_i on X_j can be estimated consistently by means of adjusting for the variable pairs $\{X_3, X_4\}$ or $\{X_4, X_5\}$; the right diagram illustrates adjustment for Z by applying the front-door criterion.

Now, the *front-door criterion* takes care of those cases in which possible back-door paths run through *unobserved* variables, which are of course not apt for being a candidate set Z possibly screening off spurious correlation in computation – unobserved variables cannot be adjusted for. Pearl's graphical solution:

Definition 2.10.3 (Pearl's Front-Door Criterion)[65]
A set of nodes Z satisfies the front-door criterion relative to an ordered pair of nodes $\langle X_i, X_j \rangle$ in a DAG G if:

(i) Z blocks all directed paths from X_i to X_j;

(ii) there are no unblocked back-door paths from X_i to Z; and

(iii) all back-door paths from Z to X_j are blocked by X_i.

Here, conditions *(i)* through *(iii)* precisely indicate such sets of nodes mediating the (otherwise unconfounded) influence of some X_i on some X_j.

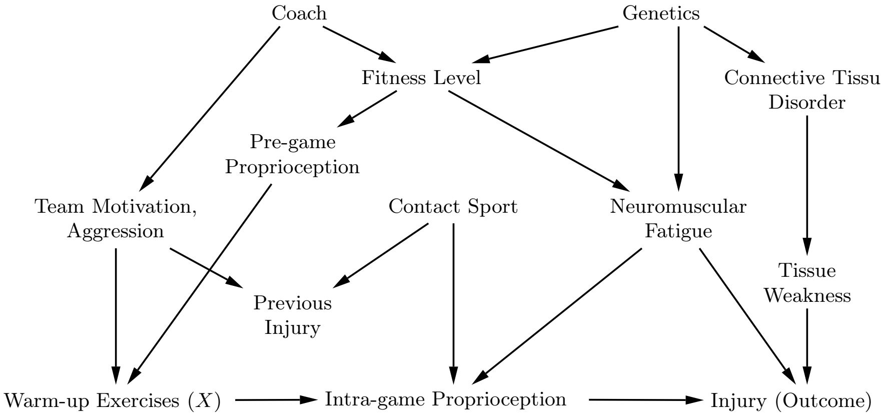

Fig. 2.11: A complex causal diagram illustrating the effect of warm-up exercises X on an athlete's susceptibility to injury Y (taken from [Shrier & Platt 2008, figure 2]).

A more complex example from medical practice, illustrated in figure 2.11, relates potential factors contributing to or preventing some athlete's susceptibility to injury while exercising the respective sport.[66] The effect of *warming up* before the game (represented by X) on the danger of *injury* (the outcome, Y) is to be tested. The mediating variable *intra-game proprioception* measures the athlete's balance and muscle control. In the upper part of the diagram the *coach* influences *team motivation* and *aggression* during the game which in turn makes an *earlier injury* more probable, just as participating in *warm-up exercises. Coach* and

[65] Definition 3.3.3 in [Pearl 2009, p. 82].

[66] Cf. for this and the following the presentation of this example case in [Shrier & Platt 2008].

genetic predisposition together contribute to the athlete's *fitness level*, and so forth. The question (in the center of the graph), if the respective game falls under the category of *contact sport* or not, also influences the probability of a *previous injury* independently of *team motivation.*

The influence of *warm-up exercises* on potential *injury* is obviously confounded by a multitude of factors. Application of the back-door criterion facilitates the search for a set of nodes Z which helps adjusting confounding factors: Those variables measuring *neuromuscular fatigue* and possible *tissue weakness* are jointly sufficient for screening off spurious influences because they intercept all back-door paths from X to the putative outcome Y. The path running through the question, if *contact sport* or not, is *inactive* without conditioning anyways, since it contains a collider node – an inverted fork. *Previous injury* is thus to be excluded from the adjusting set of variables Z, because gaining knowledge about *previous injuries* precisely opens a back-door again, thereby establishing indirect dependence between *warm-up exercises* and *susceptibility to injury.* Pearl's graphical criteria facilitate the identification of confounders and adjusting variables even in this rather complex example from medical practice.

Now, in case the effect of some variable on a second variable turns out to be identifiable in a given causal model, Pearl offers a set of rules, sound and complete, for the reduction of probabilistic expressions containing the $do(\cdot)$-operator to expressions without it. The so-called *do*-calculus enables the researcher to estimate post-intervention quantities merely from non-experimental, observational distributions. For the case that a set of variables screening off some causal flow from spurious influences can be made out – either by employing the back-door criterion or the front-door criterion – Pearl moreover presents two formulae for adjustment in [Pearl 1995], also restated in [Pearl 2009, pp. 79 ff.]. The following two respective theorems shall be given for the sake of completeness and conclude this section before the concept of *token causation* will be examined more closely below:

Theorem 2.10.4 (Pearl's Back-Door Adjustment)[67]
If a set of variables Z satisfies the back-door criterion relative to $\langle X, Y\rangle$, then the causal effect of X on Y is identifiable and is given by the formula

$$P(y \mid do(x)) = \sum_{z} P(y \mid x, z) P(z). \tag{2.18}$$

[67] Theorem 3.3.2 in [Pearl 2009, pp. 79 f.].

Theorem 2.10.5 (Pearl's Front-Door Adjustment)[68]
If a set of variables Z satisfies the front-door criterion relative to $\langle X, Y \rangle$ and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula

$$P(y \,|\, do(x)) = \sum_z P(z \,|\, x) \sum_{x'} P(y \,|\, x', z) P(x'). \tag{2.19}$$

2.11 Singular causation and the actual cause

To let the protagonists of this discussion appear once more on stage together, or to at least present their approaches in a certain concord, we take up the notion of singular causation as discussed by Pearl in *Causality*. So we are not talking about generic analysis as in "rain causes the street to get wet," but deal with situations on token level. As an example we look at the administering of some medication in a controlled clinical study.[69] When evaluating the collected data statically, it turns out that the medication neither affects the recovery of the patients positively nor negatively – on average. Furthermore, we pick out one specific patient who was administered the medication and fully recovered subsequently. Now, a causal model ought to be able to answer in particular, whether the patient's recovery occurred *due to* the treatment, *despite* the treatment, or *completely regardlessly of* the treatment. An answer cannot be given if only observational data is available, since the patient was never tested *without* administering the medication – probabilities under deviant conditions cannot be compared.

Pearl points out clearly that in such cases on token level the solution lies solely in the *counterfactual* analysis of the problem: "What would the probability of recovery have been if the medication had not been administered?" This counterfactual formulation precisely implies the deviation from observed data. The desired value can be calculated from an *alternative* model in which the counterfactual antecedent is virtually forced to be true. This very task is achieved by the $do(\cdot)$-operator in structural causal models.

To give an example: We want to learn if the intervention $\neg p$ would have brought about $\neg q$ if in fact we already know that both p as well as q have

[68] Theorem 3.3.4 in [Pearl 2009, p. 83].

[69] This example is taken from [Pearl 2009, pp. 33 f.].

occurred factually. The effect of an intervention can indeed be identified *contra* and moreover *post factum* in the methodological triple jump along the lines of natural reasoning:[70]

1. **Abduction:** Having observed p and q, i. e., our factual evidence, we infer the explanation (i. e., a hypothesis or a state of the world) going backwards (or upwards) in the network.

2. **Action:** In the context of this fixed environment we perform the local surgery $\neg p$ by intervening in the structural model (thereby transforming the graph).

3. **Prediction:** The autonomous mechanisms of the structural causal model now allow for predicting the value assignment of Q (if the causal effect of P on Q is identifiable, at all).

Performing steps 1 through 3 basically amounts to traversing the causal network twice – once *upwards* to determine the values of all exogenous variables and then *downwards* again after transforming the network by clipping all arrows pointing towards P (thereby lifting P from the influence of its direct causes). Letting the context information propagate through the network again, we are ultimately able to read off the graph the counterfactually predicted value of Q.[71]

The scheme above unifies the essential features of Pearl's agenda once more, which are conceptually and methodologically sorted out again by Christopher Hitchcock, who writes in his article *Causal Modelling* in the *Oxford Handbook of Causation*:

> *There is an important pragmatic difference between counterfactuals and interventions: we are typically interested in knowing the truth values of counterfactuals after the fact, whereas we are usually interested in evaluating the consequences of potential interventions before they are carried out.*[72]

Interventions by use of the $do(\cdot)$-operator precisely allow for such hypothetical tests in virtual, alternative experimental designs. But, if the

[70] Cf. for this and the following [Pearl 2009, p. 37 and pp. 205 ff.].

[71] This procedure is presented by Pearl in *twin networks*, where going upwards and (after transforming the model) going downwards through the same graph is interpreted as stepping sideways into the almost identical copy of the original graph where all edges directed into Q are clipped.

[72] Cf. [Hitchcock 2009a, p. 303].

do(·)-operator, in externally bringing about Lewis' "small miracles," stands in close methodological vicinity to Lewis' similarity relation, and if Pearl accuses Lewis of circularly taking some immanent notion of causation for granted when comparing two worlds with respect to their similarity to a third, must Pearl not defend himself against his own allegations, as well? – Does he not also analyze complex causal relationships by evaluating basal natural assumptions, which themselves convey causal meaning, and do so necessarily, as Pearl stresses? I would like to give the floor to Pearl himself, who formulates the following in the 2005 paper *Causes and Explanations* together with collaborator Joseph Halpern:

> *It may seem strange that we are trying to understand causality using causal models, which clearly already encode causal relationships. Our reasoning is not circular. Our aim is not to reduce causation to noncausal concepts but to interpret questions about causes of specific events in fully specified scenarios in terms of generic causal knowledge such as what we obtain from the equations of physics. The causal models encode background knowledge about the tendency of certain event types to cause other event types (such as the fact that lightning can cause forest fires). We use the models to determine the causes of single (or token) events, such as whether it was arson that caused the fire of 10 June 2000, given what is known or assumed about that particular fire.*[73]

Grounding causal analysis in such basal causal assumptions can in the thus explicated sense rightfully be called fruitful because it admits utilizing the introduced notions and methods constructively. Certainly, emphasizing truly *deterministic* causal mechanisms much rather brings acuteness into causal analysis than drawing the researcher over the notional fringes of naughtiness.

Utilizing our understanding of the mechanisms at work to learn about hypothetical alternative situations is one direction of posing questions about token causation – we might also look the other way and employ the specification of a causal model to learn which of the actually observed occurrences is the true cause of another distinct observed event. In other words: How can we find out, if some realized putative cause candidate is indeed responsible for having brought about some observed other event, maybe even post factum? To be able to deal with this sort of token (or singular) causal reasoning, Pearl advances in two steps: Firstly, our

[73] Cf. [Halpern & Pearl 2005a, p. 849].

understanding of the situation, i. e., the parameterization of the model (or the specification of the functional deterministic mechanisms), is coupled with some fixed context – the given (vector of) circumstances $\vec{u}$. Secondly, by ultimately quantifying over possible value assignments to subsets of the observed variables in V an *active causal process* is carved out, thereby marking the minimal set of contributory causes $\vec{X}$ of the token event $Y = y$ under consideration, in a sense *in virtual comparison* with different configurations of the model – different qua intervention.

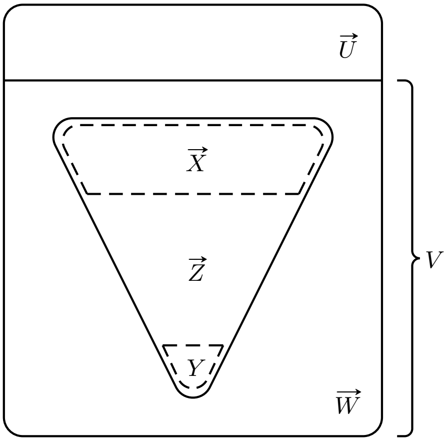

Fig. 2.12: The idea behind Pearl's definition of the actual cause, illustrated as relations between subnets of the causal model M. $\vec{U} = \vec{u}$ indicates the obtaining circumstances, the observed variables are partitioned such that $V = \vec{W} \cup \vec{Z}$ with $\vec{Z}$ signifying the *active causal process* relating $\vec{X} = \vec{x}$ and $Y = y$, where $\vec{X} \cup \{Y\} \subseteq \vec{Z}$.

Figure 2.12 illustrates the idea behind Pearl's definition of the *actual cause* (definition 2.11.1 below), relating subnets of the causal model in Venn-like presentation: The variables in the causal model M are assigned their values according to some specification of the context $\vec{U} = \vec{u}$. The tuple $\langle M, \vec{u} \rangle$ especially induces the unique assignment of values to the n-ary vector $\vec{X}$, the *total cause*, aggregating all variables $X_1, \ldots, X_n$ as *contributory causes* of the event $Y = y$. Finally, the vector $\vec{Z}$ (assuming the corresponding observed values $\vec{z^*}$) can be understood as the *active causal process* in M under the obtaining circumstances $\vec{u}$. It turns out that the definition of $\vec{Z}$ picks out (in addition to $\vec{X}$ and Y) just the variables mediating between $\vec{X}$ and Y on the paths directed from the variables in $\vec{X}$ to Y.[74]

[74] Cf. Pearl's remarks in [Halpern & Pearl 2005a, p. 854].

Definition 2.11.1 (PEARL's Actual Cause)[75]
$\overrightarrow{X} = \overrightarrow{x}$ is an actual cause *of $Y = y$ in $\langle M, \overrightarrow{u} \rangle$ if the following three conditions hold:*

1. *$\langle M, \overrightarrow{u} \rangle \models (\overrightarrow{X} = \overrightarrow{x}) \wedge (Y = y)$, that is, both $\overrightarrow{X} = \overrightarrow{x}$ and $Y = y$ are true in the actual world.*

2. *There exists a partition $\{\overrightarrow{Z}, \overrightarrow{W}\}$ of V with $\overrightarrow{X} \subseteq \overrightarrow{Z}$ and some setting $\overrightarrow{x'}$ and $\overrightarrow{w'}$ of the variables in $\overrightarrow{X}$ and $\overrightarrow{W}$ such that if the fixed context $\overrightarrow{u}$ induces the value assignment $\overrightarrow{z^*}$ for the variables in $\overrightarrow{Z}$, i. e., if $\langle M, \overrightarrow{u} \rangle \models (\overrightarrow{Z} = \overrightarrow{z^*})$, then both of the following conditions hold:*

   (a) The intervention $\langle \overrightarrow{X}, \overrightarrow{W} \rangle \leftarrow \langle \overrightarrow{x'}, \overrightarrow{w'} \rangle$ (assigning non-actual values to $\overrightarrow{X}$ and $\overrightarrow{W}$) changes $Y = y$ from true to false in the model, i. e., $\langle M, \overrightarrow{u} \rangle \not\models (Y = y)$ (in the accordingly transformed model).[76]

   (b) $\langle M, \overrightarrow{u} \rangle \models (Y = y)$ under any possible interventional assignment $\langle \overrightarrow{W'}, \overrightarrow{Z'} \rangle \leftarrow \langle \overrightarrow{w'}, \overrightarrow{z^*} \rangle$ for all subsets $\overrightarrow{W'}$ of $\overrightarrow{W}$ and all subsets $\overrightarrow{Z'}$ of $\overrightarrow{Z}$, as long as $\overrightarrow{X}$ is kept at its current value $\overrightarrow{x}$, i. e., $\overrightarrow{X} \leftarrow \overrightarrow{x}$; in other words, setting any subset of variables in $\overrightarrow{W}$ to their values in $\overrightarrow{w'}$ should have no effect on Y, as long as $\overrightarrow{X}$ is kept at its current value $\overrightarrow{x}$, even if all the variables in an arbitrary subset of $\overrightarrow{Z}$ are set to their original values in the context $\overrightarrow{u}$.

3. *$\overrightarrow{X}$ is minimal, i. e., no subset of $\overrightarrow{X}$ satisfies conditions* 1 *and* 2.[77]

In the definition of the actual cause, condition 1 simply says that we are dealing with token causation – cause and effect actually occur in the causal model M (with a full specification of the functional mechanisms) if the context $\overrightarrow{u}$ is given. Section 2 of the definition postulates a partitioning of the observed variables V into those variables ($\overrightarrow{Z}$) involved in the causal process connecting $\overrightarrow{X}$ to Y and those variables ($\overrightarrow{W}$) irrelevant

[75] Cf. PEARL's presentation in [Halpern & Pearl 2005a, p. 853] (definition 3.1 and refinements in sect. 5) and SPOHN's discussion thereof in [Spohn 2010, sect. 2]; the formulation given here differs slightly from PEARL's to maintain consistency with previously introduced conventions.

[76] The notation $\ulcorner \overrightarrow{X} \leftarrow \overrightarrow{x} \urcorner$ symbolizes the transformation of the model by iterative application of the $do(\cdot)$-operator (as introduced above) for each variable in the n-ary vector $\overrightarrow{X}$, thereby pruning the respective graph of the model and modifying all structural equations in accordance with $do(X_1 = x_1), do(X_2 = x_2), \ldots, do(X_n = x_n)$.

[77] As PEARL adds: "Minimality ensures that only those elements of the conjunction $\overrightarrow{X} = \overrightarrow{x}$ that are essential for changing $Y = y$ in $2(a)$ are considered part of the cause; inessential elements are pruned." – cf. [Halpern & Pearl 2005a, p. 853].

for the observed effect Y. Condition $2(a)$ says that changing the values of $\overrightarrow{X}$ also renders the event $Y = y$ inactive. Condition $2(a)$ also allows for changing the values of $\overrightarrow{W}$ (the variables seen as irrelevant for bringing about $Y = y$), since they might *sustain* the effect, thereby masking the influence of $\overrightarrow{X}$ on Y which we are after. On the other hand, it is to be excluded that only the assignment of $\overrightarrow{w}$ to $\overrightarrow{W}$ is responsible for the event $Y = y$, so condition $2(b)$ tightens things again by postulating that only the actual observed values of $\overrightarrow{X}$ succeed in bringing about the effect $Y = y$: Arbitrary subsets of the *irrelevant* variables in $\overrightarrow{W}$ may assume arbitrary non-actual values $\overrightarrow{w'}$. $Y = y$ will still be observed if only $\overrightarrow{X}$ is set to its actual observed value $\overrightarrow{x}$. Pearl points out that setting $\overrightarrow{X}$ to $\overrightarrow{x}$ necessarily entails the assignment $\overrightarrow{Z} = \overrightarrow{z^*}$, since $\overrightarrow{Z}$ contains the very variables mediating between $\overrightarrow{X}$ and Y – the *active causal process.*

It is important to note one thing here: If our causal model at hand is *narrow* in the sense that it only consists of the active causal process alone, the analysis is done without hassle – no ambiguities arise. For example, if the model is a simple chain, then the (total) cause of some event $Y = y$ is simply the variable X corresponding to the single parent node of Y, which is minimal in the sense of condition 3 above. If, nevertheless, the model is enriched arbitrarily by adding supposedly irrelevant information (about potential causal relations), things might get muddled. This especially shows in cases of causation by omission where some effect is produced by some specific event *not* being realized, and where it is this specific event that the researcher wants to single out and test for potential causation and not any other potential cause candidates that did not get realized either. Consider the following example, taken from [Halpern & Pearl 2005a, p. 871 (example 5.3)] (referring to an unpublished text by Hall and Paul):

> *Suppose Suzy goes away on vacation, leaving her favorite plant in the hands of Billy, who has promised to water it. Billy fails to do so. The plant dies – but would not have, had Billy watered it. . . . Billy's failure to water the plant caused its death. But Vladimir Putin also failed to water Suzy's plant. And, had he done so, it would not have died. Why do we also not count his omission as a cause of the plant's death?*

Halpern and Pearl make two suggestions: (i) Slim down the model and do away with the endogenous variable *Putin.waters.the.plant.* This would leave the analyst with only Billy's failure as potential cause candidate. (ii) If the richer structure is to be preserved, do not check all variable assignments for the putatively irrelevant variable set $\overrightarrow{W}$. Formally this amounts to defining *extended causal models* by adding sets of *allowable settings* $\mathcal{E}$ of the endogenous variables. In the case of the above example, our modeling intuition tells us that the variable setting $\mathit{Putin.waters.the.plant} = \mathit{1}$ is to be excluded from the respective set of allowable settings $\mathcal{E}$. That is, we are not considering Putin's failure to water the plant as true cause, as relevant for the situation, as informative when mentioned in discussion, etc. Both strategies are of pragmatic nature but necessary if Pearl's causal analysis is to be implemented especially in contexts where (neglected) responsibility is to be determined or guilt is to be assessed. It seems as if Pearl were loosening his own rigorous framework at this point again. But he answers:

> *Are we giving ourselves too much flexibility here? We believe not. It is up to a modeler to defend her choice of model. A model which does not allow us to consider Putin watering the plant can be defended in the obvious way: that it is a scenario too ridiculous to consider.*[78]

Ultimately – and summing up Pearl's explications – this means that the task of discerning causes (especially singular ones for explanatory purposes) comes down to enriching bare structures by *adding non-causal knowledge* and to *model-relatively* querying *lower-level basal relations* upon limiting possible settings (i. e., upon marking worlds to consider). These facets will be picked up again, exploited, and expanded further in chapter 3 now.

[78] Cf. [Halpern & Pearl 2005a, p. 871].

Chapter 3

Causality as epistemic principle of knowledge organization

as [resemblance, contiguity, causation] are the only ties of our thoughts, they are really to us the cement of the universe

David HUME, *An Abstract of a "Treatise of Human Nature"*

When Judea PEARL takes a stand on the ontic (metaphysical) status of causation in his book *Causality*, he clearly localizes causal relationships on the objective (physical) side of the pair *ontological versus epistemic (doxastic)*:[1]

> *[...] causal relationships are ontological, describing objective physical constraints in our world, whereas probabilistic relationships are epistemic, reflecting what we know or believe about the world. Therefore, causal relationships should remain unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes changes.*[2]

Nevertheless, as becomes obvious from his comments on possible, allowable settings of the variables in the causal model at hand (or on ways of

[1] Compare chapter 1, sect. 1.9, for a list of decisions to take in the process of devising a theory of causation.

[2] Cf. [Pearl 2009, p. 25].

slimming down causal models), it is the modeler's choice which comes before the assessment of causes and their effects (*within* the chosen frame): Causal inference remains model-relative in PEARL's framework, after all. Still, the holistic and monolithic path followed by an account of causation on the basis of hypothetical, possible interventions seems (i) to be very attractive to various (at least empirical) disciplines that build upon the practice of experimentation and (ii) to correspond very well with our intuitions, summed up in the observation – or rather in the *postulate* – "wiggling the cause affects the effect – and not vice versa." But on what ontological grounds can this claim be understood and made exploitable fruitfully, if a purely physically objective interpretation faces criticism?

3.1 The total system and the modality of interventions

One way of answering the question, where causation is to be localized metaphysically, is to simply deny the existence of such a relation, as Bertrand RUSSELL does in his often-quoted seminal inquiry *On the Notion of Cause* (1913):

> *[. . .] the reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.*[3]

Examining what we mean by saying that one event causes a second event, RUSSELL finds that mature science has withdrawn from the business of marking off suitable events and has rather settled on specifying variables and their corresponding measurement methods for the analysis of functional dependencies instead of causal relations.[4] RUSSELL must arrive at his concluding judgment – he bases his argument on the existence of actual events and physical processes. The analysis of everyday language

[3] Cf. [Russell 1913, p. 1].

[4] Cf. for this and the following [Dowe 2009, pp 214 f.]. There (p. 215), Phil DOWE briefly comments on this finding:

> *Russell simply makes the point that [science] focuses on functional relations between variables, a focus far removed from the kind of commonsense events that we take to be causes and effects.*

Quite contrary to that, PEARL in his framework introduces events as *random variables* and causal relations as *deterministic functional relations* between those variables, aiming at a close connection with intuitive reasoning.

with *widely defined* events (stated by expressions such as 'some person's hitting the billiard ball') ultimately yields the result that there are no real regular causes, since any event can be intercepted by way of exception and rendered void in its bringing about the effect (in this case, the billiard ball falling in some pocket) – there can always be *unmentioned* further preventative circumstances complicating our attribution of causal efficacy. On the other hand, going *narrow* trivially eliminates the concept of cause, too, as Russell argues. As soon as any cause-effect pair of events is described in all its pertaining details, saying that one event causes the other becomes trivially true, since this case of regularity only has one instance. Talk of causation can thus be exchanged for talk of determination without losing any inherent meaning.

Now, if one wants to stick to the Bayes net account of deterministic causation with interventions (going from types to tokens), there is no other option than to agree with Lord Russell: Any causal model in Pearl's sense relies on an agreement about the context, i. e., on the choice of variables that are to be excluded from modeling as *ceteris paribus conditions.* Violation of one of these silent (in principle infinitely many) conditions may render the whole model useless. As a consequence, no causal model thus construed will suffice to hold for all thinkable cases of application – for any causal model there is always a model closer to objective physical reality (by outward augmentation or zooming in).[5] On the other hand, strengthening the narrow description of events would ultimately result in infinitely narrow, point-like instantiations of certain features (following Kim's explication) such that the full Bayes net thus conceived would become infinitely dense, ultimately losing its property of being a network altogether, thus not being modifiable by structural local surgeries in Pearl's sense for testing causal directionality anymore.[6] J. L. Mackie arrives at a similar conclusion in his examination of causal priority.[7] Basing his argumentation on the ontological interpretation of causation he admits that, if one assumes total determinism, there is no way to find out about the direction of causation anymore – any specific event is apt for the explanation (or prediction) of any other event, since

[5] To avoid entering any vicious circle here, *closer* simply means *taking more events into account* – in a way, *fine-graining* the model.

[6] This remark basically combines both points made by Russell: Going narrow and storing *all potentially relevant* information in the model amounts to storing *all* information in the model (in dense arrangement). Consequently, what Nancy Cartwright in [Cartwright 2001, sect. 3a] calls "God's big Bayes net" in reference to what Wolfgang Spohn in [Spohn 2000, p. 11] calls the "all-embracive Bayesian net" cannot be a *net structure* anymore.

[7] Cf. for this and the following [Mackie 1980, pp. 189 ff.].

any two events in this system (or realizations of two distinct variables, respectively) are related in strict manner:

> *If you have too much causation, it destroys one of its own most characteristic features. Every event is equally fixed from eternity with every other, and there is no room left for any preferred direction of causing.*[8]

Moreover, assuming total determinism in the total system (along with LAPLACE, EINSTEIN, or modern deterministic interpretations of quantum mechanics) finally makes PEARL's interventionist account of causation inapplicable altogether. If there is nothing external to the system anymore, even the possibility of thinking *external interventions* becomes inaccessible if those manipulations are to have any meaning at all (perhaps some counterpart in some possible world – to stay in the laboratory picture; but alternative possible worlds would have to be *completely incompatible and incomparable* with the actual world at all times, since LEWIS' "small miracles" are excluded if *total* determinism is postulated).

Does the hunt for causes stop here? RUSSELL judges that the law of causality is only *erroneously* supposed to do no harm – he must admit, after all, that causal talk obviously does convey meaning. Pursuing to carve out the unifying content might thus still turn out to be fruitful in the end.

3.2 Subnets for epistemic subjects

When Wolfgang SPOHN ponders the metaphysical status of causal dependence, he finally arrives at the conclusion (also heading his paper from 2000) that "Bayesian Nets Are All There Is To Causal Dependence":

> *So far, I have only introduced two distinct graph-theoretical representations: one of causal dependence between variables and one of conditional probabilistic dependence. However, the core observation of each probabilistic theory of causation is that there is a close*

[8] Cf. [Mackie 1980, p. 191]. Note that MACKIE puts forward a variant of an interventionist (or at least agentive) account of causation. He makes this explicit with his first rendition of *causal priority* in [Mackie 1980, p. 190]:

> *A first approximation to an analysis of 'X was causally prior to Y', where X and Y are individual events, is 'It would have been possible for an agent to prevent Y by (directly or indirectly) preventing, or failing to bring about or to do, X' [...].*

> *connection between causal and probabilistic dependence, that the two representations indeed coincide, i.e. that each causal graph is a Bayesian net. Thereby, the Markov and the minimality condition turn into the causal Markov and the causal minimality condition.*[9]

As in Pearl's framework sketched above, some variable's Markovian parents can consequently be interpreted as the direct causes of the event represented by this variable. Spohn goes on by exploiting this:

> *[D]irect causal dependence is obviously frame-relative [...]. The relativization would be acceptable, if it concerned only the direct/ indirect distinction: what appears to be a direct causal dependency within a coarse-grained frame may well unfold into a longer causal chain within a more fine-grained frame. [...] It's worse, however. The whole notion of causal dependence is frame-relative according to [its] definition: where there appears to be a direct or an indirect causal dependency within a coarse-grained frame, there may be none within a more fine-grained frame, and vice versa.*[10]

An answer to the question what guidelines we have for building net structures of causal models that truly reflect 'real causation' might not be feasible, after all, Spohn concedes:

> *In the final analysis it is the all-embracive Bayesian net representing the whole of reality which decides about how the causal dependencies actually are. Of course, we are bound to have only a partial group of this all-embracive Bayesian net.*[11]

If this seems to be a necessary shortcoming of representing causal dependencies in suitably confined Bayesian nets, why not understand it as an essential feature of causal reasoning in the first place? The interventionist account of causation even requires the net structures of causal models to correspond to *subsystems* of the total system – expert knowledge or common sense tells us how to carve out sufficiently open and at the same time sufficiently closed subnets of our (directly or indirectly) perceived surroundings. *Open* to allow for hypothetical external interventions and *closed* to mark off the variables under consideration from those assigned merely exceptional influence as in the laboratory picture.[12] Deciding upon the set of variables considered illuminating for

[9] Cf. [Spohn 2000, p. 5].

[10] Cf. [Spohn 2000, p. 7].

[11] Cf. [Spohn 2000, p. 11]. Also see footnote 6 of this chapter for a critical remark on the term 'all-embracive Bayesian net.'

[12] Judea Pearl comments on the origins and ramifications of breaking the 'closed world assumption' in greater depth in sect. 7.5 of [Pearl 2009].

the analysis to be conducted is obviously a subjective (sometimes highly pragmatic) process that may differ from one epistemic agent to the next even if performed in compliance with rational standards. The structure of each agent's causal graph largely depends upon prior knowledge and intensional aspects to be emphasized. Interpreting the arrows in such causal graphs as conveying causal meaning ultimately amounts to accepting an *epistemic* account of causation, the proponents of which analyze causality "neither in terms of physical probabilities nor in terms of physical mechanisms, but in terms of an agent's epistemic state", as Jon WILLIAMSON summarizes in [Williamson 2009, p. 204]. WILLIAMSON marks the core of such an epistemic theory of causation:

> *[T]he proponent of the epistemic theory holds that ['A causes B'] says something about rational belief.*[13]

One of the salient advantages of the epistemic approach is the straightforward justification of causal talk across different levels (macro–meso–micro) and various domains of discourse, as WILLIAMSON continues:

> *Heterogeneity of mechanisms across the sciences is no problem because the causal relation is not analysed in terms of those [discipline-specific] mechanisms but in terms of rational belief, an account that is not specific to particular sciences.*[14]

Interventions remain relative to the formulation of the mechanisms under consideration as in PEARL's original conception. But interpreting the structure represented in a causal model as informant about an agent's epistemic state now allows for reconsidering the key concepts of a causal theory (now formulated in epistemic terms):

- *Causation* can be understood as an epistemic relation between representations of real events.
- *Causality* becomes a principle of organizing knowledge efficiently for explanation, prediction, and instruction.

Following these suggestions, the Bayesian networks at the core of each causal model need not be interpreted anymore as representing interconnected laws of physics (on some macro level) but may be re-interpreted as storing relational knowledge, which in turn enables us to augment causal models to *generic* structures of learning and communication targeted at consistent causal inference.

[13] Cf. [Williamson 2009, p. 206] where Jon WILLIAMSON refers back to Ernst MACH, who writes in *The Science of Mechanics* (1883) that "Cause and effect [...] are things of thought, having an economical office" (p. 485).

[14] Cf. [Williamson 2009, p. 206].

3.3 Organizing Data in causal knowledge patterns

One of the criticisms the purely probabilistic account of causation has to acknowledge is the charge of reducing causal relations to probabilistic ones where there may be no grounds for this simplistic transition, as summed up by Williamson:

> *[P]robabilistic dependencies may be attributable to other kinds of relationships between the variables. A and B may be dependent not because they are causally related but because they are related logically (e.g. where an assignment to A is logically complex and logically implies an assignment to B), mathematically (e. g. mean and variance variables for the same quantity are connected by a mathematical equation), or semantically (e. g. A and B are synonymous or overlap in meaning), or are related by non-causal physical laws or by domain constraints. [...] To take a simple example, if a logically implies b then* $P(b\,|\,a) = 1$ *while* $P(b)$ *may well be less than* 1*. In such a case variables A and B (where A takes assignments a and* $\neg a$ *and B takes assignments b and* $\neg b$*) are probabilistically dependent; however it is rarely plausible to say that A causes B or vice versa, or that they have a common cause.*[15]

The embedding of these relations has not been available so far, since the arrows in the graphical portion of Pearl's causal models had to be interpreted causally, the network structure had to thoroughly obey the causal Markov condition, and all events represented by random variables had to be sufficiently distinct to allow for causal inference at all. Nevertheless, I want to argue here that the way we infer causal knowledge from more basic assumptions relies to a large extent also on *non-causal* knowledge (of the sort referred to by Williamson above), which quite substantially helps arranging and connecting subnet structures of actual causal purport. Amongst the most important relations serving this purpose are node connections representing deterministic, non-directional knowledge, i. e., links that strictly correlate certain variables and along which information may be transferred instantaneously. These find no place in the Bayes net causal models defined above (especially if those are understood as sub-portions of the all-embracive net built on physical laws), but they can be introduced in the epistemically interpreted variant of those very causal models as carefully restricted augmentations. Of course, a new type of edge is necessary for representing this idea, since directed edges are already reserved for directional, asymmetrical

[15] Cf. [Williamson 2009, p. 200].

causal knowledge. So-called *epistemic contours (ECs)* shall enrich the graphical part of PEARL's causal models – however, integrating these *epistemic contours* into Bayes net causal graphs turns these graphs into semi-DAGs with undirected subnets, so-called *EC cliques*:

Definition 3.3.1 (EC Clique)
An EC clique *is a subnet in a semi-DAG (of a* causal knowledge pattern *as defined below) that is exclusively connected by undirected edges (representing epistemic contours). EC cliques are defined as* transitively closed *under the EC relation.*

This new kind of edge bars causal inference in the above Bayes net framework. The desideratum remains, namely the unification of causal and non-causal knowledge in structures that allow consistent computation of causal claims. This leads to the formulation of *causal knowledge patterns (CKPs)* targeted at facilitating the prediction of future events, the explanation of past events, and the choice of suitable actions for efficient achievement of intended goals on the basis of causal *and* non-causal data. Gaps between levels of abstraction or even between different disciplines can be bridged by making knowledge explicit in CKPs:

Definition 3.3.2 (Causal Knowledge Pattern)
A causal knowledge pattern *is a quadruple*

$$\mathcal{K} = \langle U, V, F, C \rangle$$

such that $M = \langle U, V, F \rangle$ is a causal model[16] *where*

(i) U is a set background *variables (*exogenous *variables), that are set from outside the model;*

(ii) V is a set $\{V_1, V_2, \dots, V_n\}$ of n endogenous *variables, that are determined by variables in the model – i. e., by variables in $U \cup V$;*

(iii) F is a set of causal mechanisms *for V, i. e., n functions $\{f_1, f_2, \dots, f_n\}$ (determining the value of each variable V_i in V) such that*

$$v_i = f_i(\mathbf{pa}_i, u_i)^{17}$$

with

$$f_i : \left(\prod_k Ran(\mathrm{PA}_{i_k})\right) \times Ran(U_i) \to Ran(V_i) \qquad (1 \le i \le |V|),$$

[16] Also see definition 2.7.1 of *causal model* on p. 56.

[17] The boldface $\mathbf{pa}_i$ collects for each variable V_i the values of its parent variables in a list of length $|\mathrm{PA}_i|$, i. e., $\mathbf{pa}_i := pa_1, pa_2, \dots, pa_{|\mathrm{PA}_i|}$.

where for every i: $1 \leq k \leq |\mathrm{PA}_i|$, $U_i \in U$ *(possibly combining multiple contributing and/or preventative disturbance factors into one complex variable), and* $\mathrm{PA}_i \subseteq V \backslash \{V_i\}$.[18]

(iv) C *is a set of* epistemic contours*, i. e., a set of 1-1 functions* $c_{i,j}$ *such that*

1. $1 \leq i, j \leq n \ \wedge \ i \neq j$,
2. $c_{i,j} : \mathit{Ran}(V_i) \longrightarrow \mathit{Ran}(V_j)$,
3. $\forall i, j, k(c_{i,j} \in C \wedge c_{j,k} \in C \Rightarrow c_{i,k} \in C)$,
4. $\forall i, j(c_{i,j} \in C \Rightarrow c_{j,i} \in C)$, *and*
5. $c_{j,i} = c_{i,j}^{-1}$.

Clause 3 says that the set of functions C *is* euclidean*, while 4 and 5 define* C *to be* closed under inversion.[19]

(v) A variable X *being connected to a second variable* Y *by an epistemic contour but possessing no causal mechanisms (i. e., influent arrows in the semi-DAG) is treated like an endogenous variable (since its value is determined by the value of* Y*) with one exception: If no variable in* X*'s EC clique receives its value through a causal mechanism (but only via epistemic contours), all variables in* X*'s EC clique are said to be* simultaneously exogenous*. One single variable in this EC clique receiving its value from outside conditions suffices to determine the values of all other variables in the EC clique.*

The graph $D_{\mathcal{K}}$ of a causal knowledge pattern $\mathcal{K}$ can be understood as an augmentation of the graph D_M pertaining to the causal model $M = \langle U, V, F \rangle$, which in turn is a sub-structure of $\mathcal{K}$. The set C of *deterministic epistemic contours* is represented in the graph $D_{\mathcal{K}}$ as undirected edges: The pair of contours $\{c_{i,j}, c_{j,i}\}$ is graphically rendered as the undirected edge connecting the node with the label V_i to the node with the label V_j. Such an undirected edge will in the following also simply be called *contour* – although it actually represents a *pair* of underlying inter-definable functions – since the edge in the graph symmetrically represents both corresponding functions, and context always disambiguates what is formally referred to.

[18] Also see the definition of *causal mechanisms*, 2.7.2, on p. 57.

[19] Clauses 3–5 of def. 3.3.2 are listed here w. l. o. g., since all $c \in C$ are 1-1 functions; especially 3 can be loosened to much rather express the *potential* expansion of C.

Now, epistemic contours thus defined satisfy the very desiderata listed above. They represent non-directional knowledge, thereby being capable of bridging different frameworks of description (maybe vertically on different levels or horizontally in different disciplines). Epistemic contours deterministically transfer knowledge by virtue of their definition as bijective functions – in a way marking variables that cannot be decoupled, i. e., variables that cannot be modified separately. In particular, epistemic contours are not to be deactivated by interventions, which remain defined only for directed edges (i. e., only for causal mechanisms). An epistemic contour $c_{i,j}$ between two variables marks these variables as dependent but not connected causally – a third common cause can be excluded, because intervening on either variable directly (and simultaneously, i. e., at the same stage of computation) changes the value of the other variable as well. In other words, V_i and V_j are bound intrinsically in such a way that there exists no suitable intervention to detect the direction of any "causal flow."

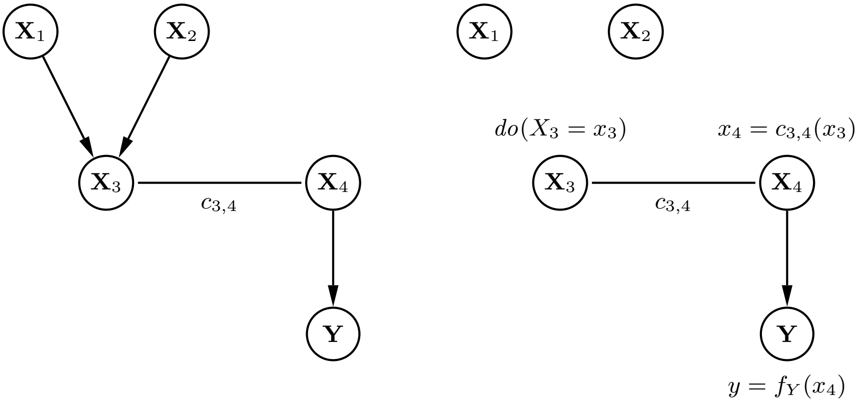

Fig. 3.1: Intervening on the variable X_3 lifts it from the (causal) influence of variables X_1 and X_2 but does not clip the link between X_3 and X_4.

Consider the left graph of figure 3.1 where two directed acyclic subnets $X_1 \rightarrow X_3 \leftarrow X_2$ and $X_4 \rightarrow Y$ are connected by the epistemic contour $c_{3,4}$ (in the graphical part of the model also denoted by the name '$c_{4,3}$'). Now, the intervention on the variable X_3 is expressed in the graph by removing the arrows connecting X_3 to its parents X_1 and X_2 but upholding the link between X_3 and X_4. This epistemic contour represents deterministic transfer of knowledge and does its job as soon as the intervention $do(X_3 = x_3)$ (i. e., the assignment of the value x_3 to the variable X_3) is performed. X_4 receives its value x_4 through the function $c_{3,4}$ and subsequently passes its value on to the causal mechanism f_Y which takes $c_{3,4}(x_3)$ as the only argument and uniquely computes the outcome y. This example illustrates with the structure X_3 — X_4 what it means to be an *EC clique*, as defined above in def. 3.3.1. To allow for consistent inference from causal knowledge patterns, the formulation of suitable restrictions on the construction and the manipulation of these structures is in order.

3.4 Causal knowledge patterns: design and manipulation

The demand for acyclicity

Just as with the directed *acyclic* graphical part of Pearl's causal models, consistent inference of causal claims from causal knowledge patterns crucially relies on these structures being *acyclic* as well.

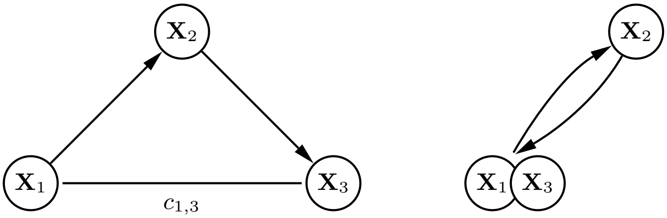

Fig. 3.2: The a-collapsibility criterion, illustrated: The left graph $D_{\mathcal{K}}$, the semi-DAG of the causal knowledge pattern $\mathcal{K}$, is collapsed to the right graph, which fails the test for acyclicity in this example.

The left graph of figure 3.2 shall serve as a motivation for the following considerations. Three variables are arranged in such a way that there is a causal chain $X_1 \twoheadrightarrow X_2 \twoheadrightarrow X_3$ and a shortcut X_1 — X_3 representing the epistemic contour $c_{1,3}$. In this example, the causal directedness of $X_1 \twoheadrightarrow X_2 \twoheadrightarrow X_3$ is rendered void by the existence of $c_{1,3}$: Intervening on X_3 should make it independent of all its predecessors, but it does not. Knowledge about X_3 propagates "backwards" over to X_1, since the epistemic contour $c_{1,3}$ does not get trimmed by the intervention performed. This way, intervening on either X_2 or X_3 influences the respective other variable and makes this pattern useless for causal analysis.

Now, The graphical criterion of *a-collapsibility* (with the 'a' denoting 'acyclic') tells the causal modeler if the pattern at hand is a proper causal knowledge pattern allowing for consistently inferring causal claims in accordance with intuition and background information.

Definition 3.4.1 (a-Collapsibility of Causal Knowledge Patterns)
A causal knowledge pattern $\mathfrak{K}$ is called a-collapsible *if pulling together all nodes of each EC clique in the semi-DAG $D_{\mathfrak{K}}$ into one single compound node per clique while leaving intact all arrows pointing at or rooting in any of the unified nodes (thereby removing all undirected edges) results in a (directed)* acyclic graph.

The right graph of figure 3.2 of our example above shows the collapsed version of the left graph. Nodes X_1 and X_3 of the EC clique $X_1 — X_3$ are pulled together into the compound node $X_{1,3}$ while all arrows connecting X_1 and X_3 to other nodes in the graph (in this case only X_2) are left intact, i. e., simply redirected into or out of the newly established compound node. The resulting graph nevertheless fails the test for acyclicity, since it cannot be defined recursively with X_2 and $X_{1,3}$ being each other's parent nodes simultaneously. Hence, the original graph $D_{\mathfrak{K}}$ is analyzed as *not a-collapsible*, which would in any case be demanded of a suitably designed causal knowledge pattern.

Causal effects in epistemically equivalent causal knowledge patterns

What it means to be a cause in the framework of causal knowledge patterns is carried over directly from the interventionist account of causation based on Bayes net causal models where a cause is an event which, when intervened on, brings about corresponding change in its effects. Interventions in causal knowledge patterns ultimately test for the direction of the causal flow, too – *epistemically* interpreted and not necessarily "push-and-pull." In this epistemic account of causation epistemic contours pass on information both ways – stably, deterministically, and not interruptibly. Epistemic contours themselves, though, mark *non-causal* relations but may nevertheless represent portions of paths that are said to be *causal.* If a causal effect is identifiable, at all, it can be computed uniquely within causal knowledge patterns (upon limiting possible settings). Pearl's concept of the identifiability of an effect, however, relies on the notion of d-separation, which is not defined for the semi-DAGs of causal knowledge patterns, yet. The extension of Pearl's criterion (as given in definition 2.6.3 on page 52) is straightforward, nonetheless.

Definition 3.4.2 (*d*-Separation for Causal Knowledge Patterns)
A path in the semi-DAG $D_{\mathcal{K}}$ of the causal knowledge pattern $\mathcal{K}$ (i. e., a sequence of nodes either connected by directed or undirected edges) is called d-separated if it is not active *in accordance with the d-separation criterion for DAGs when epistemic contours are treated as* null transitions *in the detection of chains, forks, or colliders.*

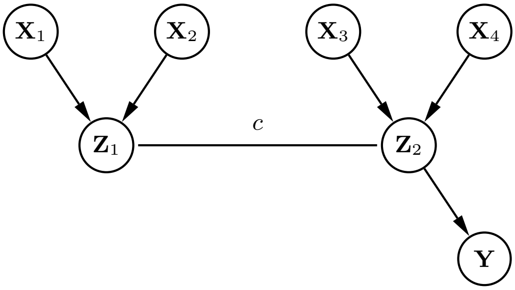

Fig. 3.3: *d*-separation in the semi-DAG $D_{\mathcal{K}}$ of the causal knowledge pattern $\mathcal{K}$. The epistemic contour along Z_1 — Z_2 (labeled c without indices, since it is the only one) presents a *null transition* for the detection of forks, chains, and collider structures.

Example (*d*-Separation in causal knowledge patterns)
Consider the semi-DAG in figure 3.3. The epistemic contour along Z_1 — Z_2 presents a null transition *for the application of the d-separation criterion in accordance with definition 3.4.2. Now, applying the criterion yields for example the following independencies (dependencies, respectively):*

1. $(X_1 \perp\!\!\!\perp X_4 \,|\, \varnothing)$*, since (by virtue of the skipped null transition) the virtual compound node $Z_{1,2}$ acts as a collider, blocking the flow of information along $X_1 \rightarrow Z_1 — Z_2 \leftarrow X_4$;*

2. $(X_1 \not\perp\!\!\!\perp X_4 \,|\, Y)$*, since Y is a descendant node of the compound collider $Z_{1,2}$;*

3. $(X_1 \perp\!\!\!\perp Y \,|\, Z_2)$*, since the value of Z_2 determines the value of Z_1 and the compound node $Z_{1,2}$ blocks the flow of information along the chain $X_1 \rightarrow Z_1 — Z_2 \rightarrow Y$;*

4. $(X_4 \perp\!\!\!\perp Y \,|\, Z_1)$*, even if Z_1 is not situated on the path $X_4 \rightarrow Z_2 \rightarrow Y$. Nevertheless, the value of Z_1 determines the value of Z_2, thereby d-separating X_4 and Y.*

This example shows in particular, that for the independencies that can be read off the graph it does not make any difference if the node representing the variable Y is connected to Z_1 or to Z_2, i. e., $Z_2 \rightarrow Y$ can be exchanged for $Z_1 \rightarrow Y$ in the semi-DAG above without loss of information if f_Y is suitably reformulated (taking the value of Z_1 as argument) or exchanged for $f'_Y := f_Y \circ c_{\rightarrow}$.[20] This observation shall be summed up in the principle of *epistemic equivalence* of causal knowledge patterns:

Definition 3.4.3 (Principle of Epistemic Equivalence of Causal Knowledge Patterns)
Two causal knowledge patterns $\mathcal{K}_1$ and $\mathcal{K}_2$ are called epistemically equivalent *if both possess the same variables and the same epistemic contours but in the graphical part possibly differ in the set of directed edges pointing towards or rooted in nodes of an EC clique in such a way that*

1. *for all nodes pointing towards a node of the EC clique the arrow tail is the same for both CKPs, while the anchor of the head may differ;*

2. *for all nodes rooted in a node of the EC clique the arrow head is the same for both CKPs, while the anchor of the tail may differ;*

3. *all nodes of this EC clique assume the same values in both CKPs if the parents of these nodes assume the same values in both CKPs;*

4. *all nodes that are children of nodes of this EC clique assume the same values in both CKPs if the nodes of this very EC clique assume the same values in both CKPs.*

The principle of *epistemic equivalence* basically states that two causal knowledge patterns might convey the same information even if they exhibit structural differences in the arrows docked onto EC cliques. In other words, inferring causal knowledge from CKPs is insensitive to along what paths information is fed into and propagated onward from EC cliques (and what the formulation of the associated causal mechanisms may be), as long as the unified sets of predecessor and successor nodes of each EC clique remain untouched. The very general formulation of the principle of epistemic equivalence emphasizes that not all combinatorially possible restructurings of a given CKP $\mathcal{K}$ (i. e., permutations of parent–child relations directly neighboring EC cliques) generate CKPs epistemically

[20] Since the epistemic contour c does not possess any indices, $c_{\rightarrow}$ indicates one function of the pair of epistemic contours $c_{\rightarrow}/c_{\leftarrow}$ with $c_{\rightarrow} : Ran(Z_1) \rightarrow Ran(Z_2)$.

equivalent to $\mathcal{K}$. E. g., if a node X_i of some EC clique within $\mathcal{K}$ acts as a collider node, i. e., X_i is a child of more than one parent node in PA_i, and the causal mechanism f_i acts like a logical *OR* switch, then all CKPs epistemically equivalent to $\mathcal{K}$ where *one* of the arrows pointing at X_i is redirected to point towards X_j have *all* arrows originally taking part in the causal mechanism f_i redirected to X_j (i. e., f_j takes over all the arguments of f_i). Nonlinear belief propagation in accordance with the mechanism of the *OR* switch cannot be ensured in any other way.

Docking arrows onto EC cliques

Just as when choosing suitable variables for embedding into causal models within Pearl's framework, the choice of how to dock influent arrows onto EC cliques is ultimately left to the causal analyst, too. As long as there is only one arrow pointing towards one of the nodes of a specific EC clique, the answer to the question, where to anchor the arrow head, boils down to the precise formulation of the causal mechanism that corresponds to this very single influent arrow. With the intended cases of application in mind this decision is made naturally in general, since if, e. g., different nodes of an EC clique represent observed events formulated in languages of different disciplines, a single influent arrow will be anchored to the node of the same language as the pertaining causal mechanism. In general, this principle extends to multiple arrows entering an EC clique "from above" and analogously to one or more arrows rooted in an EC clique and pointing towards other nodes "below."

One more word about the Markovian assumption and the integration of epistemic contours is in order here: Figure 3.1 exhibits the structure $\{X_1, X_2\} \rightarrow X_3 \text{ — } X_4$ where X_4 is treated as an endogenous variable in accordance with definition 3.3.2, i. e., it receives its value by assignment through $c_{3,4}$ with the value of X_3 as its only argument. X_3, however, does not comply with the Markov assumption in that its value is set by f_3 and by $c_{4,3}$ at the same time. This does not pose a problem here – X_3 can not be assigned incompatible values, since X_4's value is not set before X_3's when the system of equations for the network is solved step by step. This is always the case with EC cliques that are set by one sole path "from above."

A causal knowledge pattern behaves differently, though, in the case of multiple arrows entering one and the same EC clique. Now, one variable might be assigned contradictory values *at the same time*, i. e., it might be set through a causal path “from above” and additionally receive incompatible information via an epistemic contour. E. g., the variable Z_1 in figure 3.3 contradicts the Markovian assumption in exactly this way: z_1 is calculated by drawing on the pertaining causal mechanism f_{Z_1}, which only takes the obtaining values of Z_1’s parents as its arguments. This assignment by f_{Z_1} might be contradicted by the equation for the 1-1 relation $c_{\leftarrow}$ setting Z_1 by computing Z_2. The contradiction, formally:

(i) Let $f_{Z_1}(x_1, x_2) = \min(x_1, x_2)$ and $f_{Z_2}(x_3, x_4) = \max(x_3, x_4)$;

(ii) let $c_{\leftarrow} = c_{\rightarrow} = \mathrm{id}$;

(iii) let $x_1 = x_2 = x_3 = 0$ and $x_4 = 1$
with $X_1, \ldots, X_4 \in \{0, 1\}$ dichotomous variables;

$$\left.\begin{array}{ll} \text{(iv)} & z_1 = f_{Z_1}(0,0) = \min(0,0) = 0 \\ \text{(v)} & z_2 = f_{Z_2}(0,1) = \max(0,1) = 1 \Rightarrow z_1 = c_{\leftarrow}(z_2) = \mathrm{id}(1) = 1 \end{array}\right\} \lightning$$

This contradiction arises, since $\{X_1, X_2\}$ and $\{X_3, X_4\}$ are independent in the first place, which can be verified by the extended criterion of d-separation for causal knowledge patterns – the mechanisms f_{Z_1} and f_{Z_2} might, e. g., represent independent experimental designs independently resulting in and accounting for two strictly correlated observed phenomena (e. g., measured or labeled differently) *if the experiments are actually performed*. In other words, each causal history entering an epistemic contour supports one possible explanation of the occurrence described by this very epistemic contour. Explanation is understood here as the act of post factum singling out a set of variables of the causal knowledge pattern $\mathfrak{K}$ and naming the obtaining values to answer the question “Why did x occur?” or singling out a sub-pattern of $D_{\mathfrak{K}}$ (and naming the corresponding variables’ values) to answer the question “How was x produced?”[21]

Now, the establishment of the epistemic contour c is only justified in the first place if there is the possibility of assigning values to the exogenous variables $X_1, \ldots, X_4$ such that the system of equations for the causal knowledge pattern can be solved consistently, at all. One distinguished case, the non-interventional consistent *initial situation*, will be marked by *default assignment* as the *default case*. The basic assumption beneath

[21] Note that the notion of explanation is in this sense always relative to a causal knowledge pattern $\mathfrak{K}$ which might either be induced by the question itself or made explicit in the answer.

this is that the default of the epistemic contour under consideration has corresponding defaults in all its exogenous predecessor variables *coupled* in marking this situation.[22] Of course, what the default value of an EC clique is, differs from one context to another. It is a highly intensional concept, after all. But so is the concept of causal analysis in causal knowledge patterns with epistemic contours. The concept of *default* does not have to remain obscure, though – on the contrary, its integration contributes to the computability and to the applicability of causal knowledge patterns. Christopher Hitchcock, who makes the point that occurrences *deviating from normality* are much rather attributed causal efficacy than those *following the normal course of events*,[23] be quoted here for an elaborate view on what *defaults* and *deviants* essentially are:

> *As the name suggests, the default value of a variable is the one that we would expect in the absence of any information about intervening causes. More specifically, there are certain states of a system that are self-sustaining, that will persist in the absence of any causes other than the presence of the state itself: the default assumption is that a system, once it is in such a state, will persist in such a state. Theory – either scientific or folk – informs us which states are self-sustaining in this way. For example, Newtonian physics tells us that an object's velocity is self-sustaining, whereas its acceleration is not. Thus the default is that the object will maintain the same velocity. The default may depend upon the level of analysis. Consider, for example, a variable whose values represent the state of an individual – alive or dead. It is a plausible principle of folk biology that an individual will remain alive unless something causes her to die, hence it would treat 'alive' as the default value of the variable. But from the perspective of a physiologist, remaining alive requires an amazing effort on the part of complex, delicate systems, as well as interactions with the environment; hence death might be viewed as the default state. Perhaps a case could be made for allowing only genuine laws of nature to determine default values of variables, but if we disallow folk theories, we are not likely to arrive at a theory that accords with folk intuitions. Note also that the default value of a variable may not be an intrinsic feature of the state that is represented. That is, we could have two individuals in the very same state, while one is in a deviant state and the other in a default state.*[24]

[22] Turning this assumption upside down means that if any of the EC clique's variables is assigned a deviant value through its pertaining causal mechanism, this specific variable must have at least one exogenous predecessor exhibiting an efficacious deviant value, too – either observed or set by intervention.

[23] See Hitchcock's considerations on this issue in [Hitchcock 2009a].

[24] Cf. [Hitchcock 2007, p. 506] – quoted here without footnotes.

The following considerations shall make explicit under what circumstances a causal knowledge pattern is of use to the causal researcher in need of explanation, what restrictions are required of such a pattern if it is to be used for prediction, and how information transfer via epistemic contours can promote cross-framework counterfactual reasoning.

Maintaining consistency in the observational case

The following refers w. l. o. g. to causal knowledge patterns with one epistemic contour for the sake of simplicity. Observation will always yield consistent results if there is only one common orphan predecessor node for all nodes in an EC clique – the epistemic contour would not be justified otherwise, it makes explicit precisely this feature of the causal knowledge pattern under consideration.[25] Consider the left graph of figure 3.4 where observing $X_2 = x_2$ will feed the initial value into the system of equations represented by $D_{\mathfrak{K}}$, which will then take care of consistent propagation of belief to X_1 and X_3.

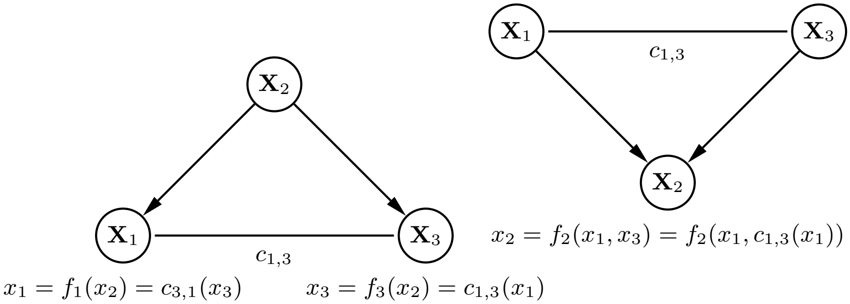
X2
X1
X3
$c_{1,3}$
$x_1 = f_1(x_2) = c_{3,1}(x_3)$
$x_3 = f_3(x_2) = c_{1,3}(x_1)$
X1
X3
$c_{1,3}$
X2
$x_2 = f_2(x_1, x_3) = f_2(x_1, c_{1,3}(x_1))$

Fig. 3.4: If the semi-DAG $D_{\mathfrak{K}}$ of a causal knowledge pattern $\mathfrak{K}$ is to be used for consistent causal inference, restrictions on analysis and intervention by the $do(\cdot)$-operation apply.

If the set of orphan predecessor nodes of a certain EC clique consists of more than one node, though, these nodes are self-evidently d-separated and their paths into the EC clique semantically independent in general. Nevertheless, in accordance with the explications above, there has to be some default situation that renders solving the system of equations

[25] Note that this is the only case where the orphan predecessor nodes of the EC clique under consideration are (trivially) non-d-separated. Nevertheless, if hidden disturbances are not excluded from the analysis, these would (at least partly) have to be analyzed as dependent.

consistent and entails compatible values in the 1-1 assignment within the EC clique, thereby licensing the integration of the epistemic contour in the first place. Consider for example the graph $D_{\mathfrak{K}}$ in figure 3.3 with the numerical example (given on page 100). The default situation is marked by the assignment of 0 to the variables within the EC clique Z_1 — Z_2 and can be predicted from the observation that the EC clique's orphan predecessors' values are all 0, too, where 0 values might in turn be the presumed defaults of all the exogenous variables. Nevertheless, if only X_4 were observed as assuming 1, solving the system of equations for the causal knowledge pattern would yield inconsistent assignments. Still, the deviant observation $X_4 = 1$ would be attributed causal efficacy since it stirs the stable equilibrium of the self-sustaining default situation. If observing solely the sub-pattern through which $X_4 = 1$ enters the EC clique leads to a unique prediction of Z_1 and Z_2, this sub-pattern would be drawn on in explaining the effect (i. e., the change in Z_1 and Z_2), whereas other paths rooting in variables that persevere in default state would be dismissed as irrelevant. This concept is summed up in the following definition of *explanatory dominance.*

Definition 3.4.4 (Principle of Explanatory Dominance)
A set of variables Z_1 in a certain EC clique connecting the set of variables C is said to weakly explanatorily dominate *the set of variables $Z_2 \subseteq C \backslash Z_1$ in world ω iff Z_1's variables*

(i) all show compatible values (relative to the epistemic contours),

(ii) exhibit a value incompatible with at least one of the values in Z_2,

(iii) and at least one of their orphan predecessors shows a deviant value with effective influence *on Z_1 (i. e., were it set to a different value, at least one variable in Z_1 would be assigned differently, too).*[26]

Moreover, Z_1 is said to strictly explanatorily dominate *the set Z_2 iff, in addition to (i)–(iii), Z_2's orphan predecessors (excluding such on paths through the epistemic contour itself) only assume default values.*[27] *If Z_1 is the maximal set strictly explanatorily dominating the set $C \backslash Z_1$, the causal history of any of the variables in Z_1 may support explanation of the obtaining values of Z_1, while all other histories are marked as* irrelevant*.*

[26] This condition neglects deviants whose influence is absorbed or overridden by specific causal mechanisms, but attributes relevance to multiple deviants that would be efficacious if they did not cancel each other precisely.

[27] This ensures with the explications above that Z_2 also receives its default value and that it does not receive this default value from multiple exogenous deviants cancelling each other precisely.

The principle of explanatory dominance thus draws on background knowledge about defaults and deviants to support (predictive) determination of the values of EC clique variables. While the weak version simply says that if Z_1 dominates Z_2, Z_2 cannot be used as assignment target, the strict version implies that Z_1 in fact is a good candidate for providing the desired value (and marks Z_2's history as irrelevant for explanation in case C is partitioned by Z_1 and Z_2). Graphically, the implications of this principle may be expressed in the semi-DAG by pruning all *irrelevant* causal histories' entry links into the EC clique. In the case of our example above (observing $X_4 = 1$ in the causal knowledge pattern represented by figure 3.3), the singleton $\{Z_2\}$ would be said to strictly explanatorily dominate (as the maximal set with this property) the set $\{Z_1\}$, since the deviant observation in X_4 would be drawn on for explaining the changed value of Z_2, while the causal mechanism f_{Z_1} would be denied *explanatory power* – graphically expressed by pruning $X_1 \rightarrow Z_1$ and $X_2 \rightarrow Z_1$ (which makes Z_1 an endogenous variable receiving its value only from $c_{\leftarrow}(z_2)$ and dismisses the critical inconsistent case addressed and barred by the Markov assumption/restriction in the first place).

Finally, if more than one of the exogenous predecessor variables of a certain EC clique exhibit deviant values (in some observed world ω), the respective causal knowledge pattern only retains its value for prediction and explanation in case the EC clique variables with deviant exogenous predecessors receive compatible values. If this compatibility is not ensured, the causal knowledge pattern becomes useless for causal inference (in ω) – none of the deviating causal histories of the respective EC clique will be preferred above the others, and no recommendation for suitable (graphical) restructuring can be read off the data at hand to possibly re-establish the Markov condition (adjusted for semi-DAGs as explicated above). This might lead to dismissing (or refining) the considered causal knowledge pattern, ultimately, since the course of events ω calls for an analysis differently structured (in more detail, respectively).

Interventions in causal histories of EC cliques and multiple revisions in the graph[28]

Consider the left semi-DAG in figure 3.4 where node X_2 is connected to the EC clique X_1 — X_3 by the superstructure $X_1 \leftarrow X_2 \rightarrow X_3$ such that both variables X_1 and X_3 receive their values from the corresponding

[28] I am especially thankful to Manfred Schramm for helpful advice on how to pin down the ideas in this section.

causal mechanisms each of which only compute the value of X_2 for this task:

$$\begin{aligned} x_1 &= f_1(x_2) &= c_{3,1}(x_3) \\ x_3 &= f_3(x_2) &= c_{1,3}(x_1) \end{aligned}$$

Now, if X_3 were to be intervened on atomically, the local surgery removing the directed edge $X_2 \rightarrow X_3$ should amount to lifting X_3 from the influence of X_2 – but without further refinement it does not: The modification of the structural equations only renders the causal mechanism for X_3 void, but it leaves the deterministic epistemic contour intact such that X_3 might now be assigned its value by two contradicting equations. X_3 is still influenced indirectly by X_2 in virtue of the epistemic contour $c_{1,3}$ along which the "flow of information" remains unblocked. To be able to answer the question "What is the causal effect of *doing* x on some epistemic contour?", the influence of setting the variable X to the value x on the variables in the respective EC clique must be uniquely computable. In other words, determining the values of the variables in some EC clique upon performing $do(X = x)$ requires that

(i) the effect of *doing* $X = x$ on the variables in this very EC clique is identifiable, and

(ii) the affected variables in the EC clique (i. e., those with X as a predecessor) are not explanatorily dominated by any set of variables that are not affected (such a dominating set would be any set of variables whose orphan predecessors contain deviants possibly leading to deviant values in the EC clique).

PEARL's criterion of the identifiability of causal effects (see definition 2.10.1) must consequently be extended suitably to be applicable to epistemic contours in causal knowledge patterns, too.

Definition 3.4.5 (Identifiability of Causal Effects on Epistemic Contours)
The effect of the intervention $do(X = x)$ on some EC clique is said to be identifiable *iff every variable in the EC clique that has X amongst its predecessors becomes (in the graphical representation) d-separated from* PA_X *upon removing all links from* PA_X *into X.*

This criterion rules out the case of mixed influence – observational and interventional at the same time – on variables in the respective EC clique. If any path from deviant *observation* into the EC clique is interrupted by the intervention, then this manipulation, hypothetically performed, may account for change in the EC clique. The values affected by the intervention will now determine the values of all other variables in this EC clique according to the following considerations, which shall be split up into three cases, namely

A. *direct* interventions on EC clique variables,

B. single interventions *above* EC cliques, and

C. arbitrary *compound* interventions.

Case A. Direct interventions on EC clique variables. Any two variables connected by an EC clique are strictly dependent, i. e., especially in the case of an EC clique of simultaneously exogenous variables it becomes obvious that the Markov assumption of mutual independence is overridden (locally for these variables) but can in principle be regained by postulating that setting one variable in this EC clique sets all variables in this EC clique in accordance with the pertaining epistemic contours.[29] In the case of inner epistemic contours, whatever the causal mechanisms may be for each of the entry points into a specific EC clique, setting one variable effectively means that the complete EC clique is lifted from the influence of the joint set of parents of all its variables. One variable could not be set to the desired value if some other causal mechanism were to interfere, but if the structural surgery is carried out and the assignment performed, this must mean that no other influence overrides this intervention. Graphically, intervening on one variable within an EC clique cuts all links into this very EC clique, which receives its compatible values from the single $do(\cdot)$-affected variable. Applying this principle to the example above, intervening in the left graph of figure 3.4 by *doing* $X_3 = x_3$ will not only remove $X_2 \rightarrow X_3$ but also cut out $X_2 \rightarrow X_1$. The value of X_1 is consequently assigned by $c_{3,1}(x_3)$ after setting X_3 to x_3.

Compound interventions, if directly performed on EC cliques, are of course required to not oppose each other. The following example shall give an illustration of how causal inference would be rendered impossible if such opposing interventions were not ruled out. After that, the special case of *ladder structures* will be discussed in a second example below.

[29] Observing *incompatible* values in simultaneously exogenous EC clique variables contradicts the CKP design in the first place and would naturally lead to dismissing this structure.

Example (Opposing interventions on EC clique variables)
Consider the pair of epistemic contours $c_{i,j}/c_{j,i}$ *represented in the semi-DAG* $D_{\mathfrak{K}}$ *of the causal knowledge pattern* $\mathfrak{K}$ *as the undirected edge* c *in the structure* $X_i \overset{c}{-} X_j$*. This minimal EC clique is associated with the pair of mutual assignments*

$$\begin{array}{rcll} x_i & = & c_{j,i}(x_j) & \quad \text{and} \\ x_j & = & c_{i,j}(x_i). & \end{array}$$

Setting either variable to a constant value by external intervention simultaneously determines *the value of the second variable in accordance with the system of equations above. Intervening on both variables at the same time by joint manipulation causes trouble (in general) if the interventions are performed independently, since* X_i *and* X_j *might be assigned values that are not compatible with the epistemic contour* $c_{i,j}$ *anymore (and* $c_{j,i}$*, respectively), as in the following situation where the two dichotomous variables* X_i *(taking distinct values* x_i *or* x'_i*) and* X_j *(taking distinct values* x_j *or* x'_j*) are set simultaneously.*

(i) $c_{i,j} = \{\langle x_i, x_j\rangle, \langle x'_i, x'_j\rangle\}$

(ii) $c_{j,i} = c_{i,j}^{-1}$

(iii) Simultaneous manipulations (compound intervention):
$do(X_i = x_i)$ *and* $do(X_j = x'_j)$

$$\left.\begin{array}{l} \textit{(iv) with } do(X_i = x_i)\textit{: } X_j \textit{ assumes } c_{i,j}(x_i) = x_j \\ \textit{(v) with } do(X_j = x'_j)\textit{: } X_i \textit{ assumes } c_{j,i}(x'_j) = x'_i \end{array}\right\} \lightning$$

Lines (iv) and (v) make the contradiction obvious: Setting the variables in opposition *to each other ('opposed' relative to* $c_{i,j}/c_{j,i}$*) makes the epistemic system collapse and renders further consistent inference impossible. If hypothetical compound interventions are to be performed, at all, they have to be performed in mutual dependence, i. e., by (systematically or pragmatically) suitable restrictions embedded in the intuitions on which the epistemic contours are formulated in the first place.*[30]

Example (Intervening in ladder structures)
Consider the left diagram of figure 3.5, which exhibits a ladder structure*, possibly connecting two frameworks of description with epistemic*

[30] This is quite in analogy with the systematical and mathematical constraint in Pearl's causal models that no variable can be intervened on by multiple interventions setting opposing values *simultaneously*.

contours bridging corresponding atomic events in both systems (framework A and B) along the development over time – the causal chains $X_1 \to X_3 \to X_5$ and $X_2 \to X_4 \to X_6$ are multiply connected in this graph. At each stage of the system's development knowledge may be exchanged both ways.[31] *The causal histories on each side of the ladder are (on a higher level) as tightly connected as their components. Now, intervening on X_2 by setting it to x_2 should yield the same course of the world as simply observing $X_2 = x_2$. X_2 is an exogenous variable, after all (which also means that X_1 and X_2 are* simultaneously exogenous*). The value of X_1 must only be computed by drawing on $c_{2,1}$, in other words,* doing *x_2 also determines x_1 and lifts both variables from the influence of any potential latent background variable. If X_4 is to be intervened on, however, merely pruning the link $X_2 \to X_4$ will not suffice for lifting X_4 from the influence of X_2, since the path $X_2 — X_1 \to X_3 — X_4$ remains unblocked. Again, X_3 and X_4 are marked as variables that cannot be decoupled and are only to be modified simultaneously. Consequently,* doing *x_4 must also lift X_3 from the influence of its parent variables (X_1 in this case) and set X_3 to $c_{4,3}(x_4)$.*

Case B. Single interventions above EC cliques. If the influence of some variable to be intervened on on a specific EC clique is mediated by causal mechanisms (i. e., by directed edges in the graph), then causal reasoning can only be carried out non-paradoxically in case that (i) the effect (on the EC clique) to be made out is identifiable and (ii) the value brought about in the affected EC clique variables is not explanatorily dominated by any set of non-affected variables in the same EC clique.

Example (Testing effects on EC cliques with independent histories) *Consider the middle diagram of figure 3.5 where the EC clique $X_5 — X_6$ is influenced along two separate causal histories that are purposely not linked by further epistemic contours (e. g., between X_1 and X_2). These strands might represent two alternative causal paths, e. g., two different experimental designs producing strictly correlated observations X_5 and X_6 (by actively deviating from the passive default situations in X_1 or X_2). $X_5 — X_6$ acts like a logical* OR*: Intervening on X_3 will remove the directed edge $X_1 \to X_3$ and virtually* decouple *the causal histories of X_5 and X_6 (for X_2 assuming its* default *value) by graphically removing the entry link $X_4 \to X_6$ (whose pertaining causal history – default in character – can be called* irrelevant *in the sense of def. 3.4.4).*

[31] The epistemic contours depicted in the graph of figure 3.5 are labeled with the name of only one of the functions they represent – $c_{1,2}$ also signifies $c_{2,1}$ in accordance with definition 3.3.2 above.

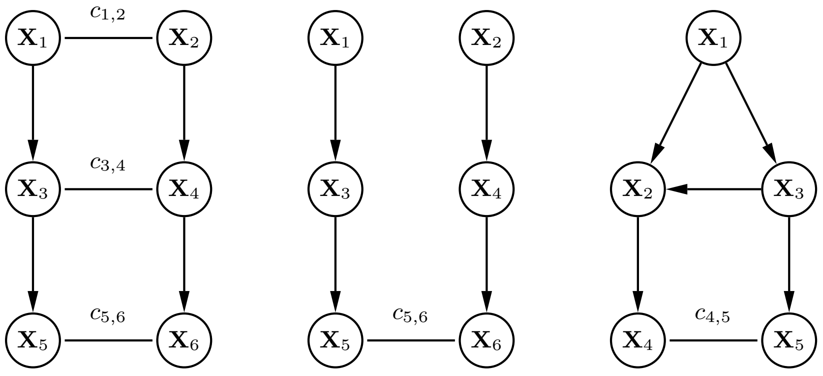

Fig. 3.5: In ladder structures (illustrated in the left graph) information is exchanged between both chains at each stage. The middle and the right graph show examples of differently structured causal histories leading up to X_4 and X_5.

Example (Testing effects on EC cliques with converging histories)
The right diagram of figure 3.5 shows an epistemic contour X_4 — X_5 whose variables' causal histories converge in X_1. Now, intervening on X_1 does not pose any problem for the value assignment to X_4 and X_5, since the formulation of the systems of equations for each causal history warrants consistent outcomes. Intervening on X_3, however, cuts the link $X_1 \twoheadrightarrow X_3$ and might result in contradicting (incompatible) value assignments to X_4 and X_5, because X_4 is still potentially influenced by X_1 as well. The effect of setting X_3 to x_3 is not identifiable *in accordance with definition 3.4.5 above: Pruning $X_1 \twoheadrightarrow X_3$ does not lift X_4 from the influence of* PA_3 *(X_3's parents), and X_3 is at the same time an element of the set of X_4's predecessors.*[32]

Intervening on X_2, however, will result in removing the edges $X_1 \twoheadrightarrow X_2$ and $X_3 \twoheadrightarrow X_2$ simultaneously such that the effect of $do(X_2 = x_2)$ on the EC clique X_4 — X_5 is analyzed as identifiable. *Now, if X_1 exhibits its default value, X_5 will also, and the set $\{X_5\}$ will consequently not explanatorily dominate the set $\{X_4\}$ of variables affected by* doing x_2. *The causal history of X_5 will be rendered void by removing the edge $X_3 \twoheadrightarrow X_5$*

[32] This analysis is analogous to applying Pearl's criterion for the identifiability of causal effects (see definition 2.10.1) together with the back-door criterion (see definition 2.10.2) in Bayes net causal models without undirected edges; the analogy can be seen directly when X_4 and X_5 are pulled together into a compound node.

(thereby re-establishing the extended Markov restriction for world ω with X_1 at its default value and attributing explanatory power to X_4 with its pertaining history).

Case C. Arbitrary compound interventions. Performing multiple interventions simultaneously can be done by securing the above conditions (i. e., *identifiability* and *non-dominatedness*) under the additionally imposed restriction that – as in the case of intervening on EC cliques directly – arbitrary interventions must not lead to opposing values within the same EC clique, i. e., the set of affected variables in some specific EC clique may only show compatible values upon intervening. Then, as before, all remaining causal histories (if they do not produce a set of explanatorily dominating variables) may be *deactivated* by removing their entry links to the EC clique under consideration. Again, this ensures consistent causal inference and explicates graphically what *decoupling experimental designs* means.

In conclusion, all the above cases show how consistent inference from type causal structures becomes partly relativized to token causal findings in that the default situations mark certain subsets of possible worlds (i. e., courses of events) and thereby facilitate prediction, explanation, and – ultimately – compact formulation of target-oriented strategy even in the case of independent alternative causal histories.

From epistemic contours onwards

When considering edges directed away from EC cliques, things stand differently. The right diagram of figure 3.4 shows the situation where information from the EC clique is jointly used for the computation of X_2's value x_2. In this case, X_2 seems to be jointly causally influenced by both of its parents, and intervening on X_2 correctly lifts it from the influence of its parents, but the formulation of the causal mechanism pertaining to X_2 can be reduced in the following manner:

$$\begin{aligned} x_2 &= f_2(x_1, x_3) \\ &= f_2(x_1, c_{1,3}(x_1)) = f_2'(x_1) \\ &= f_2(c_{3,1}(x_3), x_3) = f_2''(x_3) \end{aligned}$$

This shows that one of the arrows pointing towards X_2 is *superfluous* for the propagation of knowledge, i. e., this piece of structural information does not give us any additional computational information we did not have before. The value of X_2 can be calculated from one of the nodes within the EC clique alone. Since all epistemic contours are just as stable and autonomous as the causal mechanisms in the causal knowledge pattern, one of the functions of the pair $c_{1,3}/c_{3,1}$ can thus be coded into f_2 directly, which basically makes f_2 a function of x_1 alone (of x_3, respectively), as given above by f_2' (by f_2'', respectively). One of the links $X_1 \twoheadrightarrow X_2$ or $X_3 \twoheadrightarrow X_2$ can thus be called a *pseudo*-link just as a directed node connection that is removed in the process of refining the model – going from the fully connected graph to a slimmer version – aiming at specifying the examined situation in the most informative way for the derivation of meaningful causal claims. In the case of slimming down the fully connected directed graph just as in the case of deleting superfluous *pseudo*-links, the principle of *Occam's razor* and good implementation practice tell us that introducing (or upholding) redundant information is to be avoided.[33] I believe this idea also appertains to the features of the economical principle of knowledge organization called causality in this account. As a structural rule, collider nodes in a semi-DAG linked to more than one parent node of the same EC clique may in many cases be reduced away for reasons of economy (unless associations along different paths are precisely to be emphasized as in cases of decision making, see section 4.1).

Mimicking hypothetical interventions and learning by abduction

The above explications elaborate how contradictory conclusions from knowledge represented in causal knowledge patterns can be avoided by adding one more ingredient – default assignments. If observation is overridden by external manipulation, though, testing for compatible values in epistemic contours helps the researcher rule out contradictory experiments (relative to a subset of possible worlds) or experimental designs altogether (if two independent designs never yield consistent values or only trivially in one marked possible world). The question, what manipulations go together well and what manipulations are to be avoided for which obtaining observations is answered by the above rules.

[33] Note that introducing epistemic contours is not understood as falling under this verdict in the first place, since the variables connected by epistemic contours still convey *intensional* knowledge and *additional connotations* that might be exploited in the further augmentation of the causal knowledge pattern under consideration. This does not hold for superfluous directed edges as in the example, since they can be reduced mathematically. Also see Pearl's short remark on model preference and Occam's razor in [Pearl 2009, sect. 2.3, pp. 45 ff.].

Turning this very question around, one might consider epistemic contours that bridge technical or descriptive frameworks within one causal knowledge pattern such that, e. g., $X_i \overset{c}{—} X_j$ represents this inter-framework bridge with X_i belonging to framework A and X_j belonging to framework B. When manipulating the direct causes of X_i within framework A, a typical answer we might look for now is what the corresponding intervention in framework B would be in order to bring about X_j. Causal knowledge patterns serve as *oracles of abduction* for this kind of counterfactual deliberations.

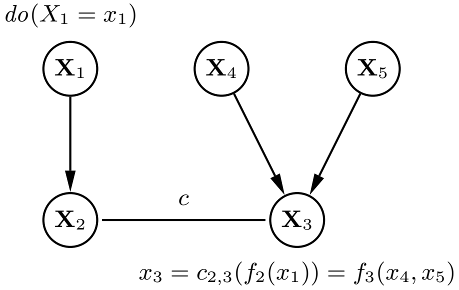

Fig. 3.6: Intervening on the variable X_1 in the semi-DAG given here yields knowledge about possible interventions “across c” that would be necessary to bring about some realization of X_3 directly through joint manipulation of X_4 and X_5.

The illustration in figure 3.6 shows a semi-DAG with the epistemic contour c possibly bridging two descriptive frameworks A and B with X_1, X_2 situated in framework A and X_3, X_4, X_5 in framework B. Intervening on X_1 by *doing* $X_1 = x_1$ enables us to read off the graph the value of X_3 computable by $x_3 = c_{2,3}(x_2)$ (where x_2 is calculated through $x_2 = f_2(x_1)$). The given causal knowledge pattern marks X_4 and X_5 as direct causes of X_3, setting X_3's value according to the causal mechanism f_3, such that $x_3 = f_3(x_4, x_5)$ when the arrows pointing towards X_3 are *not* removed but kept intact. Having all this knowledge at hand makes the causal knowledge pattern an informant about hypothetical interventions in framework B that entail the same assignment to X_3 as the one just brought about indirectly by hypothetical intervention within framework A. The set of possible, simultaneous compound interventions on X_4 and X_5 entailing X_3's realization $c_{2,3}(f_2(x_1))$ is the set

$$\overset{-1}{f_3}\,[c_{2,3}(x_2)]$$

which is the set of inverse images of $c_{2,3}(x_2)$ under f_3, defined by

$$\overset{-1}{f_j}\,[c_{i,j}(v_i)] = \{\mathbf{pa}_j \mid f_j(\mathbf{pa}_j) = c_{i,j}(v_i)\}$$

with f_j being the causal mechanism for V_j and $\mathbf{pa}_j$ representing the (vector of) values of all parents (i. e., direct causes) of V_j.[34] Of course, iterated application of this abductive step will yield information about causal histories of any length.[35]

3.5 Reviewing the framework

Someone objecting to the implementation of epistemic contours in an extension of the manipulationist Bayes nets framework might argue that the nodes of one EC clique represent events that are not distinct – quite on the contrary, they are even strictly correlated – and should as such not be included in a well-designed causal model. The critic might add, there is nothing more to causal dependence than given in Bayesian nets, and epistemic contours or the intuition behind those should merely be guidelines in the modeling process, which ultimately yields well-known standard Bayes net causal models with all nodes denoting extensionally disjoint regions in spacetime.[36] This all might be crucial considerations in the business of equipping robots or inference machines with basic constraints on what to read off the noise acquired through their sensors – on whatever level of abstraction these might operate. In the case of tracing the mechanics of human causal reasoning (expressed in everyday language or specialized jargon), things lie differently. Pearl himself emphasizes the recourse to basal causal assumptions when the epidemiologist builds a causal model or when we discuss politics in private. If the intuitions behind epistemic contours serve as guidelines in this modeling process, they might as well be embedded in a suitably modified framework that allows for making explicit how we arrive at the blueprints of interrelated events structured by causal knowledge. Large portions of what we know about the relations of the events surrounding us is built upon non-causal data, i. e., as in the current proposal, deterministic non-directional knowledge. Epistemic contours, as introduced

[34] This notation is adapted from [Link 2009, p. 433] where Link defines the set "Urbildmenge von B unter f" as $\overset{-1}{f}\,[B] := \{x \mid f(x) \in B\}$.

[35] See also sect. 2.11 for a presentation of Pearl's adaptation of the concept of abduction to the analysis of *actual causes* in *epistemically related* twin networks.

[36] For a discussion of discerning events and times in the business of causal modeling see [Hitchcock (forthcoming)].

above, are included in shared bodies of information either (i) tentatively until theories are unified or concepts matched and labeled with the same name tag (thereby reducing the contents of the model to extensionally disjoint denotata) or (ii) purposely to mark intensions, aspects, perspectives, jargon, or frameworks or (iii) even because spatio-temporal locality or contiguity seems to be violated in the setting under consideration and the purely mechanistic framework fails to yield some insightful rendition. The embedding of non-causal knowledge for causal inference and for communicating (indirect) causal relations is facilitated by construing causal knowledge patterns in the unified formal system proposed above.[37]

The approach given here accounts for causal relations as epistemic relations by which knowledge is organized efficiently. It can be understood as a uniform account that does not treat causal claims pluralistically within their respective domains but on the contrary facilitates the unification of heterogeneous levels or disciplines in the same formal structure. Causal inference remains frame-relative with the shape of those frames being due to our cognitive faculty of carving out subsystems from what we perceive around us.[38] This does not put off events to the realm of pure imagination (or mental representation) but allows us to retain a solid event realism such that causal relations epistemically hold between representations of real (i. e., physically ontological) events. Following Pearl, analysis builds upon type knowledge and goes from there to token claims. It does so non-reductively in one sense – with causation being defined in terms of basal causal assumptions – and reductively in another – by deriving higher-level causal claims from lower-level ones. Causality becomes an epistemic principle of organizing knowledge efficiently by ordering it deterministically – always with the possibility of also evaluating probabilistic causal claims as propagation of belief blurred by unmeasured influences. Relying on the Bayes net framework, the CKP toolbox can readily be applied to the same settings as causal models. It will, however, be able to also treat examples that incorporate events which are entangled in epistemic manner or only turn out to be so in virtue of the extended capabilities of causal knowledge patterns.

[37] Point (i) above contains that in the CKP framework knowledge used for making causal claims can be explicated formally as a basis for experts to start disentangling (or unifying) variables – maybe of different levels of explanation – to tell us how things work physically ontologically if this is the aim of research and if the situation under consideration permits such an analysis, at all.

[38] Questions of model evocation and model revision must be put aside here to further examinations in induction, abstraction, and belief revision.

Chapter 4

Modeling with causal knowledge patterns

> What our eyes behold may
> well be the text of life but
> one's meditations on the text
> and the disclosures of these
> meditations are no less a part
> of the structure of reality
>
> Wallace STEVENS, *Three Academic Pieces – no. 1*

4.1 Causal decision theory, or: Of prisoners and predictors

Decision theory in general examines the rational principles guiding the decisions that aim at the attainment of one's goals. *Causal* decision theory does so by taking one's act's consequences into account – rationally choosing an option must be based on the available knowledge about the causal relations in the respective situation, so the argument goes. One of the principles taken to be a measure for rationality is the option of maximizing the utility of the outcome, i. e., by making the outcome equal or better than if one had chosen a different alternative for action. Probabilities and utilities are used to compute an act's expected utility such that – as emphasized in *causal* decision theory – dependence between acts and outcomes are understood as of causal (asymmetrical) character

– contrary to a merely *evidential* theory of decision making. A second principle of rationality dictates choosing the course of action that is better, regardless of what the world is like. This principle of dominance seems to be in conflict with the above-mentioned principle of expected-utility maximization in the curious case of Newcomb's paradox.

NEWCOMB, NOZICK, and a problem

Referring back to the physicist William NEWCOMB, who first formulated this dilemma for decision theory, Robert NOZICK elaborates on – as he calls it – Newcomb's problem, in which two principles of rational choice seemingly conflict each other, at least in the numerous renditions in the vast literature on this topic.[1]

In Newcomb's problem some human-like agent plays a game against some daemon predictor that influences the course of the game upon predicting his opponent's move. The agent may choose to take either one or two boxes in front of him – either box 1 only or box 1 and 2 together. In doing so he has no knowledge about the contents of the opaque box 1, but he can see one thousand dollars lying in box 2. If the daemon predicts that the agent will take only one box (i. e., box 1), he will put one million dollars in the opaque box 1. The daemon will put nothing in box 1, though, if he foresees the agent taking both boxes. The prediction is reliable, or as NOZICK introduces the predictor, "[o]ne might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about [the agent's] choice in the situation to be discussed will be correct."[2] Moreover, the agent has perfect knowledge of all these features of the decision game he finds himself in.[3]

The possible outcomes of the game are presented in table 4.1 where the rows stand for the agent's options, the columns partition the world in possible states, and each cell contains the sum our agent receives upon choosing an action in some state of the world.

[1] Cf. [Nozick 1969] for the original presentation of the paradox and [Weirich 2008] for an overview on various suggestions of how to solve the Newcomb case.

[2] Cf. [Nozick 1969, p. 114].

[3] Note that for reasons of simplicity this presentation of the Newcomb game situation slightly (but inessentially) differs from the way NOZICK originally presents it in [Nozick 1969].

	prediction: one-boxing	prediction: two-boxing
take box 1	\$ 1M	\$ 0
take box 1 and 2	\$ 1M + \$ 1T	\$ 1T

Table 4.1: Possible outcomes in Newcomb's problem for the options of taking box 1 only (taking boxes 1 and 2, respectively) and for correct and incorrect predictions made by the daemon.

Now, what makes Newcomb's case so problematic is the fact that the choice of action seems to depend on the choice of the principle one applies in *rationalizing* the situation. Two principles seem to be concurring candidates in reasoning about Newcomb's problem, which – although unrealistic – seems to trigger solid intuitions about the decision theoretic norms to be applied here.[4] The rationales of maximizing expected utility and of choosing dominating options are defined in the following.

Definition 4.1.1 (Maximum Expected Utility Principle)[5]
Among those actions available to a person, he should perform an action with maximal expected utility*.*
The expected utility $EU(A)$ of an action A yielding the exclusive outcomes $O_1, \ldots, O_n$ with probabilities $P(O_1), \ldots, P(O_n)$ and corresponding utilities $U(O_1), \ldots, U(O_n)$ is calculated by the weighted sum

$$\sum_{i=1}^{n} P(O_i) \times U(O_i).$$

Definition 4.1.2 (Dominance Principle)[6]
If there is a partition of world states such that, relative to it, action A weakly dominates action B, then A should be performed rather than B.
Action A weakly dominates *action B for person P iff, for each state of the world, P either prefers the consequence of A to the consequence of B, or is indifferent between the two consequences, and for some state of the world, P prefers the consequence of A to the consequence of B.*

[4] Nozick himself obviously put the story on the test bench: "I should add that I have put this problem to a large number of people, both friends and students in class. To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that theses people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly." – cf. [Nozick 1969, p. 117].

[5] This definition is adapted from [Nozick 1969, p. 118].

[6] This definition is adapted from [Nozick 1969, p. 118].

Let us take 'reliable' (as ascribed to the daemon's faculty of foreseeing future events) at face value and compute the expected utility for the outcome of each specific course of the game – the unit of the expected utility being dollars in our case. Assuming a reliable daemon basically amounts to saying that the act of taking one or both boxes and the prediction of this very act are highly correlated such that acts in states of the world with incorrect predictions receive a probability of 0, whereas matching acts and predictions receive the probability of 1. Table 4.2 shows the expected utilities for all four thinkable courses of the game with one option clearly to be preferred over all others: The agent should take only the opaque box and can then be certain of winning \$ 1M, which clearly supercedes the alternatives as *maximum expected utility.*

	prediction: one-boxing	prediction: two-boxing
take box 1	\$ 1M	\$ 0
take box 1 and 2	\$ 0	\$ 1T

Table 4.2: Computing expected utilities in the case of a perfectly reliable prediction yields the utility of \$ 0 for all cells representing incorrect predictions. Maximizing this expected utility amounts to choosing only box 1.

Pondering a different approach to maximizing the outcome of the game, Nozick tweaks the story a little: The predictor did make his prediction a week ago, and it is now the agent's turn to make up his mind and take either only the opaque box 1 or on top of that also the transparent second box 2, which contains one thousand dollars openly visible to the agent. The money is already there and will not be taken out of the boxes anymore after the agent has made a decision. So, regardless of the daemon's prediction, adopting the *principle of dominance* forces the agent to take both boxes – he will always end up with one thousand dollars more than if he had only taken one box. Taking both boxes even *strictly dominates* the act of taking only one box as can be read off table 4.1 by comparing an entry in the second line to the entry in the first line within the same partition of the world's states.

Obviously, the principle of maximizing expected utilities and the principle of dominance yield opposing recommendations to the deliberating agent. While standard *evidential* decision theory seems to lean towards *one-boxing* (taking an agent's act as a sign of what the prediction must have been), *causal* decision theorists clearly position themselves on the side of *two-boxing* (rejecting backward causation and understanding the agent's deliberate decision as cutting any connection between act and prediction). When Judea Pearl within his interventionist account of causal reasoning discusses *model-internal* observed acts and *model-altering* actions from outside, he also comes to reflect upon the conceptual difficulties hidden in Newcomb's problem:

> *The confusion between actions and acts has led to Newcomb's paradox (Nozick 1969) and other oddities in the so-called evidential decision theory, which encourages decision makers to take into consideration the evidence that an action would provide, if enacted. This bizarre theory seems to have loomed from Jeffrey's influential book* The Logic of Decision *(Jeffrey 1965), in which actions are treated as ordinary events (rather than interventions) and, accordingly, the effects of actions are obtained through conditionalization rather than through a mechanism-modifying operation like do*(x).[7]

When Pearl goes on by comparing the maxims of evidential and causal decision theory, he baldly comments in a footnote:

> *I purposely avoid the common title "causal decision theory" in order to suppress even the slightest hint that any alternative, noncausal theory can be used to guide decisions.*[8]

To reconcile the dominance principle with the expected-utility principle – and hence to dissolve the paradox in Newcomb's case – has been the aim of quite a few proposals, which nevertheless arrive at different conclusions.

Conditionals and causal graphs

In *A Theory of Conditionals* (1968) Robert Stalnaker suggests a formal framework for analyzing the truth of counterfactual statements (subjunctive conditionals) quite similar to Lewis' proposal sketched above – 'If A, then B' is assigned a truth value in accordance with the following informal condition:

> *Consider a possible world in which A is true, and which otherwise differs minimally from the actual world.* 'If A, then B' is true (false) just in case B is true (false) in that possible world.[9]

[7] Cf. [Pearl 2009, p. 108].

[8] Cf. [Pearl 2009, p. 108, footnote 1].

[9] Cf. [Stalnaker 1968, p. 169].

The subjunctive connective '>' is subsequently equipped with the more formal semantical rules

$$\begin{aligned}&A > B \text{ is true in } \alpha \text{ if } B \text{ is true in } f(A,\alpha) \text{ and}\\ &A > B \text{ is false in } \alpha \text{ if } B \text{ is false in } f(A,\alpha),\end{aligned}$$

where α is a possible world, the *base world*, and $\beta = f(A,\alpha)$ represents the *selected world* minimally differing from the actual world in which B is evaluated (with f being the selection function operating on a suitable similarity ordering of possible worlds).

Now, in his *Letter to David Lewis* (1972) STALNAKER suggests a way of calculating expected utilities in the Newcomb problem that uses probabilities of subjunctive conditionals instead of standard conditional probabilities.[10] The expected utility of some action A would then be computed the following way:

$$EU(A) = \sum_{i=1}^{n} P(A > S_i) \times U(A \,\&\, S_i),$$

where n signifies the amount of states S the world is partitioned into, i. e., $n = 2$ for the two possible predictions 'one-boxing' ($i = 1$) and 'two-boxing' ($i = 2$). As STALNAKER argues, the agent's action does not *cause* the daemon's prediction made in the past, and hence the probability of the conditional equals the probability of the prediction alone. But this sets all probability terms in the sum formula above to equal values – the utilities can just be read off the corresponding cells in table 4.1. Two-boxing's expected utility will always be greater then one-boxing's expected utility. Following Robert STALNAKER's suggestion of interpreting the involved probabilities *causally*, the maximization of expected utility and the dominance principle recommend taking the same action: two-boxing.

Applying causal decision theory to Newcomb's problem has been criticized by many authors – mainly because it yields the counter-intuitive recommendation of taking both boxes, which nevertheless remains as the only rationally explained choice given the circumstances of Newcomb's problem with decisions screening off acts from any previous events, as causal decision theorists claim. In his seminal book *The Foundations of Causal Decision Theory* James JOYCE clearly states his position on the issue:

[10] Cf. for this and the following [Weirich 2008, sect. 2.2].

> *When the evidential and the causal import of actions diverge [. . .], the evidential theory tells decision makers to put the pursuit of good news ahead of the pursuit of good results. Many philosophers, I among them, see this as a mistake. Rational agents choose acts on the basis of their* causal efficacy*, not their auspiciousness; they act to* bring about *good results even when doing so might betoken bad news.*[11]

While, e. g., David Lewis and Brian Skyrms in their accounts mark attainable situations by building causal information into states of the world and thereby reconcile the above otherwise diverging principles of rational choice in the recommendation of two-boxing, Ellery Eells in his considerations arrives at the same conclusion without drawing on the notion of causality. He claims that mere reflection on the available *evidence* will force the agent to rationally go for both boxes – even more direct without the recourse to any causal theory. Quite in this line of reasoning Richard Jeffrey also eliminates any hint of a causal nexus between the events in Newcomb's problem for the sake of a less metaphysically charged analysis. Pondering the Newcomb case Jeffrey seems to oscillate between one-boxing and two-boxing to later arrive at the conclusion that the story, presented this way, is a somehow *illegitimate* decision problem with the freely deliberating agent not capable of freeing his decision from being correlated with the predictor's prediction.[12] Terry Horgan and Paul Horwich take the Newcomb plot at face value and promote one-boxing, simply because one-boxers ultimately take more money home, as the story is told. Paul Weirich diagnoses dryly: "The main rationale for one-boxing is that one-boxers fare better than do two-boxers. Causal decision theorists respond that Newcomb's problem is an unusual case that rewards irrationality. One-boxing is irrational even if one-boxers prosper."[13]

Having developed his ranking theory as a tool for epistemology and causal analysis,[14] Wolfgang Spohn positions himself on the side of causal (vs. evidential) decision theory and had been a strong advocate of two-boxing for a long time before he started "Reversing 30 Years of Discussion" by presenting an elaborate argumentation "Why Causal Decision Theorists Should One-Box."[15]

[11] Cf. [Joyce 1999, p. 146].

[12] Cf. e. g. [Joyce 2007].

[13] Cf. [Weirich 2008, sect. 2.5].

[14] Cf. Spohn: Ranking Theory (forthcoming).

[15] The quotations here refer to the title of [Spohn (forthcoming)].

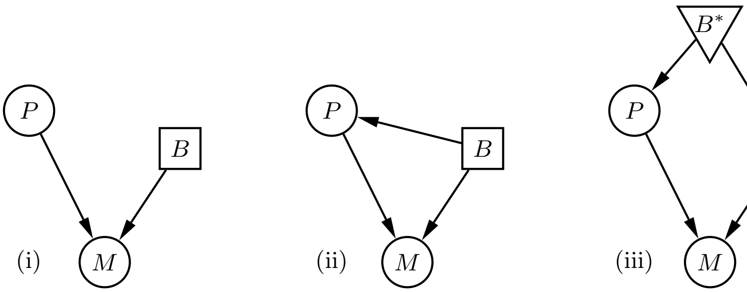

Fig. 4.1: Wolfgang Spohn discusses the usual *manipulated (mutilated) causal graph* (i) employed by causal decision theorists for the analysis of the Newcomb problem, the *decision graph* (ii) for the same situation, and the *reflexive decision graph* (iii) augmented by the decision node B^*.

Figure 4.1 illustrates the golden thread in Spohn's chain of reasoning. Time evolves from top to bottom in all three diagrams. The left diagram (i) shows the standard rendition used by causal decision theorists for the analysis of the Newcomb problem – this *mutilated causal graph* contains the node P representing the daemon's prediction as the first event in time before *action* node B (representing the agent taking one or two boxes) and the bottom node M (for monetary outcome). The diagram is *mutilated* quite in agreement with Pearl's interventionist framework: The hypothetical local surgery, i. e., the intervention on B, prunes any arrows possibly pointing towards B, thereby freeing this node from the influence of any other node in the model and making the corresponding variable an exogenous one. The course of action can now be chosen on the basis of this *decision graph*, in which the *wiggled* variable is graphically represented by the square node. This rendition follows the two decision theoretic principles highlighted by Spohn in this context: "acts are exogenous" and – derived from the first – "no probabilities for acts." Of course, Spohn's *acts* have to be interpreted as Pearl's *actions* (i. e., *acts* in *mutilated* models). Whatever the connection between nodes P and B might have been in some graphical rendition of the original causal relations understood as representing the Newcomb plot (e. g., with P as a direct cause of B), graph (i) in figure 4.1 represents the variables' dependencies once the agent deliberately takes action. P and B are d-separated (by the collider in $P \rightarrow M \leftarrow B$), which makes the choice of taking both boxes rational – whatever has been put into the boxes (based upon the prediction early in the game) will not become less by choosing either one or, alternatively, two boxes (later in the game).

Spohn declares himself dissatisfied with this analysis and brings up the mind-bugging questions about the reliability of the daemon, again:

> *What about the remarkable success of the predictor that suggests that given you one-box it is very likely that she will have predicted that you will one-box, and likewise for two-boxing? How do they enter the picture? They don't. [Causal decision theorists] do not deny them, but they take great pains to explain that they are not the ones to be used in practical deliberation calculating expected utilities; and they diverge in how exactly to conceive of the subjective probabilities to be used instead.*[16]

If the causal graph contained one more arrow from B to P, making the agent's action a direct cause of the daemon's prediction (as illustrated in figure 4.1, diagram (ii)), we would inevitably introduce backward causation into the analysis. Spohn wants to avoid this but interprets graph (ii) as the *decision-guiding pattern* which the agent uses to choose between alternative actions – in Spohn's terms: the ordinary decision graph for Newcomb's problem. How are the causal relations laid out, however? If neither the prediction causes the agent's act nor this act can cause the daemon's prediction, we have to infer the existence of an earlier third event as a common cause of both P and B – quite in accordance with Reichenbach's *Common Cause Principle.* Spohn's straightforward suggestion is to understand the *decision situation* the agent finds himself in as the common cause in question. This decision situation B^* (as introduced into graph (iii) in figure 4.1) might consist of all the agent's beliefs, prior knowledge, or rational principles the agent may not even be aware of (the daemon is, however) but which he will without fail employ in deciding about his strategy B when standing before the two boxes. In particular, B^* also contains the full ordinary decision-guiding pattern (ii), which makes graph (iii) a *reflexive decision graph* containing a reduced version of itself.[17] Making this move, Spohn openly rejects the "acts are exogenous" principle. An agent's strategic deliberation about alternative courses of action does *not* decouple the act from past or future events – he might, quite on the contrary, make his deliberations depend on (i. e., graphically speaking, link them to) *predecessor nodes in the diagram.* He might, on top of that, also be aware of the probabilities of different actions he may choose from, knowing what he *usually does* or intentionally *avoids in normal cases* etc. There might be *probabilities*

[16] Cf. [Spohn (forthcoming), p. 4].

[17] Spohn gives clear rules for the step-wise reduction of a reflexive decision graph to its ordinary counterpart possibly containing backward links – cf. [Spohn (forthcoming), sect. 3].

for the agent's act, after all. Querying SPOHN's reflexive decision graph on the ground of all these considerations ultimately yields the recommendation of one-boxing – after reflecting on the current situation (in B^*), the rational agent must come to the unequivocal conclusion that deciding to one-box and acting accordingly simply maximizes the utility of his act B.

Let us compare SPOHN's analysis with PEARL's causal maxims, once more. The ordinary decision graph (as displayed in figure 4.1.ii) fully complies with what PEARL would devise for strategic reasoning, i. e., a graph that simulates possible outcomes of hypothetical interventions. Setting B tells us the value of M. B is an exogenous variable such that the "acts are exogenous" principle is adhered to – act and action amount to the same consequence in this case. The evidential and the causal approach perfectly concord in this diagram, were it not for the directed backward edge $B \rightarrow P$. This is the reason for PEARL to think directly in terms of the mutilated graph (given in figure 4.1.i) and for SPOHN to call diagram 4.1.ii not *causal* but *reduced, ordinary decision graph.* In the further step of construing the reflexive decision graph 4.1.iii, SPOHN must reject the "acts are exogenous" principle and convincingly argues for his case: The hypothetical intervention on the variable B must not be performed within the reflexive decision graph. This graph makes explicit what it means for the agent to be rational, i. e., he acts on his knowledge, principles, and rational considerations given in B^*. Pruning the link $B^* \rightarrow B$ would make the agent plainly *irrational* and *ignorant* of his own situation, since the deliberation process is *pushed into the model.*

Technical answers to questions about how to properly reduce reflexive decision graphs to their ordinary, structural counterparts can all be found in SPOHN's explications. Conceptual questions remain, however.[18] Firstly, the introduction of a common cause for B and P essentially adds to the Newcomb's story the idea of being (perhaps physically determinately) *pre-disposed.* In a way, this metaphysically overloads the already artificially construed plot with another element just by drawing on REICHENBACH's principle of the common cause. Moreover, it forces SPOHN to set apart the agent's inclinations to take certain actions from the acts themselves. Decision making in the game is consequently *re-interpreted as only discovering one's previously fixed inclinations* (where discovery is

[18] I am thankful to Wilken Steiner for valuable discussions of Newcomb's problem and SPOHN's treatment of it.

not something brought about actively, e. g., such that it would manifest itself in hypothetical test interventions, but simply a feature of persistent rationality becoming *evident*). This rendition seems very far from the much more intuitive interventionist framework, which merely requires the agent to bear a *confined mini laboratory* in his head and turn the knobs therein – knowledge about the mechanisms will yield unique virtual outcomes and guide decision making. Nevertheless, SPOHN's complex reflexive decision graph does rest in its core on the very simple ordinary reduced decision graph (figure 4.1.ii) to which the whole burden of explanation is shifted, which shall be looked at more closely in the following. What can be the content of this reduced graph, after all? If the link $B \rightarrow P$ is dismissed as causal relation, of what nature can it be? If it, on the other hand, does stand for some hidden causal connection and is dismissed as backward causation, it must represent a causal link through some obscure common cause. If this common parent node of both P and B is the decision situation again – just as in the reflexive graph on the meta level – analysis enters an infinite regress at this point. Only the interventionist approach could prevent this from happening by pruning $B \rightarrow P$, but then this would already apply on the upper level in the reflexive decision graph and conflict with SPOHN's final conclusion. If the supposed common cause in figure 4.1.ii is interpreted as some irreducible obscure past event or state whose existence just has to be acknowledged and whose link to B shall not be interrupted, then how would it be possible to perform hypothetical test interventions on this very node to virtually maximize the outcome? If reflecting on this graph ultimately comes down to just *observing* the propagation of values, then, one has to conclude, SPOHN's suggestion is constrained to stay within evidential reasoning.

Foreseeing acts, foreseeing actions[19]

What the backward link $B \rightarrow P$ in graph 4.1.ii can possible mean shall in the following be made explicit within the CKP framework, thereby ideally revealing more about the nature of the paradox and hopefully illuminating some more features of how we reason with (non-)causal knowledge. The causal knowledge pattern in figure 4.2 traces the story of Newcomb's problem by only referring to the events that actually are in the narration. The problem is not treated by tweaking the story but by choosing a framework fit to accommodate all relevant concepts.

[19] I have greatly benefitted from discussing Newcomb's problem and the concept of rationality with Olivier Roy for whose comments on this section I am very thankful.

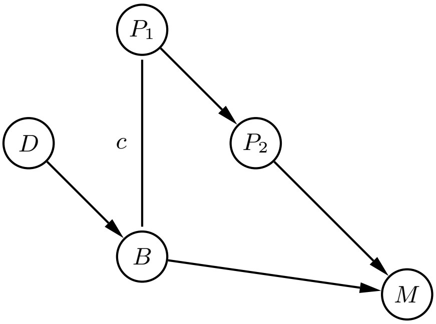

Fig. 4.2: Newcomb's problem with the act of taking one or two boxes (B) deterministically connected to the daemon's reliable prediction ($P1$) by an epistemic contour (c) in this causal knowledge pattern.

Our human-like agent deliberates about the situation he finds himself in and decides what to do (D), namely if he takes one box or both boxes (B). The daemon predicts what the agent will do (P_1) and prepares the boxes accordingly (P_2). The monetary outcome (M) should finally reward the rational agent. Time evolves from top to bottom in the diagram.[20] The vertical positioning of P_2 is inessential for the analysis of the situation (P_2 could as well come after B if the game is set up in a way that the agent only writes down his choice on a sheet of paper secretly in step B). The daemon's prediction together with its reliability is interpreted in this causal knowledge pattern as an undirected 1-1 relationship. Neither would we say that the agent's act genuinely causes the prediction of this very act, nor does it sound right to say the prediction causes the predicted event.[21] But there is more in the pattern: B is not directly linked to the daemon's preparation of the boxes P_2 – this connection is mediated by the prediction P_1, which has direct causal influence on P_2 in turn. This is quite in agreement with SPOHN's analysis that the causal structure of the Newcomb problem should exhibit some node previous to both players' acts in the game that at the same time takes care of the bidirectional transfer of belief. P_1 and P_2 are separated in the causal knowledge pattern for this very reason. On the other side, D (the human-like agent's decision situation) and B (his concrete move in the game – either taking one or both boxes) are separated, as well, to

[20] Note that this diagram graphically reverses SPOHN's rendition where time evolves from bottom to top.

[21] Moreover, as is argued here, drawing on REICHENBACH's *Common Cause Principle* for an explication of 'prediction' is precisely a source of counter-intuitive inference.

disentangle conceptually what it means for the agent to spontaneously and possibly unforeseenly *change his mind.* This is a much-discussed issue in the literature and does pose additional problems if the modeling allows for the agent changing his mind and the daemon's prediction referring to the 'wrong' decision. Not so in the suggested causal knowledge pattern, which links the prediction P_1 to the agent's final act B however often he may have made up or changed his mind before actually taking only one or, after all, both boxes. In other words, pondering courses of action must focus on B bearing the whole burden of explanation in the process of finding the best strategy for the maximization of the outcome. This is exactly as Nozick tells the story.

The modeling does not draw on the insertion of backward links that would signify backward causal flow. Nevertheless, *information* is transferred back in time along the epistemic contour c, thereby formally grasping the very meaning of 'prediction.' c will not get cut off by any local surgery of the graph. By suitably applying hypothetical test interventions the following contents can be read off the causal knowledge pattern – quite in accordance with intuition:

- The agent's decision (D) causes his act (B) – in general: any causal history of B naturally influences the agent's act causally;
- the agent's decision (D) is also interpreted as causing the daemon's peculiar prediction (P_1) and thereby also as causing the daemon's particular move in the game (P_2);
- intuition also conforms with the claim that the agent's taking one or two boxes (B) causes his antagonist's preparation of the boxes – the predictor *reacts* to (B), after all;
- nevertheless, the agent's act (B) does *not cause* its own peculiar prediction (P_1) but *determines it uniquely* and – looking at the pattern from above – *simultaneously* though *backwards through time.*

Now, especially the last point reveals the core of the paradox and localizes the difficulties in reasoning about the causal relations involved. Any attempt of solving the artificial plot of Newcomb's problem hinges on the question how to embed the concept of reliably predicting future events into the formal analysis (if such an analysis is not denied in the first place exactly because of the fictional character of the narration). The causal knowledge pattern above presents the prediction as the very thing it is – an *image* of the agent's act. *Backward links* are excluded from this rendition while querying the pattern does yield *indirect causal* *claims referring back across time.* This interpretation would of course not stand physically ontologically based scrutiny, but it conforms with our concepts of prediction (of future events) and reaction (to facts just learned of). How pieces of knowledge are organized and beliefs propagated is shown in the causal knowledge pattern devised here. Obviously, the "acts are exogenous" principle insisted on by Judea Pearl is relativized in applying causal knowledge patterns to problems of decision theory. The epistemic contour c is not deactivated by intervening on B, while the one directed edge $D \twoheadrightarrow B$ is removed by the external action $do(B = b)$ – quite in Pearl's sense B and P_1 become jointly exogenous (in accordance with definition 3.3.2). To sort the terms involved here: The *act* B becomes *exogenous* by virtue of the *action* $do(B = b)$, which is itself *external.*[22] If the prediction of events is formalized within a model (a causal knowledge pattern, respectively), *foreseeing acts* can be made explicit, while *foreseeing actions* cannot be given graphical expression. Reflecting on the Newcomb situation and performing hypothetical manipulations on the basis of integrating causal and non-causal knowledge finally guides the agent (who is aware of the setting) towards the correct decision. Resorting to reflexiveness is not necessary for virtually maximizing the outcome. The conclusion must be one-boxing.

As a last remark in this part on causal decision theory, David Lewis shall be mentioned here once more. He examines another paradoxical puzzle of strategic thinking and finds in 1979 that the "Prisoners' Dilemma Is a Newcomb Problem", too.[23] The story in this particular dilemma shall be outlined briefly. Two suspects are caught by the police, that do not have sufficient evidence for conviction and therefore question the prisoners separately and (also separately) promise immediate release if the prisoners betray the respective other prisoner by confessing. However, if both confess, each serves a sentence of three months – in case both remain silent, each serves one month. Table 4.3 summarizes the situation compactly. If prisoner A applied the *principle of dominance* to his situation, he would of course confess, thereby always being off better than if he remained silent. If both prisoners think alike in this respect, however, they will be doomed to a sentence of another three months in prison. This is what makes the situation a strategic dilemma: Attributing the same (degree of) rationality to both prisoners does not entail the best outcome. If they include in their deliberations the ascription

[22] For clarification: *exogenous* remains a model-internal property of nodes (i. e., variables, respectively), whereas *external* marks transformations of causal structures.

[23] The quotation refers to the title of [Lewis 1979].

of *like-mindedness* to their fellow inmate, both of them should remain silent. If this ascription is reliable enough (or even deterministically certain), e. g., because of some commitment to the same gang code, then the prediction in Newcomb's problem and this theoretical simulation (the ascription) in the prisoners' dilemma essentially amount to the same thing – "[i]nessential trappings aside, Prisoners' Dilemma is a version of Newcomb's Problem, *quod erat demonstrandum.*"[24]

	B stays silent	B confesses
A stays silent	Each serves 1 m	A serves 1 y, B goes free
A confesses	A goes free, B serves 1 y	Each serves 3 m

Table 4.3: Each of the prisoners could go free or serve a sentence of one month, three months, or a year – depending on their strategic decisions.

A *common* causal knowledge pattern might be used to capture all (non-)causal relations as in the above rendition of Newcomb's problem – quite naturally and without introducing further metaphysical assumptions about possible background variables. In fact, tilting the time axis in figure 4.2 by 90 degrees (such that time evolves from left to right) yields the skeleton of the prisoners' plot (of course, D and P_2 are particular ingredients of Newcomb's problem and inessential for the current examination). c represents the mutual ascription of like-mindedness of both prisoners, who *must* decide to cooperate during their simultaneous (but separate) questioning to achieve the joint best result. May the Newcomb case be some fictional construction, Lewis makes the case for analyzing the prediction of future events and the ascription of like-mindedness to one's antagonist in terms of the same underlying pattern:

> *Some have fended off the lessons of Newcomb's Problem by saying: "Let us not have, or let us not rely on, any intuitions about what is rational in goofball cases so unlike the decision problems of real life." But Prisoners' Dilemmas are deplorably common in real life. They are the most down-to-earth versions of Newcomb's Problem now available.*[25]

[24] Cf. [Lewis 1979, p. 239].

[25] This final quotation borrows the concluding paragraph from [Lewis 1979, p. 240]. I agree with Lewis on the point that situations of strategic deliberations of the kind exemplified here are "the most down-to-earth versions of Newcomb's Problem" – because there is *nothing more to know* than already said – in contrast to cases of so-called *medical Newcomb problems* where research might in most cases yield additional information and knowledge about true common causes whose influence would indeed be rendered void by free deliberation/active intervention.

4.2 Meaningful isomorphisms

Augmenting standard Bayes nets by adding epistemic contours might at first seem reducible again, as sketches of causal models (in Pearl's sense) are refined and incorporated into scientific bodies of explanation or into strategic groundwork for policy making. The postulate of only admitting *extensionally distinct* events in the analysis poses problems, though, as soon as *intensional distinction* becomes necessary or different approaches towards measuring the same phenomenon need to be emphasized and unified in one frame. Epistemic contours pave the way for such enhanced modeling. As an additional structural component in causal knowledge patterns these 1-1 functions

- represent non-directional knowledge,
- are capable of bridging frameworks of description,
- deterministically transfer knowledge *simultaneously*,
- mark variables that cannot be decoupled (i. e., *set* separately),
- and are not deactivated by interventions, in particular.

Synonymy, strict semantical and conceptual dependencies, or logical and mathematical relations present problems for the standard Bayes net approach and are purposely excluded (by the expert modeler) from integration into Pearl's causal models. Representing such relationships as directed edges would subject them to possible local atomic surgeries, which would immediately yield paradoxical inferences. Including non-causal knowledge into the analysis, where information about isomorphic relations is available, does essentially support causal inference, though, and can be computed consistently in the framework of causal knowledge patterns – exemplary cases shall be considered in the following.

Synonyms

Two synonyms refer to the same phenomenon or to the same observation when they denote events, and are therefore modeled as just one node in the graph of a causal model representing just one variable that can be labeled differently but for which the specification of the method of measurement fixes the extensional meaning. Epistemic contours can be used to accommodate more than one *signifiant* of one and the same event in the causal knowledge pattern, thereby bridging jargon, levels of specialization, frameworks of distinct interests, differing aspects of the same research object, or intended systems of neighboring theories –

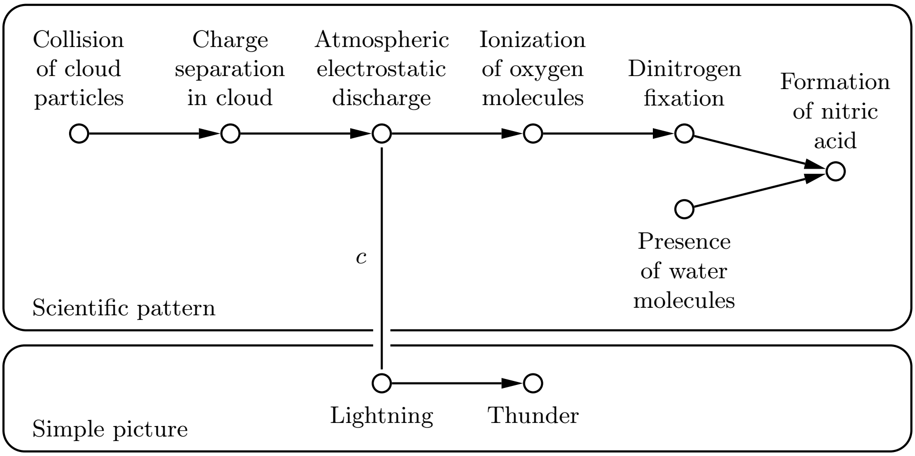
Collision of cloud particles
Charge separation in cloud
Atmospheric electrostatic discharge
Ionization of oxygen molecules
Dinitrogen fixation
Formation of nitric acid
Presence of water molecules
c
Scientific pattern
Lightning
Thunder
Simple picture

Fig. 4.3: Frameworks of differing specialization are bridged by the epistemic contour *c* representing an isomorphic relation between *atmospheric electrostatic discharge* and *lightning*.

perhaps in the process of synthesizing.[26] Figure 4.3 explicates this use of epistemic contours: Two originally separate causal models are connected through an epistemic contour *c*. The upper framework shows the fine-grained model of some bio-chemist who wants to trace the formation of nitric acid in the atmosphere. In the graph of his causal model the directed edges represent causal mechanisms across the disciplines (quite in the sense of Jon Williamson, see also p. 90). Electro-magnetic phenomena are linked to chemical processes in this scientific pattern, which does not pose any difficulty to the proponent of the epistemic account of causation. The lower causal model in figure 4.3 illustrates the simple picture of how lightning and thunder might be arranged in some naive (regularity) account (which possibly might not even be embeddable into some more fine-grained scientific rendition). These two frameworks do not talk about the same sets of phenomena, nor do they lie on the same

[26] Such epistemic bridges might lead to the discovery of more bridges between the respective frameworks. On the other hand, formalizing assumptions about the existence of possible isomorphisms in this context can be of help for refuting these very assumptions and for clarifying blurred differences between originally distinguished theoretical terms or cross-disciplinary "false friends." Pragmatic examples can be found in pedagogical theories: One of the principles of neuro-didactics states that learning is more effective if existing prior knowledge is activated. How new information is connected with/augmented by/embedded in such prior knowledge or how it is to be re-ordered/aligned can be formalized in recourse to twin CKPs.

level of specialization, and neither do their modelers "pursue common interests." Nevertheless, as in this example, if the modelers agreed on a meta level to the fact that one of the phenomena their models are about always *co-occurs*, the epistemic contour c shows how the different frameworks can be aligned side by side for mutual information exchange. *Thunder* can subsequently be explained through the chain rooting in *Collision of cloud particles*.

Logical and mathematical dependencies

Any association between the description of two events that comes in the form of a parametric equation can in principle be represented by an epistemic contour to facilitate causal reasoning across methods of measuring in cases where the introduction of a common cause seems far-fetched, artificial, or controversial. Scale translations (on the same level of measurement) can thus be given formal expression within causal knowledge patterns, as well as unit conversions (e. g., of currencies) and geometric transformations. Entities on both sides of such an epistemic contour extensionally 'measure one coin' but intensionally emphasize 'its different sides' – especially when these different characterizations invoke differing causal claims. E. g., temperature (in ranges) and color (in name codes) of metal can be aligned side by side but might each be linked to different effects or even to different causes – according to the experimental setup. But knowledge about either tells the experimenters more about the effects across the deterministic undirected edge in the causal knowledge pattern.[27] Inversely proportional behavior is yet another candidate for modeling by epistemic contours (if the equation contains the total quantity as a parameter). When considering a certain country, the percentage of sealed soil determines what remains within the countries boundaries as a source of biomass and as a storage of nutrients, substances, water, etc. Neither causes the other – intervening on one *determines* the opposite. Both might be roots of differing causal chains, though, according to how knowledge is organized in some causal knowledge pattern, i. e., within some epistemic subject.

[27] Transferring knowledge about the temperature over to knowledge about the color might in addition require knowledge about the properties of the piece of metal under consideration. These properties are part of the parametric formulation of the isomorphism represented by the epistemic contour – they can be seen as a kind of ceteris paribus conditions if the experimental setup examines *normal metal* under *normal conditions*. Parameters that can be modified through surgeries in the model are excluded from the framework devised here but might motivate a possible further extension.

Semantical and conceptual dependencies

Epistemic contours are also applicable to cases where interlevel causal claims are to be rendered explicit – as one example of representing *semantical dependencies* by the introduction of an epistemic contour. When Carl Craver and William Bechtel talk about "levels of mechanisms," they have in mind what Pearl would call zooming into (or zooming out of) a given model to present the situation under consideration in a more fine-grained (or, respectively, in a more coarse-grained) variant.[28] In their sense, a mechanism is a process as structured in a given causal model (in Pearl's sense) and can as such be related to a more detailed or less detailed structure describing the same situation:

> *[L]evels of mechanisms are a species of compositional, or part-whole, relations. In contemporary debates about reduction and interlevel causation, it is common for authors to talk about 'levels of aggregation,' 'levels of organization,' 'levels of complexity,' and 'mereological levels.' Such descriptions apply to levels of mechanisms as well. Higher levels of mechanisms are aggregated (i. e., built up from) or composed from parts that are organized into more complex spatial, temporal, and causal relations.*[29]

In their discussion of both top-down and bottom-up causation they claim that causation across levels is described by what they call *mechanistically mediated effects*, which "are hybrids of constitutive and causal relations in a mechanism, where the constitutive relations are interlevel, and the causal relations are exclusively intralevel." They maintain further that the "[a]ppeal to top-down causation seems spooky or incoherent when it cannot be explicated in terms of mechanistically mediated effects."[30] In saying that, Craver and Bechtel refer back to David Lewis who fixes intuitions about the distinctness of cause and effect in his influential *Causation as Influence*:

> *C and E must be distinct events – and distinct not only in the sense of nonidentity but also in the sense of nonoverlap and nonimplication. It won't do to say that my speaking this sentence causes my speaking this sentence or that my speaking the whole of it causes my speaking the first half of it; or that my speaking causes my speaking it loudly, or vice versa.*[31]

[28] See also [Machamer *et al.* 2000] for a detailed overview of the concept of *mechanism* in the sciences.

[29] Cf. [Craver & Bechtel 2007, p. 550].

[30] Cf. [Craver & Bechtel 2007, p. 547].

[31] This quotation is from the unabridged version of [Lewis 2000] as reprinted in [Collins *et al.* 2004].

The least disputed cases of genuine top-down or bottom-up causal relations, where changing the system prima facie causally influences (at least one of) its parts or vice versa, are figures of pars-pro-toto and totum-pro-parte reasoning. These cases can straightforwardly be translated into the framework of causal knowledge patterns by marking the level transition with an epistemic contour that represents the *constitution* (in Craver's and Bechtel's words) or the (mutual) *constraint* (as suggested by Max Kistler as a conceptual refinement in his comment[32] on Craver and Bechtel). Two examples shall be considered in the following.[33]

The *general's heart attack* can be understood as a genuine pars-pro-toto figure, where the defect of the heart as part of the general's body *determines* the general's state of being alive or dead. The biological intra-organism level and the level of the organism as a whole are linked here through 1-1 functional information exchange, not by any causal process. The general remains alive as long as (and only as long as) his heart continues beating. This isomorphism is expressed in the pertaining causal knowledge pattern as a simple epistemic contour, consequently. The story of *Ignatius and his hotdogs* on the other hand presents a clear totum-pro-parte case where Ignatius maneuvers his hotdog cart to the corner of the street to market his hotdogs there. Craver and Bechtel ask: "What caused the hotdogs (and the molecules in the hotdogs, and the atoms comprising the molecules, and so on) to arrive at the corner? Ignatius." Although on different levels, Ignatius' pushing the cart causes the cart to move – and its parts and contents along with it. They are simply "carried along for the ride."[34] An epistemic contour marks the 1-1 non-causally interrelated positions of cart and hotdogs. Making Ignatius move the cart ("using a $do(\cdot)$-operation on him") will ultimately move his goods as well, which licenses the above claim about Ignatius as causee on the lower level, too.

4.3 Epistemic contours and the Markov assumption, revisited

Introducing epistemic contours as bridges of non-directional knowledge transfer into structures of causal reasoning was possible because these structures were understood as schemata of knowledge organization

[32] See [Kistler 2010] for Kistler's explications.

[33] For this and the following cf. [Craver & Bechtel 2007, pp. 557 ff.].

[34] Cf. [Craver & Bechtel 2007, p. 558] for both quotations.

shaped by the structuring power of the epistemically interpreted principle of causality. The interventionist characterization of causation could be maintained by explicitly overriding the requirement that all variables have to be modifiable separately – this is not the case for variables of an EC clique, where intervening on one distinguished variable strictly forbids any opposing intervention on other variables in the same EC clique. Epistemic contours precisely postulate non-interruptibility of the deterministic functional connection they stand for. Their integration clashes with the Markov assumption, because such epistemic contours might contradict the assignment of values through the causal mechanisms at work. By introducing the epistemic principle of *explanatory dominance* consistency is taken care of, again. Explicating causal with closely intertwined non-causal knowledge in one unifying network makes the underlying assumptions concrete, transparent, and operative on the surface. Finally, the formal framework of causal knowledge patterns offers a means for consistently deriving higher-level causal claims from basal data of different types and might offer insight into dialectics of communication and processes of learning. Adding structural ingredients and intensional markers of default knowledge to standard Bayes net causal models along with the rules for implementation counters Pearl's skepticism:

> *The Markovian assumption [...] is a matter of convention, to distinguish complete from incomplete models. By building the Markovian assumption into the definition of complete causal models [(def. 2.7.1)] and then relaxing the assumption through latent structures [(see p. 66)], we declare our preparedness to miss the discovery of non-Markovian causal models that cannot be described as latent structures. I do not consider this loss to be very serious, because such models – even if any exist in the macroscopic world – would have limited utility as guides to decisions. For example, it is not clear how one would predict the effects of interventions from such a model, save for explicitly listing the effect of every conceivable intervention in advance.*[35]

What it means for an event A to cause some distinct event B is explained above in terms of doxastic structures – how one predicts the effects of interventions in causal knowledge patterns as compounds of causal and non-causal knowledge is described in interventionist vocabulary by extending the rules for the $do(\cdot)$-operation. Drawing upon bodies of epistemically organized relations precisely guides the epistemic subject to decisions that might not be explainable as straightforwardly from plain Bayes net structures.

[35] This quotes [Pearl 2009, p. 61] without footnotes.

Appendix A

Random variables (stochastic variables)

The Σ**-random variable** V **over** Ω is an $\langle \mathcal{F}, \mathcal{X} \rangle$-measurable (total) function of the outcome of a statistical experiment, mapping possible outcomes to values (realizations, e. g., real numbers). The meaning of the random variable lies in the linkage between the outcome of an experiment and its mathematical representation:[1]

$$V : \Omega \to \Sigma$$

such that

$$\begin{array}{llll} V(\omega) = \sigma & \text{or in short:} & V = \sigma & \text{or also} \\ V(\omega) = v & \text{or in short:} & V = v & \text{(as commonly used),} \end{array}$$

with a probability space $\langle \Omega, \mathcal{F}, P \rangle$ and an observation space $\langle \Sigma, \mathcal{X} \rangle$, as explained in the following.

The **probability space** is a triple $\langle \Omega, \mathcal{F}, P \rangle$, where Ω is the sample space $Dom(V)$ of a random process (sometimes also S for 'sample space' or U for 'universe'), and $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is the set of events (where each event is a set containing zero or more outcomes), the event algebra, a σ-algebra (σ-field or also borel field) over the set Ω, by definition a nonempty collection of subsets of Ω (including Ω itself) that is closed under complementation and countable unions of its members.

[1] Cf. for this and the following e. g. [Fahrmeir *et al.* 2000].

E. g., for a given sample space $\Omega = \{a, b, c, d\}$, $\mathcal{F}$ might be the subset of $\mathcal{P}(\Omega)$ specified as $\{\varnothing, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$. If we have $\mathcal{F} = \mathcal{P}(\Omega)$ in the case of a finite sample space Ω, V is always measurable.

The probability function (the measure) $P : \mathcal{F} \to [0, 1]$ defines a measure over $\mathcal{F}$, satisfying the Kolmogorov axioms:

(K1) $P(A \in \mathcal{F}) \geq 0$,

(K2) $P(\Omega) = 1$,

(K3) $P(\bigcup_i A_i) = \sum_i P(A_i)$ for any countable sequence of pairwise disjoint (i. e., mutually exclusive) events $A_1, A_2, \ldots$ $(\in \mathcal{F})$.

The **measurable observation space (state space)** $\langle \Sigma, \mathcal{X} \rangle$ typically couples the real numbers $\mathbb{R}$, the integers $\mathbb{N}$, or any finite set of values with a suitable σ-algebra $\mathcal{X}$ over Σ with $\mathcal{X} \subseteq \mathcal{P}(\Sigma)$:

> for real-valued (continuous) random variables: $V : \Omega \to \mathbb{R}$;
>
> for discrete random variables yielding values of countable sets, e. g., of the set of natural numbers: $V : \Omega \to \mathbb{N}$;
>
> and for dichotomous random variables: $V : \Omega \to \{0, 1\}$.

In this representation **events** are subsets of some sample space Ω, which are also often written as propositional formulas containing random variables, e. g., $\{\omega \,|\, u_1 \leq V(\omega) \leq u_2\}$, or shorthand: $\{\omega \,|\, u_1 \leq V \leq u_2\}$. For the sample space $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{n-1}\}$ with size n, the singletons $\{\omega_0\}, \{\omega_1\}, \ldots, \{\omega_{n-1}\}$ are called 'atomic events.' Events $A \subseteq \Omega$ are determined by the random variable V, e. g., through formulations of the following kind:

$$\begin{aligned}
\{V = v\} &:= \{\omega \in \Omega \,|\, V(\omega) = v\}, \\
\{V \leq v\} &:= \{\omega \in \Omega \,|\, V(\omega) \leq v\}, \\
\{u_1 \leq V \leq u_2\} &:= \{\omega \in \Omega \,|\, u_1 \leq V(\omega) \leq u_2\}, \\
\{V \in I\} &:= \{\omega \in \Omega \,|\, V(\omega) \in I\},
\end{aligned}$$

where I is some specific interval.

Example (Tossing a coin twice)
A double coin toss may be modeled in the following probability space $\langle\Omega,\mathcal{F},P\rangle$*:*

- Ω *is the set of four possible outcomes:* $\{\langle \mathit{Heads}, \mathit{Heads}\rangle, \langle \mathit{Heads}, \mathit{Tails}\rangle, \langle \mathit{Tails}, \mathit{Heads}\rangle, \langle \mathit{Tails}, \mathit{Tails}\rangle\}$;
- $\mathcal{F} = \mathcal{P}(\Omega)$*;*
- *for a fair coin, all possible atomic events are assigned equal probability:* $P(\{\langle N_1, N_2\rangle\}) = \frac{1}{4}$ *for* $N_1, N_2 \in \{\mathit{Heads}, \mathit{Tails}\}$*.*

The random variables X_1*,* X_2*, and* V *are defined as follows:*

1. $X_1 : \Omega \to \mathbb{R}$ *such that* $\langle N_1, N_2\rangle \mapsto 0$*, if* $N_1 = \mathit{Heads}$*,* 1 *otherwise;*
2. $X_2 : \Omega \to \mathbb{R}$ *such that* $\langle N_1, N_2\rangle \mapsto 0$*, if* $N_2 = \mathit{Heads}$*,* 1 *otherwise;*
3. $V : \Omega \to \mathbb{R}$ *such that* $V(\omega) = X_1(\omega) + X_2(\omega)$ *for any* $\omega \in \Omega$*;*

and $\mathcal{X}$ *is the borel algebra over the real numbers* $\mathbb{R}$*.*

Nota. For many applications it is not necessary – maybe not even possible – to find an underlying sample space (as it is in the examples of tossing a coin or rolling a dice). Nevertheless, parameters of interest (e. g., stock yield) may formally be interpreted as random variables V in the form of functions as well: Let $\Omega \subseteq \mathbb{R}$ be the set of possible values of such a V with the assignment $\omega = v = V(\omega)$ (for any $\omega \in \Omega$), i. e., V formally becomes the identity function.

Appendix B

Technicalities: Implications of *d*-separation

The sprinkler example in chapter 2 demonstrates how the conditional dependencies represented by the graph can be recovered through the use of the d-separation criterion. Pearl fixes this idea in the following theorem due to Verma and Pearl in [Verma & Pearl 1988].[1]

Theorem B.0.1 (Probabilistic Implications of *d*-Separation)[2]
If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.

We shall have a look at the first part of definition B.0.1 and formalize it (referring back to what Pearl says about *Markov Compatibility* in prose) through the following formula:

$$\forall gp \forall XYZ \Big(G(g) \wedge P(p) \wedge C(g,p) \wedge (X \perp\!\!\!\perp Y \mid Z)_{G[g]} \Rightarrow (X \perp\!\!\!\perp Y \mid Z)_{P[p]} \Big), \quad \text{(B.1)}$$

where G is to be read as *is.a.directed.acyclic.graph*, P is to be read as *is.a.probability.distribution*, and C means *are.compatible* (in accordance with definition 2.6.2). Compatibility requires the existence of a factorization of the joint probability function p under consideration *as dictated* by the corresponding graph g. Moreover, the lower index $_{G[g]}$ indicates independence (i. e., the graphical – hence $_G$ – notion of d-separation) in

[1] See also [Geiger *et al.* 1990].

[2] In [Pearl 2009, p.18]: theorem 1.2.4.

the graph g, whereas the lower index $_{P[p]}$ indicates probability-theoretic – hence $_P$ – independence between the random variables p ranges over.[3] Although PEARL actually uses X, Y, and Z in one formula referring to *nodes in a graph* (by the lower index $_G$) and to *variables in joint distributions* (by the lower index $_P$) at the same time, this use of variables needs to be looked at carefully again, especially when quantifying X, Y, and Z. E. g., X cannot simply refer to nodes, since nodes cannot be independent in a probability-theoretic sense, as suggested in the consequent of formula B.1. On the other hand, X cannot strictly refer to random variables, since there is no explanation as to what it means for a random variable to be d-separated (as suggested by the term with the lower index $_G$). The question remains: What does X refer to if we still want to use it in quantified formulae and attribute some meaning to it? One possible answer might be that X merely refers to a rather abstract *label* that only gets evaluated by the construct $(\cdot \perp\!\!\!\perp \cdot \,|\, \cdot)$ according to the lower index, thus shifting the problem of denotation to the question, how exactly the notion of *compatibility* links the nodes in a graph to the corresponding random variables of a certain joint probability function. Following this suggestion a possible reading of a term such as $(X \perp\!\!\!\perp Y \,|\, Z)_G$ might be: *The nodes I, J, and M are arranged in the graph in such a manner that the nodes which are functionally assigned the labels X and Y are d-separated by the node which is functionally assigned the label Z.* A similar reading applies to terms as $(X \perp\!\!\!\perp Y \,|\, Z)_P$. Now, if the graph in which the d-separation statement is evaluated is *compatible* with the joint probability function where the conditional independence term is evaluated (as demanded in the antecedent of formula B.1), then the functional assignment of *labels* to *nodes in the graph* is interlinked with the functional assignment of the same *labels* to *random variables of the joint probability function.* This expresses the intention of the notion of *Markov Compatibility.*[4]

Since the definition of *Markov Compatibility* (definition 2.6.2) relies on the explication of *Markovian Parents* (definition 2.6.1), which in turn uses a certain ordering of the variables under consideration, we should

[3] This notation extends PEARL's use of the lower index – he merely considers unquantified formulae, such as $(X \perp\!\!\!\perp Y \,|\, Z)_G \Rightarrow (X \perp\!\!\!\perp Y \,|\, Z)_P$.

[4] To formalize these remarks, the connection between d-separation in the graph and conditional independence between random variables must be restated with function terms in the following manner: $(f(X) \perp\!\!\!\perp f(Y) \,|\, f(Z))_G \Rightarrow (h(X) \perp\!\!\!\perp h(Y) \,|\, h(Z))_P$, where f is a function from abstract labels to nodes and h is a function from abstract labels to random variables.

be able to ground the idea of *variable interlinking* on a more basal notion by employing an ordering of the variables, too. In [Verma & Pearl 1988] PEARL introduces the notion of a *causal list* (or *causal input list* when referring to the algorithmic import) for this very purpose. Such a causal list is based on a specific *dependency model* that provides the variables – in our case the joint probability distribution we are examining:[5]

Definition B.0.2 (Causal List)
A causal list of a dependency model contains two things: an ordering of the variables and a function that assigns a tail boundary *to each variable x. For each variable x let U_x denote the set of all variables which come before x in the given ordering. A tail boundary of a variable x, denoted B_x, is any subset of U_x that renders x independent of $U_x - B_x$. A unique DAG can be generated from each causal list by associating the tail boundary of the variable x in the list with the set of direct parents of any node x in the DAG.*[6]

From such a causal input list an edge-minimal graphical representation can be derived algorithmically.[7] The following theorem postulates the existence of a causal input list under given circumstances:

Theorem B.0.3 (Existence of a Causal List)[8]
If M is a dependency model which can be perfectly represented by some DAG D, then there is a causal list L_Θ which generates D.

Example. The right graph in figure 2.5 (page 49) could have been built from the following causal list L with the variable ordering Θ and the tail boundary function B:

$$\begin{aligned}
L &= \langle \Theta, B\rangle, \qquad \text{where} \\
\Theta &= \langle X_2', X_3', X_4', X_5'\rangle, \\
B &= \{\langle X_2', \{\}\rangle, \langle X_3', \{\}\rangle, \langle X_4', \{X_2', X_3'\}\rangle, \langle X_5', \{X_4'\}\rangle\}.
\end{aligned}$$

The given explication is not the only possibility, since, e.g., X_3' could be listed before X_2' in the variable ordering Θ – the only requirement being that, if X_i is an ancestor of X_j in the graph (which is an unambiguous relation in any DAG), then $X_i <_\Theta X_j$.

[5] Cf. [Verma & Pearl 1988, p. 71].

[6] As PEARL adds: *An equivalent specification of a causal list is an ordered list of triplets of the form $I(x, B_x, R)$, one triplet for each variable in the model, where R is $U_x - B_x$.*

[7] Cf. [Verma & Pearl 1988, p. 71].

[8] In [Verma & Pearl 1988, p. 72], with proof.

Having lifted the compatibility requirement $C(g,p)$ of our antecedent in formula B.1 to a more generic level, we can proceed with the universally quantified most inner implication of formula B.1: $(X \perp\!\!\!\perp Y \,|\, Z)_G \Rightarrow (X \perp\!\!\!\perp Y \,|\, Z)_P$. This statement is of course directed from the graphical representation g to the underlying dependency model p, our probability distribution, since we are examining *soundness*. It implies that all independencies read off from the graph g are also present in the probability distribution p, but – in general – not all independencies of p are represented by g. Graphs with this property (relative to a given dependency model) are called *I-maps* (of that model).[9]

Definition B.0.4 (I-map)[10]
It is not always necessary nor feasible to have an exact representation of a dependency model; in fact, an efficient approximation called an I-map *is often preferred to an inefficient perfect map. A representation R is an I-map of a dependency model M iff every independence statement represented by R is also a valid independence of M. Thus, R may not represent every statement of M, but the ones it does represent are correct.*

One remark, before we turn to the '88 version of the proof of soundness: The probability distributions we are dealing with here are so-called *graphoids* due to the list of four common graphoid properties they obey:[11]

$$\begin{array}{llr}
 & & \text{(B.2)} \\
\text{symmetry} & (X \perp\!\!\!\perp Y \,|\, Z) \Longleftrightarrow (Y \perp\!\!\!\perp X \,|\, Z) & \text{(a)} \\
\text{decomposition} & (X \perp\!\!\!\perp YW \,|\, Z) \Longrightarrow (X \perp\!\!\!\perp Y \,|\, Z) & \text{(b)} \\
\text{weak union} & (X \perp\!\!\!\perp YW \,|\, Z) \Longrightarrow (X \perp\!\!\!\perp W \,|\, ZY) & \text{(c)} \\
\text{contraction} & (X \perp\!\!\!\perp W \,|\, ZY) \;\&\; (X \perp\!\!\!\perp Y \,|\, Z) \Longrightarrow (X \perp\!\!\!\perp YW \,|\, Z) & \text{(d)}
\end{array}$$

where X, Y, and Z represent three disjoint subsets of objects (e.g., variables or attributes) and the notation YW is a shorthand for $Y \cup W$.

Having gathered these notional explications we can proceed to PEARL's proof of theorem B.0.1, for which he uses a formalization different from formula B.1, given in the following theorem:

Theorem B.0.5 (Connection between Causal List and I-map)[12]
If M is a graphoid and L_Θ is any causal list of M, then the DAG generated by L_Θ is an I-map of M.

[9] This pertains to undirected and directed graphs – each with corresponding separation criteria.

[10] Cf. [Verma & Pearl 1988, p. 70].

[11] Cf. [Verma & Pearl 1988, pp. 69 ff.].

[12] In [Verma & Pearl 1988, p. 72]: theorem 2.

Employing the explications above we see that the proof of theorem B.0.1 can be reduced to the proof of theorem B.0.5: Our dependency model M is a joint probability distribution p obeying the four graphoid axioms in equation B.2. What we required of the *Markov Compatibility* statement $C(g, p)$ above, we see now encoded in the demand for existence of such a causal list L_Θ generating the DAG g, which now needs to be shown to be an I-map of p.[13]

Proof. To prove the soundness of d-separation we induct on the number of variables in the graphoid M. Let Θ have k variables. In the inductive step we will have to show for some initial segment of Θ with length n that the DAG generated from it is an I-map of M, assuming we have already proven the DAG generated from the initial segment of Θ with length $n - 1$ to be an I-map of M $(n \leq k)$.

Induction Basis. If our graphoid M merely consists of one variable, the DAG generated by L_Θ is an I-map of M trivially.

Induction Hypothesis. Let $L_{\Theta'}$ be based on L_Θ in such a manner that only some initial segment of Θ with length n (called Θ' in the following) is considered in the declaration of boundaries by B'. Since we are following the ordering Θ', we will be concerned with the last variable in this ordering, v. Let $L_{\Theta'-v}$ be the causal list $L_{\Theta'}$ formed by removing v from Θ' and all entries containing v from B' (this will only be a single entry, since v cannot appear in any boundaries, yet, by method of construction).[14] Moreover, let the DAG generated from $L_{\Theta'-v}$ be $G' - v$. Graphically, expanding $G' - v$ to G' will mean the addition of the v node and its *incident* edges (v cannot be parent to any other node in the DAG G' at this step in the construction process, guaranteed by the ordering of variables being consistent with the parentship relation in the DAG). Last, let $M_{G'}$ be the dependency model (a graphoid) corresponding to G' and $L_{\Theta'}$ (with n variables).

We suppose that the DAG $G' - v$ is an I-map of M.[15]

[13] The proof given here follows [Verma & Pearl 1988, pp. 72 ff.] and [Geiger *et al.* 1990, pp. 517 f.].

[14] In the following, these derived concepts will also be referred to as $\Theta' - v$ and $B' - v$.

[15] This in turn entails that the DAG $G' - v$ is also an I-map of $M_{G'-v}$, since the same $n - 1$ variables appear in both, $M_{G'-v}$ contains a thorough list of all independencies between these variables, and no independence information will be *overwritten* on the way to building up M by adding further variables to $M_{G'-v}$. M contains all d-separated triplets of $G' - v$, and so does $M_{G'-v}$ as the minimal case.

Inductive Step. By the induction hypothesis we can assume that $M_{G'-v} \subseteq M$, i. e., there are no independencies in $M_{G'-v}$ which are not contained in M as well.

The following schema symbolizes the step from $n-1$ to n and lists the target objects for each of the aforementioned concepts:

$$G' - v \xrightarrow{\text{adding the node } v \text{ together with all its incident links}} G'$$

$$M_{G'-v} \xrightarrow{\text{adding the variable } v \text{ plus all independence triplets containing } v} M_{G'}$$

$$L_{\Theta'-v} \xrightarrow{\text{extending } \Theta' - v \text{ by } v \text{ and } B' - v \text{ by the boundary assignment for } v} L_{\Theta'}$$

Now, each $M_{G'}$ triplet T of the form $(A \perp\!\!\!\perp B \,|\, C)$ falls into one of three categories: Either the newly added variable v does not appear in T, at all, or it appears in the first position A (in the second position B, respectively – by symmetry, as in equation B.2.a) or in the third position (i. e., C) of T. These three cases will have to be treated separately for *all* such triplets T of $M_{G'}$. We have to make sure that *no* triplet, introduced by the addition of v to $L_{\Theta'-v}$ and evaluated in the graph, is *off* M, to finally conclude that $M_{G'} \subseteq M$:

Case 1. If v does not appear in T, T must be of the form $(X \perp\!\!\!\perp Y \,|\, Z)$, where X, Y, and Z are three disjoint subsets of variables (none containing v). If T is in $M_{G'}$ it must already have been in $M_{G'-v}$, since otherwise there would have been at least one active path in $G' - v$ (between X and Y when Z is instantiated) which would have been deactivated by adding v. But the mere addition of nodes and further links cannot deactivate formerly active paths in a DAG. We know that $G' - v$ is an I-map of $M_{G'-v}$, so T must be an element of $M_{G'-v}$, which in turn is a subset of M, hence T is in M.

Case 2. The sub-case of v appearing in the first position of the triplet can be treated equally to the sub-case of v appearing in the second position by symmetry. The following argument goes for the first position, i. e., we are considering a triplet T of the form $(Xv \perp\!\!\!\perp Y \,|\, Z)$. Again, X, Y, and Z are three disjoint subsets of variables. Let $\langle v, B, R \rangle$ be the last triplet in $L_{\Theta'}$.[16]

[16] Here, B denotes the *Tail <u>B</u>oundary*, R the *<u>R</u>est*, i. e., the set of preceding variables separated from v by the set B.

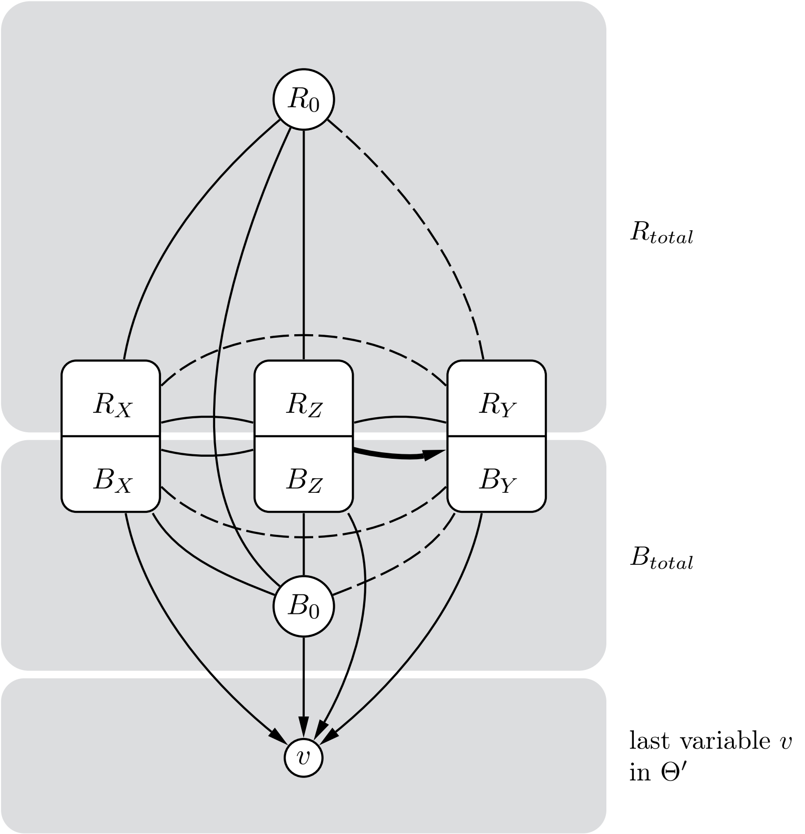

Fig. B.1: Graphical, schematic overlay of the two independence statements $(v \perp\!\!\!\perp R_{total} \,|\, B_{total})$ and $(X \perp\!\!\!\perp Y \,|\, Z)$.

Figure B.1 displays the schematic overlay of T and the last triplet in $L_{\Theta'}$: The background (mirroring the last entry in $L_{\Theta'}$) is divided into three areas – $\{v\}$, R_{total}, and B_{total} – in which v, X, Y, and Z have to be accommodated. X, Y, and Z itself are partitioned into $R_X \cup B_X$, $R_Y \cup B_Y$, and $R_Z \cup B_Z$, thereby marking possible overlaps with B and R (set-theoretically interpreted as the possible existence of shared elements). B_0 and R_0 collect all nodes not named explicitly, so that $B_{total} = B_X \cup B_Y \cup B_Z \cup B_0$, $R_{total} = R_X \cup R_Y \cup R_Z \cup R_0$ (both of which are partitions), and $U = \{v\} \cup B_{total} \cup R_{total}$. Undirected edges indicate the (possible) existence of *active paths* between two sets of nodes.[17]

[17] Note that sets of nodes dividing two active paths are not deactivating the compound path, in general, the only difference being v itself, which acts as a collider node for any pair of incident links due to the method of construction and hence deactivates any traversing paths.

Additionally, edges with arrowheads between two sets of nodes exclude directed edges pointing in the inverse direction.

By definition of the causal list $L_{\Theta'}$, we have given that there are no active paths between R_{total} and v. On the other hand, *all* nodes in B_{total} are necessarily connected to v by method of construction. Moreover, all nodes within B_{total} might be interconnected, as well as all nodes in R_{total}. Also, the last entry in $L_{\Theta'}$ does not exclude any active paths between a node in B_{total} and a second node in R_{total}, so these have to be included, too.

By application of *decomposition* (equation B.2.b) to $(Xv \perp\!\!\!\perp Y \mid Z)$ we get

$$(X \perp\!\!\!\perp Y \mid Z) \in M, \tag{B.3}$$

because v does not occur in this independence statement. Following this statement, we can take four edges (marked by dashed lines) out of the schema in figure B.1:

1. The path $R_Y \rightarrow R_0$ has to be deleted – otherwise it would possibly open an active path $Y \rightarrow R_0 — X$ thereby contradicting equation B.3.

2. The path $R_Y \rightarrow R_X$ has to be deleted, too, because all paths from Y to X have to be intercepted by Z.

3. The path $B_Y \rightarrow B_X$ must be omitted by analogy.

4. Finally, the path $B_Y \rightarrow B_0$ would possibly open the path along $B_Y \rightarrow B_0 — B_X$ and thus circumvent Z on the way from Y to X. It has to be cancelled to avoid contradicting equation B.3, again.

The only path that remains untouched by the above considerations is $B_Y \rightarrow v$, because v acts as a collider due to the method of construction (i. e., it cannot be parent to any other node, yet) and deactivates any traversing paths. Hence this path must stay in the graph as a possible link.

Since we find in the graph G' the independence statement $T = (Xv \perp\!\!\!\perp Y \mid Z)$, we also find $(v \perp\!\!\!\perp Y \mid Z)$ (by *decomposition* as in equation B.2.b). Moreover, we know that B_Y has to be connected to v by an arrow pointing towards v. Now, the only way Z alone would d-separate B_Y from v would be by *functionally determining* B_Y (indicated by the bold arrow in the diagram), since we have the chain

$Z — B_Y \twoheadrightarrow v$ with Z not lying on the path from B_Y to v. *Functional determination* means that the value of any node in B_Y is fixed once we know the value of (all nodes in) Z, i. e., B_Y has no other parent nodes than Z, and any interconnections with other sets of nodes have to be directed edges *emanating* from B_Y: Z is *screening off* B_Y from any influences of non-descendants.[18] This can easily be seen, because if we have $(v \perp\!\!\!\perp B_Y \,|\, Z)$ and functional determination of B_Y by Z, we can infer $(v \perp\!\!\!\perp B_Y \,|\, B_Y)$ by collapsing the path $Z \twoheadrightarrow B_Y$. The last statement can be paraphrased: *Once we know the value of B_Y, learning the value of B_Y (which is not new to us at that point) does not change our degree of belief in a certain value of v.* In fact, this holds for any node or set of nodes replacing v in that context. This in turn means that $(B_Y \perp\!\!\!\perp U \backslash (B_Y \cup Z) \,|\, Z)$ must already have been an element of $M_{G'-v} \subseteq M$.[19]

We thus get in particular

$$(B_Y \perp\!\!\!\perp Xv \,|\, Z) \in M. \tag{B.4}$$

Since all constraints imposed onto the schema so far hold in M, we can now read off directly from the resulting graph that

$$(R_Y \perp\!\!\!\perp Xv \,|\, ZB_Y) \in M. \tag{B.5}$$

Referring back to the declaration of the partitions in the schema and applying *contraction* (equation B.2.d) to equations B.4 and B.5, we conclude $(B_Y R_Y \perp\!\!\!\perp Xv \,|\, Z) \in M$, which yields $(Xv \perp\!\!\!\perp Y \,|\, Z) \in M$.

Case 3. If v appears in the third entry of the triplet, then T must be of the form $(X \perp\!\!\!\perp Y \,|\, Zv)$. As we saw in case 1 above, the addition of v and its incident links alone cannot serve to deactivate a formerly active path in G'. So $(X \perp\!\!\!\perp Y \,|\, Z)$ already holds in $G' - v$ and therefore also in G'. Together with the *weak transitivity property* of DAGs these two independence statements result in $(Xv \perp\!\!\!\perp Y \,|\, Z)$, or $(X \perp\!\!\!\perp Yv \,|\, Z)$ by symmetry. In case 2 we saw that triplets of this form must also be in M. Finally, applying the *weak union conversion* (equation B.2.c) to $(Xv \perp\!\!\!\perp Y \,|\, Z)$ yields $T \in M$. ⊠

[18] The notion of *functional determination* is explicated in [Geiger *et al.* 1990, pp. 517 f.] where Pearl's proof of the soundness of d-separation is given reformulated. Case 2 of the proof, as stated here, follows this alternative route, too.

[19] We are considering the general case of B_Y not being empty. Functional determination nevertheless also holds for the case $B_Y = \varnothing$.

Due to the asymmetry in the I-mapness relation, not all independencies contained in a dependency model M can necessarily be read off from a graph G, even if M and G are compatible. This is the content of the next theorem concluding the discussion about the probabilistic implications of d-separation and referring back to where we started with *Markov Compatibility*:

Theorem B.0.6 (Implications of Markov Compatibility)[20]
For any three disjoint subsets of nodes (X, Y, Z) in a DAG G and for all probability functions P, we have:

(i) $(X \perp\!\!\!\perp Y \mid Z)_G \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)_P$ *whenever G and P are compatible; and*

(ii) *if $(X \perp\!\!\!\perp Y \mid Z)_P$ holds in all distributions compatible with G, it follows that $(X \perp\!\!\!\perp Y \mid Z)_G$.*

[20] In [Pearl 2009, p.18]: theorem 1.2.5.

References

Albert, Max. 2007. The propensity theory: a decision-theoretic restatement. *Synthese*, **156**(3), 587–603.

Allwein, Gerard and Jon Barwise (eds). 1996. *Logical Reasoning with Diagrams.* Oxford University Press.

Baumgartner, Michael. (forthcoming). Interventionism and Epiphenomenalism. *Canadian Journal of Philosophy.*

Bayes, Thomas. 1763. Facsimiles of two papers by Bayes I. An essay toward solving a problem in the doctrine of chances, with Richard Price's forward and discussion. *Phil. Trans. Royal Soc. London*, **53**, 370–418.

Beebee, Helen. 2009. *Causation and Observation. In:* [Beebee *et al.* 2009]. Chap. 22, pages 471–497.

Beebee, Helen, Christopher Hitchcock, and Peter Menzies (eds). 2009. *The Oxford Handbook of Causation (Oxford Handbooks).* Oxford University Press.

Boolos, George S., John P. Burgess, and Richard C. Jeffrey. 2002. *Computability and Logic.* 4th edn. Cambridge University Press.

Burgess, Simon. 2004. The Newcomb Problem: An Unqualified Resolution. *Synthese*, **138**(2), 261–287.

Cartwright, Nancy. 2001. What is Wrong with Bayes Nets? *Monist*, **84**(2), 242.

—— 2004. Causation: One Word, Many Things. *Philosophy of Science*, **71**(5), 805–819.

—— 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics.* 1st edn. Cambridge University Press.

Chang, Hasok and Nancy Cartwright. 1993. Causality and Realism in the EPR Experiment. *Erkenntnis*, **38**(2), 169–190.

Choi, Sungho. 2003. The Conserved Quantity Theory of Causation and Closed Systems. *Philosophy of Science*, **70**(3), 510–530.

Collins, John, Ned Hall, and L. A. Paul (eds). 2004. *Causation and Counterfactuals (Representation and Mind).* MIT Press.

Craver, Carl F. and William Bechtel. 2007. Top-Down Causation Without Top-Down Causes. *Biology and Philosophy*, **22**(4), 547–563.

Davidson, Donald. 1963. Actions, Reasons, and Causes. *Journal of Philosophy*,

60(23), 685–700.

de Pierris, Graciela and Michael Friedman. 2008. Kant and Hume on Causality. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, CSLI, Stanford University.

Dechter, Rina, Hector Geffner, and Joseph Halpern (eds). 2010. *Heuristics, Probability and Causality. A Tribute to Judea Pearl*. College Publications.

Dowe, Phil. 2009. *Causal Process Theories. In:* [Beebee *et al.* 2009]. Chap. 10, pages 213–233.

Ertel, Wolfgang. 2009. *Grundkurs Künstliche Intelligenz*. 2nd, revised edn. Vieweg+Teubner.

Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot, and Gerhard Tutz. 2000. *Statistik. Der Weg zur Datenanalyse*. 3rd edn. Springer-Lehrbuch. Springer-Verlag, Berlin – Heidelberg – New York.

Garrett, Don. 2009. *Hume. In:* [Beebee *et al.* 2009]. Chap. 4, pages 73–91.

Gauthier, David. 1988. In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality). *Proceedings of the Aristotelian Society*, **89**, 179–194.

Geiger, Dan, Thomas Verma, and Judea Pearl. 1990. Identifying Independence in Bayesian Networks. *Networks (New York, NY)*, **20**(5), 507–534.

Gillies, Donald. 2000. Varieties of Propensity. *The British Journal for the Philosophy of Science*, **51**(4), 807–835.

—— 2001. Critical Notice on *Causality: Models, Reasoning, and Inference* by Judea Pearl. *The British Journal for the Philosophy of Science*, **52**(3), 613–622.

—— 2002. Causality, Propensity, and Bayesian Networks. *Synthese*, **132**(1-2), 63–88.

—— 2005. An Action-Related Theory of Causality. *The British Journal for the Philosophy of Science*, **56**(4), 823–842.

Glennan, Stuart. 2009. *Mechanisms. In:* [Beebee *et al.* 2009]. Chap. 15, pages 315–325.

Gärdenfors, Peter. 2008 (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Reprint edn. Studies in Logic: Mathematical Logic and Foundations, no. 13. College Publications.

Halpern, Joseph Y. and Judea Pearl. 2005a. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, **56**(4), 843–887.

—— 2005b. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, **56**(4), 889–911.

Healey, Richard. 2009. *Causation in Quantum Mechanics. In:* [Beebee *et al.* 2009]. Chap. 33, pages 673–686.

Hitchcock, Christopher. 1995. The Mishap at Reichenbach Fall: Singular vs. General Causation. *Philosophical Studies*, **78**(3), 257–291.

—— 1996. Causal Decision Theory and Decision-theoretic Causation. *Noûs*, **30**(4), 508–526.

—— 2004. Causal Processes and Interactions: What Are They and What Are They Good For? *Philosophy of Science*, **71**(5), 932–941.

—— 2007. Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*, **116**(4), 495–532.

—— 2009a. *Causal Modelling. In:* [Beebee *et al.* 2009]. Chap. 14, pages 299–314.

—— 2009b. Structural Equations and Causation: Six Counterexamples. *Philosophical Studies*, **144**(3), 391–401.

—— 2010. Probabilistic Causation. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

—— (forthcoming). Events and Times: A Case Study in Means-Ends Metaphysics. *Philosophical Studies.*

Horwich, Paul. 1985. Decision Theory in Light of Newcomb's Problem. *Philosophy of Science*, **52**(3), 431–450.

Huber, Franz. 2009. Ranking Functions. *In: Encyclopedia of Artificial Intelligence.* Hershey.

Hume, David. 1748. *An Enquiry Concerning Human Understanding.* The University of Adelaide Library 2004 (derived from the Harvard Classics Volume 37, 1910 P.F. Collier & Son.).

Humphreys, Paul. 1985. Why Propensities Cannot be Probabilities. *Philosophical Review*, **94**(4), 557–570.

Hájek, Alan. 2010. Interpretations of Probability. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

Jaynes, E. T. 1989. Clearing up Mysteries – The Original Goal. *Pages 1–27 of:* Skilling, J. (ed), *Maximum Entropy and Bayesian Methods.* Kluwer Academic Publishers, Dordrecht.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge University Press.

—— 2002. Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **110**(1), 69–102.

—— 2007. Are Newcomb Problems Really Decisions? *Synthese*, **156**(3), 537–562.

Kistler, Max. 2010. Mechanisms and Downward Causation. *Philosophical Psychology*, **22**(5), 595–609.

Lakoff, George. 1990. *Women, Fire, and Dangerous Things – What Catego-*

ries Reveal about the Mind. Reprint edn. University Of Chicago Press.

Lauth, Bernhard and Jamel Sareiter. 2002. *Wissenschaftliche Erkenntnis.* mentis Verlag GmbH, Paderborn.

Lewis, David. 1973a. Causation. *Journal of Philosophy*, **70**(17), 556–567.

—— 1973b. *Counterfactuals.* 2nd edn. Wiley-Blackwell.

—— 1979. Prisoners' Dilemma is a Newcomb Problem. *Philosophy & Public Affairs*, **8**(3), 235–240.

—— 1980. A Subjectivist's Guide to Objective Chance. *Chap. 13, pages 263–293 of:* Jeffrey, Richard C. (ed), *Studies in Inductive Logic and Probability*, vol. 2. Berkeley: University of Berkeley Press.

—— 1986a. *Philosophical Papers: Volume II.* Oxford University Press.

—— 1986b. *Postscripts to "Causation". In:* [Lewis 1986a]. Pages 172–213.

—— 2000. Causation as Influence. *Journal of Philosophy*, **97**(4), 182–197.

Link, Godehard. 2009. *Collegium Logicum – Logische Grundlagen der Philosophie und der Wissenschaften – Band I.* mentis Verlag GmbH, Paderborn.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. Thinking About Mechanisms. *Philosophy of Science*, **67**(1), 1–25.

Mackie, John Leslie. 1965. Causes and Conditions. *American Philosophical Quarterly*, **2**(4), 245–264.

—— 1980. *The Cement of the Universe: A study of Causation.* Clarendon Paperbacks. Oxford University Press.

Mccurdy, Christopher S. I. 1996. Humphrey's Paradox and the Interpretation of Inverse Conditional Propensities. *Synthese*, **108**(1), 105–125.

Menzies, Peter. 2004. Causal Models, Token Causation, and Processes. *Philosophy of Science*, **71**(5), 820–832.

—— 2009a. Counterfactual Theories of Causation. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

—— 2009b. *Platitudes and Counterexamples. In:* [Beebee *et al.* 2009]. Chap. 17, pages 341–367.

Mumford, Stephen. 2009. *Causal Powers and Capacities. In:* [Beebee *et al.* 2009]. Chap. 12, pages 265–278.

Nozick, Robert. 1969. Newcomb's Problem and Two principles of Choice. *Pages 114–146 of:* Rescher, Nicholas (ed), *Essays in Honor of Carl G. Hempel.* Dordrecht: Reidel.

Paul, L. A. 2009. *Counterfactual Theories. In:* [Beebee *et al.* 2009]. Chap. 8, pages 158–184.

Pearl, Judea. 1982. Reverend Bayes on Inference Engines: a Distributed Hierarchical Approach. *Pages 133–136 of: Proceedings of the Second National Conference on Artificial Intelligence. AAAI-82.* Pittsburgh, PA: AAAI Press.

—— 1995. Causal diagrams for empirical research. *Biometrika*, **82**(4), 669–688.

—— 2000a. *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

—— 2000b. *The Logic of Counterfactuals in Causal Inference (Discussion of 'Causal Inference without Counterfactuals' by A. P. Dawid).* Tech. rept. R-269. Cognitive Systems Laboratory – Departments of Computer Science and Statistics – University of California, Los Angeles, CA 90024.

—— 2009. *Causality: Models, Reasoning, and Inference.* 2^{nd} edn. Cambridge University Press.

—— 2010. Review of N. Cartwright *Hunting Causes and Using Them* (R-342, September 2008). *Economics and Philosophy*, **26**, 69–77.

POPPER, Karl R. 1959. The Propensity Interpretation of Probability. *The British Journal for the Philosophy of Science*, **10**(37), 25–42.

REICHENBACH, Hans. 1956. *The Direction of Time.* University of Los Angeles Press.

RUSSELL, Bertrand. 1913. On the Notion of Cause. *Pages 1–26 of: Proceedings of the Aristotelian Society.* New Series, vol. 13. The Aristotelian Society.

SCHAFFER, Jonathan. 2008. The Metaphysics of Causation. *In:* ZALTA, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

SCHULZE, Wolfgang. (forthcoming). Cognitive Transitivity. The Motivation of Basic Clause Structures.

SCHUMACKER, Randall E. and Richard G. LOMAX. 2004. *A Beginner's Guide to Structural Equation Modeling.* 2^{nd} edn. Psychology Press.

SHAPIRO, Lawrence A. and Elliott SOBER. 2007. Epiphenomenalism - the Do's and the Don'ts. *In:* WOLTERS, Gereon and Peter K. MACHAMER (eds), *Studies in Causality: Historical and Contemporary.* University of Pittsburgh Press.

SHRIER, Ian and Robert PLATT. 2008. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, **8**(1), 70.

SLOMAN, Steven A. 2005. *Causal models: how people think about the world and its alternatives.* Oxford Scholarship Online. Oxford University Press.

SONG, Jae Jung. 1996. *Causatives and Causation: A Universal-Typological Perspective.* Longman Linguistics Library. Addison Wesley Publishing Company.

SOSA, Ernest (ed). 1974. *Causation and Conditionals (Readings in Philosophy).* Oxford University Press.

SPOHN, Wolfgang. 1983. *Eine Theorie der Kausalität.* Habilitationsschrift (LMU München).

—— 2000. Bayesian Nets Are All There Is To Causal Dependence. *Pages 157–172 of:* GALAVOTTI, M. C. *et al.* (eds), *Stochastic Dependence and*

Causality. CSLI Publications, Stanford.

——— 2001. *Deterministic Causation. In:* [Spohn *et al.* 2001]. Pages 21–46.

——— 2006. Causation: An Alternative. *The British Journal for the Philosophy of Science*, **57**(1), 93–119. Reprinted in: [Spohn 2008].

——— 2008. *Causation, Coherence and Concepts: A Collection of Essays.* Boston Studies in the Philosophy of Science. Springer-Verlag, Berlin – Heidelberg – New York.

——— 2009. A Survey of Ranking Theory. *In:* Huber, Franz and Christoph Schmidt-Petri (eds), *Degrees of Belief.* Springer-Verlag, Berlin – Heidelberg – New York.

——— 2010. *The Structural Model and the Ranking Theoretic Approach to Causation: A Comparison. In:* [Dechter *et al.* 2010]. Chap. 29.

——— (forthcoming). Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box. *Synthese.*

Spohn, Wolfgang, Marion Ledwig, and Michael Esfeld (eds). 2001. *Current Issues in Causation.* Paderborn: mentis Verlag GmbH, Paderborn.

Stalnaker, Robert C. 1968. *A Theory of Conditionals. In:* [Sosa 1974]. Chap. XII, pages 165–179.

Suppes, Patrick. 1987. Propensity Representations of Probability. *Erkenntnis*, **26**(3), 335–358.

van de Laar, Tjeerd. 2006. Dynamical Systems Theory as an Approach to Mental Causation. *Journal for General Philosophy of Science*, **37**(2), 307–332.

Verma, Thomas and Judea Pearl. 1988. Causal Networks: Semantics and Expressiveness. *In: Proceedings of the 4th Annual Conference on Uncertainty in Artificial Intelligence (UAI-88).* New York: Elsevier Science.

Watkins, Eric. 2009. *Kant. In:* [Beebee *et al.* 2009]. Chap. 5, pages 92–107.

Weatherson, Brian. 2009. David Lewis. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

Weirich, Paul. 2008. Causal Decision Theory. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, CSLI, Stanford University.

Williamson, Jon. 2009. *Probabilistic Theories. In:* [Beebee *et al.* 2009]. Chap. 9, pages 185–212.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation (Oxford Studies in the Philosophy of Science).* Oxford University Press.

——— 2009. *Agency and Interventionist Theories. In:* [Beebee *et al.* 2009]. Chap. 11, pages 234–264.

Register of names

Roland Poellinger • `http://logic.rforge.com` • 2011/12