Causality and Unification: How Causality Unifies Statistical Regularities

Gerhard Schurz

Received: 22/06/14
Final Version: 25/11/14

BIBLID 0495-4548(2015)30:1p.73-95
DOI: 10.1387/theoria.11913

ABSTRACT: Two key ideas of scientific explanation –explanation as causal information and explanation as unification– have frequently been set into mutual opposition. This paper proposes a "dialectical solution" to this conflict, by arguing that causal explanations are preferable to non-causal ones, because they lead to a higher degree of unification at the level of explaining statistical regularities. The core axioms of the theory of causal nets (TC) are justified because they offer the best if not the only unifying explanation of two statistical phenomena: screening off and linking up. Alternative explanations of the two phenomena are discussed and it is shown why they don't work. It is demonstrated that although the core axioms of TC are empirically vacuous, extended versions of TC have empirical content by means of which they can generate independently testable predictions.

Keywords: Unification, explanation, causality, theory of causal nets, screening off, linking up

RESUMEN: Con frecuencia se han planteado como contrapuestas dos ideas clave en la explicación científica (explicación como información causal y explicación como unificación). El presente artículo propone una "solución dialéctica" argumentando que las explicaciones causales son preferibles a las no-causales porque aquellas comportan un mayor grado de unificación en la explicación de regularidades estadísticas. Los axiomas centrales de la teoría de redes causales (TC) están justificados porque ofrecen la mejor, si no la única, explicación unificada de dos fenómenos estadísticos: neutralización (*screening off*) y vinculación (*linking up*). Se discuten las explicaciones alternativas de estos dos fenómenos y se razona por qué no funcionan. Se demuestra además que aunque los axiomas centrales de TC son empíricamente vacuos, las versiones extendidas de TC tienen un contenido empírico gracias al cual pueden generar predicciones independientemente contrastables.

Palabras clave: unificación, explicación, causalidad, teoría de redes causales, neutralización, vinculación

1. Introduction: Causality or unification? A dialectical solution

Since the earliest times of mankind, human beings of all cultures have explained regular connections of events in term of causal connections. Evolutionary psychology teaches us that the construction of causal models is a characteristic ability of homo sapiens (Tomasello 1999), and cognitive psychologists have found that causal modelling appears spontaneously in very early stages of childhood (Sperber et al. 1995). In accordance with these findings, it is a deep-seated common sense intuition that all regular connections between events that we observe have their explanation in terms of cause-effect relations. Elaborating this intuition, a variety of philosophers of science have argued that all good scientific explanations are causal (e.g. Railton 1978, Salmon 1984, and Strevens 2008), or more generally, that the metaphysics of causality is the basis of our understanding of the world (cf. Beebee et al. 2009, parts II and III).

These intuitions stand in stark contrast to the perennial difficulties philosophers have had when trying to *justify* causality as something ontologically real. According to Hume's fundamental sceptical challenge, the classification of correlated events into causes and effects doesn't correspond to anything real "out there in the world" – only the correlations are real, while causality is a mere habit of our cognitive system. Since then, many philosophers have renewed and elaborated Hume's challenge (famously Russell 1912/13 and more recently Norton 2009 and Psillos 2009). Following these criticisms of causality, there is an opposing camp of philosophers of science who argue that the characteristic mark of good scientific theories and explanations is not their causal nature, but their power of *unification*, i.e., their ability to predict and explain a variety of empirical phenomena in terms of a small number of basic principles. The idea that *unification* is the main goal of scientific theories has been articulated by Mach (1883, 586f) and Whewell (1837). More recently, it has been proposed that unification is the major feature of scientific explanations, by Friedman (1974), Kitcher (1981), Schurz/Lambert(1994), Schurz (1999) and de Regt (2006). These authors see the main goal of explanation as to provide a deeper *understanding* of the empirical facts by means of unifying them.

These two paradigms of scientific explanation —explanation as causal information versus explanation as unification— have frequently been set into mutual opposition (cf. Kitcher 1989, Salmon 1989, 180-186, de Regt 2006, see Schurz 2014). In this paper I will propose a "dialectical" solution of this conflict: I will argue that one reason why causal explanations are preferable to non-causal explanations is that they lead to a higher degree of unification. I don't say that this is the only reason: a second reason for the preferability of causal over non-causal explanation is that only the former but not the latter can inform us about practical possibilities of producing intended effects by the right kinds of actions or interventions. However, causality cannot be reduced to intervention possibilities, since not all parts of nature can be changed by human interventions. What this paper intends to show is that independently from this practical advantage the great theoretical advantage of the assumption of causal principles lies in their unification power in the explanation of empirical regularities. More precisely, we want to show that there exists a general *theory of causality*, abbreviated as *TC*, which offers the best and most unifying explanation (among all available explanations) of two (in)stability properties of statistical dependencies with regard to conditionalization: *screening off* and *linking up*. In other words, we think that the general principles of TC can be justified by means of an *inference to the best explanation* (IBE). Since the principles of TC provide unified higher-level explanations of statistical regularities and their (in)stability properties, explanations of single events whose general premises offer causal information contribute more to this unification than single event explanations whose general premises do not offer causal information (we return to this argument in sec. 4).

The theory of causality (TC) that we have in mind is the *theory of causal nets* that has been developed by SGS (= Spirtes, Glymour, and Scheines 2000) and Pearl (1988, 2009), with forerunners such as Reichenbach (1956), Blalock (1961), and Suppes (1970). This paper intends to contribute to this theory by showing that directed cause-effect relations as axiomatized by TC are the *best* —in the sense of most unifying— explanation of screening off and linking up. We consider alternative explanations and highlight their problems and disadvantages. To this end, we understand probabilities in the *statistical* sense, as dispositions of repeatable events to produce inductively inferred limiting frequencies; this objective interpretation of probability is important for our attempt to justify the attribution

of causal connections, by their power to explain probabilistic dependencies as objective features of the world (as opposed to subjective features of beliefs). It is an intended consequence of this view that causal claims involving singular events have to be backed up by probabilistic regularities.

Our account makes use of mathematical variables. A variable is a function X: D → Val(X) from a domain D of individuals to its value space Val(X) = {x_1, x_2, ...}, which is a family of properties or a set of numbers. If X denotes colour, for example, then Val(X) = {red, green,...} and X assigns a colour X(d) to every individual d ∈ D. For example, that the object d has the colour green is expressed by "X(d) = x_2", where "x_2" denotes "green" (etc.). We also admit that D consists of n-tuples of individuals, e.g. individuals at certain time-points. Simple dichotomic property-pairs are represented by binary variables X_F with value space {F, ¬ F} (e.g., {red,not-red}). We make use of the following notational conventions:

X,Y,... are variables and U,V,... sequences of variables.

Lower-case letters "x" (or "x_i") stand for values of X, and lower-case "u" (or "u_i") for sequences of values of the respective variables in U.

$P(X_1, ..., X_n)$ is a (statistical) probability distribution over a suitable algebra AL over the space of values, i.e. P: $AL(Val(X_1) \times ... \times Val(X_n)) \rightarrow [0,1]$).

"P(x)" abbreviates "P({x})" and represents "P(X(α)=x)", i.e. the probability that X takes value x in the underlying domain D (the individual variable α is bound by P).

Likewise, "P(S)" abbreviates "P(X(α) ∈ S)", "P(¬ x)" abbreviates "P(X(α) ∉ x)", "P(x,y)" abbreviates "P(X(α) = x ∧ Y(α) = y)", and "P(x|y)" abbreviates "P(X(α) = x|Y(α) = y)", i.e. the conditional probability of x given y, provided P(y) > 0.

Two variables X,Y are said to be probabilistically dependent (DEP(X,Y)) iff at least *some* of their values are dependent; they are probabilistically independent (INDEP(X,Y)) iff *all* of their values are independent. More generally, probabilistic (in)dependence between X and Y, conditional on a sequence of variables U, is defined by any one of the following equivalent formulations (a) – (c):

(1) DEP(X,Y|U) *iff* (a) ∃x,y,u: P(x|y,u) ≠ P(x|u) and P(y,u) > 0, or
(b) ∃x,y,u: P(y|x,u) ≠ P(y|u) and P(x,u) > 0, or
(c) ∃x,y,u: P(x,y|u) ≠ P(x|u) × P(y|u).

INDEP(X,Y|U) *iff* not DEP(X,Y|U), i.e. *iff* ∀x,y,u: P(x|y,u) = P(x|u) or P(y,u) = 0 (with equivalent formulations as above).

The equivalence of (a) with (b) makes clear that probabilistic dependencies are always *symmetric*. *Unconditional* dependence is defined as dependence conditional on the empty sequence: DEP(X,Y) iff DEP(X,Y|∅), and likewise for unconditional independence. Moreover, the definition of probabilistic dependence is generalized to sequences of variables U,V,W via the following definition: DEP(U,V|W) iff ∃u,v,w: P(u|v,w) ≠ P(u|w) and P(v,w) > 0.

Probabilistic (in)dependence between *values* of variables is defined as follows: DEP(x,y|u) iff P(x|y,u) ≠ P(x|u) and P(y,u) > 0 and INDEP(x,y|u) iff P(x|y,u) = P(x|u) or P(y,u) = 0. If P(x|y,u) > P(x|u), then x depends positively on y, and if P(x|y,u) > P(x|u), then x depends negatively on y.

2. Causality obtained from an inference to the best explanation

2.1. Causality as a theoretical concept: a comparison with Newtonian force

According to the findings of contemporary (post-positivistic) philosophy of science, scientific theories contain theoretical concepts such as atom, force, etc.[1] Theoretical concepts are not definable in terms of observable phenomena; instead, they offer unified explanations of them in terms of hidden structures. Nor is a primitive theoretical concept definable by a single theoretical principle or axiom. Rather, its semantic content is characterized by a theory, or at least by the *core* of a theory, to which this concept belongs. Classical physics, for example, stipulates gravitational forces as unobservable causes of the trajectories of physical bodies. The "meaning" of "gravitational force" is not determined by a single definition, but by the joint effect of the synthetic axioms of Newtonian mechanics which, when combined, explain a large variety of empirical phenomena. We suggest that causality should, in precise analogy to force, be understood as a theoretical concept whose meaning can be explicated by the core axioms of the theory of causality, TC. We thus assume that the empirical (or non-theoretical) concept of TC is the notion of a probability distribution over a set of variables, whose properties are to be explained by assuming theoretical cause-effect relations between these variables according to the principles of TC.

In order to be empirically significant, it is not sufficient for theoretical concepts and the axioms characterizing them to offer "some" explanations of the empirical phenomena. These explanations must not be entirely post-facto, but should be able to generate use-novel empirical content, i.e. potentially novel predictions by which they are independently testable.

Of course, there is no guarantee that a theoretical concept that offers unifying explanations does also refer to a really existing entity. Purely *instrumental* interpretations of theoretical concepts as useful means of unifying empirical phenomena are always possible. But the more explanatory and predictively successful a theory becomes, the more plausible it is to assume that the theoretical concepts that produce this success actually *do* refer to something real. The concept of force in Newtonian physics is explanatory successful to an admirably large extent. In this paper we attempt to show that the concept of causality can also be justified by appeal to its success in offering unifying explanations.

The question of the empirical content of TC will be briefly tackled in section 3. The focus of this section will be a demonstration of how TC can be justified by appeal to its explanatory power. The decisive question which we must answer for this purpose is: *What does causality explain?*

The answer cannot be that every empirical regularity is explained by a corresponding causal power. For every regularity, one can postulate a corresponding causal connection that "explains" it *post facto*. Causal "explanations" of this sort would amount to a mere metaphysical duplication of empirical regularities that can neither achieve explanatory unification nor generate new empirical content. Therefore Ockham's razor dictates their elimination.

Causality is also not needed to explain why observed regularities are inductively projectible, as some philosophers have suggested (cf. Fales 1990, ch. 4). The inductive projectibility of regularities is already explained by assuming that they are backed up by *lawlike*

[1] Cf. Carnap (1956), Lewis (1970), Sneed (1971), Balzer et al. (1987), Papineau (1996), French (2008).

connections (Armstrong 1983, part 1). Causality, however, goes *beyond* inductive projectibility or lawlikeness: regularities connecting the joint effects of a common cause, for instance, may be perfectly lawlike although they are obviously non-causal.

To withstand Hume's sceptical challenge one has to answer the question of why cause-effect relations are *needed at all*, instead of simply accepting lawlike regularities as primitive facts. Our answer is that cause-effect relations yield the best available explanation for two otherwise mysterious (in)stability properties of probabilistic regularities in regard to conditionalization: screening off and linking up.

2.2. Explaining screening off

Since screening off and linking up are the major explananda of causal relations, we have to characterize them in an empirical, i.e. *purely probabilistic* way, without presupposing causal notions. We first turn to screening off:

(2) X and Y are *screened off* by Z *iff* (i) DEP(X,Y) and (ii) INDEP(X,Y|Z).

Examples:

(2.1) Barometer reading (X) storm coming (Y) atmospheric pressure (Z)
(2.2) Light switch (X) Light bulb (Y) Electric current (Z)

The probabilistic dependence between X and Y disappears when one conditionalizes on arbitrarily chosen but *fixed* values of a third variable Z. Condition (2) implies the probabilistic dependencies DEP(X,Z) and DEP(Y,Z).[2] We assume the usual case that these dependencies are not screened off by Z. Moreover, we focus on *robust* (or faithful) cases of screening off in which the disappearance of the probabilistic X-Y dependence after conditionalization on Z is *stable* under small changes of the involved conditional probabilities (we shall see in section 3.2 that most cases of screening off are robust in this sense.)

Intuitively, we immediately interpret the correlations in (2.1) and (2.2) as produced by directed causal relations: We believe that we "know" that screening off occurs because Z is a common cause in (2.1) and an intermediate cause in (2.2). In order to achieve a philosophical justification of causality, however, we must *free our mind* from prefabricated causal intuitions and assume for a moment that we only know the variables' probability distributions. If we do that, we are confronted with a riddle: *Why* does the X-Y correlation disappear when fixing Z's value?

The *best* available explanation of robust screening off phenomena —in fact, the only good explanation we can think of— is the following: the two dependencies between Z and X and between Z and Y directly reflect causal connections,[3] while the dependence between X and Y results from these causal connections and is thus *mediated* (or *transmitted*) by Z.

---

[2] *Proof:* We have $P(X|Y) = \Sigma_z P(X|z,Y) \times P(z|Y)$. Assume the contrary: INDEP(Z,Y), i.e. $\forall z$: $P(z|Y) = P(z)$. Moreover $\forall z$: $P(X|z,Y) = P(X|z)$ by the screening off condition (3)(ii). So we continue: ... $= \Sigma_z P(X|z) \times P(z) = P(X)$. Thus INDEP(X,Y) follows, contradicting the assumption. (For INDEP(Z,X) the proof is similar, only X and Y exchange their role.)

[3] The claim that the causal connection between X and Z in fig. 1 is "direct" is relative to the set of variables {X,Y,Z}.

This situation is depicted in fig. 1. If we consider subsets of individuals with different X-values, these individuals will have differently distributed Y-values *only because* they have differently distributed Z-values. So if we conditionalize on a subdomain of individuals with fixed Z-values, individuals with different X-values will no longer have differently distributed Y-values, i.e. the probabilistic dependence will no longer be transmitted from X to Y.

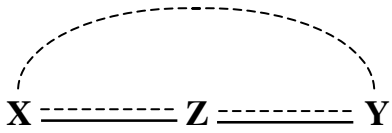

Fig. 1. Explanation of screening off by binary causal relations ("..." stands for "probabilistic dependence" and "–" for "direct causal connection").

Note that explaining screening off only requires assuming an *undirected binary* "causal" dependence relation; no direction of causation is needed so far. Directed causation is, however, required for discriminating screening off from linking up. Moreover, when we call the dependence relation "causal" we presuppose, of course, that the involved variables refer to *pairwise analytically* (logically, mathematically) *independent* attributes or types of events. If variables are analytically dependent, then the appropriate explanation of a screening off relation is not a causal, but a semantic connection between variables (for illuminating examples cf. Williamson 2005, sec. 4.2).

Before we turn to linking up and directed causation, we show why prominent alternative attempts fail in explaining screening off.

First of all, *duplication accounts* cannot explain screening off. Duplication accounts come in two varieties: (i) Humean-reductionistic (causality is "nothing but" correlation) and (ii) naive-metaphysical (every correlation is "backed up" by a corresponding causal connection). Both types of account would postulate a direct causal connection between *every* two correlated variables X and Y, as shown in fig. 2(a). But then they cannot explain why Z screens off X from Y; this fact would remain mysterious. It is precisely the assumption that *not all* correlations correspond to direct causal connections which explains screening off.[4]

Fig. 2. (a) Duplication accounts cannot explain why DEP(X,Y) vanishes. (b) The blocking theory cannot explain why DEP(X,Y) vanishes for all Z-values on which one conditionalizes.

[4] There are more refined versions of duplication accounts which have similar problems. According to a referee of this paper, a Humean duplication account should assert that *causality is nothing but* the pattern of probabilistic (in)dependence. However, this doesn't seem to be correct, since the same type of probabilistic pattern —namely a conditional correlation— may be produced by different causal patterns (common causes, indirect causes, common effects, etc.).

A second alternative explanation of screening off would be the *blocking* account: Some Z-values can block the causal connection between X and Y, as depicted in fig. 2(b). But this hypothesis cannot explain why the X-Y correlation vanishes when conditionalizing on *arbitrary* Z-values. It seems that the only explanation that works is the one given above: Z screens X off from Y because Z mediates X's dependence on Y.

A final objection might point out that it is impossible even to *ask* for an explanation of screening off without already presupposing causal notions, because all explanations *are* causal. However, we don't understand the explanation of screening off in a causal sense for otherwise we would end up in an infinite regress. In accordance with many philosophers of science, we assume that there is a non-causal sense of "explanation" which consists in unification and in the generation of potential predictions.

2.3. Explaining linking up

Let us now turn to the phenomenon of linking up. Some sets of variables {X,Y,Z} have probability distributions that feature exactly the opposite (in)stability properties of screening off. We call this phenomenon "linking up" and define it again in a purely probabilistic (i.e. non-causal) way:

(3) X and Y are linked up by Z *iff* INDEP(X,Y) and DEP(X,Y|Z).

Example:

Angle of the sun (X) length of a tower (Y) length of its shadow (Z)

Two independent variables X and Y become linked up by Z *iff* they become dependent after conditionalization on some values of Z. The position of the sun, for example, is not correlated with the height of a tower, but it becomes correlated if we conditionalize on the shadow's length. If the tower's shadow is long, for instance, we can infer that the solar altitude must be low if the tower is short. As in screening off scenarios, (3) implies DEP(X,Z) and DEP(Y,Z). Again we focus on robust cases of linking up.

Let us once more put aside prefabricated causal intuitions. Then we face a second riddle: Why do two formerly independent variables X and Y become correlated when we conditionalize on certain Z-values? Undirected causal relations cannot explain *both* screening off and linking up. To explain linking up, Z must again act as a *mediator* between X and Y. So the undirected causal relations in the linking up scenario must have the *same* structure as in the screening off scenario depicted in fig. 1. But if the causal structure should be able to explain both screening off and linking up, it cannot be the same in these two cases, because the two phenomena involve opposite probabilistic (in)stability effects.

The best available explanation for screening off *and* linking up —again the only good explanation we can think of— is to assume that causal relations are *directed*: In what follows "X → Y" expresses that X exerts a causal influence on Y that is "direct", i.e. unmediated relative to the given the set of variables V. The way that this direct causal influence is physically realized is not specified by TC. However, two assumptions are required that are precisely formulated in sec. 3.1 and *informally* stated as follows:

— *Productivity (P):* "Ceteris absentibus" X → Y implies a probabilistic dependence between X and Y, and

— *Causal connection (C):* Probabilistic dependencies are the result of directed causal connections, which transmit probabilistic influence from causes to effects, but not from effects to causes.

We can now explain screening off *and* linking up phenomena as follows. In both cases, Z mediates between X and Y. So we have three possible directed causal structures as candidates for explaining these phenomena:

(a) $X \to Z \to Y$ (or $X \leftarrow Z \leftarrow Y$): Z is an intermediate cause (between X and Y).
(b) $X \leftarrow Z \to Y$: Z is a common cause (of X and Y).
(c) $X \to Z \leftarrow Y$: Z is a common effect (of X and Y).

The first two structures explain screening off; the third one explains linking up.

Explaining screening off

(a) Intermediate cause ($X \to Z \to Y$): Y depends on X because a change of X-values causes a change of Z-values which, in turn, causes a change of Y-values (DEP(X,Y)).
(b) Common cause ($X \leftarrow Z \to Y$): Y depends on X because changes of X-values are caused by changes of Z-values which also cause changes of Y-values (DEP(X,Y)).

In case (a) as well as case (b), X-value variations can lead to Y-value variations only due to Z-value variations; thus fixing the value of Z renders X and Y independent (INDEP(X,Y|Z)).

The logical structure of both explanations is this: From $X \to Z$ in case (a), or from $Z \to X$ in case (b), we infer DEP(X,Z) by (P); likewise we infer DEP(Z,Y) from $Z \to Y$ and (P). For explaining DEP(X,Y) we must assume that the the causal models (a) and (b) satisfy the following condition of *dependence transitivity* (DT): $\exists x,y: \Sigma_{z \in Val(Z)} P(y|z) \times P(z|x) \neq \Sigma_{z \in Val(Z)} P(y|z) \times P(z)$. (DT) is not probabilistically valid (except for binary variables); rather (DT) is a kind of faithfulness assumption (see sec. 3). Given (C), (DT) is a sufficient (and necessary) condition for DEP(X,Y) to hold in the causal structures (a) and (b) (a proof of this is given in (11.2)(a), sec. 3.1). INDEP(X,Y|Z) follows from condition (C) applied to the causal structures (a) and (b); this is obvious from (C)'s precise formulation in sec. 3.1, since X is not d-connected with Y given Z is fixed. This completes the explanation.

Explaining linking up:

(c) Common effect ($X \to Z \leftarrow Y$): Y doesn't depend on X because a change of X-values causes a change of Z-values which, however, is not accompanied by a change of Y-values, because value-changes are not transmitted from an effect to its cause. Fixing Z to certain values will render X and Y dependent (DEP(X,Y|Z), as explained in the sun-tower-shadow example (3).

The logical structure of this explanation starts again with the observation that by axiom (P), $X \to Z$ and $Z \leftarrow Y$ imply DEP(X,Z) and DEP(Z,Y), respectively. By (C), no probabilistic influence of a cause X on its effect Z can be transmitted to Z's other cause Y; so "ceteris absentibus" X and Y are probabilistically independent, i.e. INDEP(X,Y). In order to explain DEP(X,Y|Z) we must assume the condition of *dependence-overlap* (DO): $\exists x,y,z: Dep(x,z) \wedge Dep(z,y)$. Like condition (DT), condition (DO) is a kind of faithfulness assumption (a proof that INDEP(X,Y) and (DO) imply DEP(X,Y|Z) for the causal structure (c) is given in (11.2)(b), sec. 3.1). This completes the explanation.

We now turn to the discussion of some further alternative explanations in terms of directed causal arrows. A well-known alternative is the time-honoured world view of *occasionalism* that was held as an alternative to the causal-naturalistic world view. Occasionalism claims that correlations between (kinds of) events are not to be explained by cause-effect relations between these events; rather all events are the direct effect of God's will. Thus, God is the common cause of all events which are correlated because of their being joint effects of this cause. However, if this were true, it would be impossible to screen off two correlated events from each other by fixing the value of a third event: this possibility can only be explained by assuming direct cause-effect relations between the events themselves (see fig. 3).

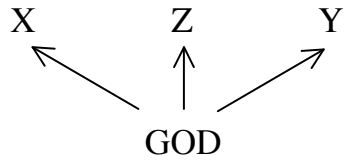

Fig. 3. Alternative explanation by occasionalism: This world-view cannot explain why it is possible to screen off X from Y by conditionalizing on Z.

A defender of occasionalism may point out that screening off occurs in fig. 3 because God wants only certain but no other *combinations* of the values of the variables X, Y, and Z. This would mean that God causes the variables of our world not by separated causal arrows, but *holistically*, in the sense that we have to draw just *one* causal arrow from God to the entire system of variables: GOD → (X,Y,Z). This is a possible but less informative and less unifying explanation, since the explanatory work is no longer done by the causal structure, but entirely by the probability distribution P(GOD,(X,Y,Z)).

With help of directed arrows, we are even able to explain why certain cases of screening off and linking up are *non-robust*. A non-robust scenario in which Z screens off X from Y could, for example, be explained by the causal structure in fig. 4(a): the positive conditional dependence due to $X \rightarrow^{+} Z \ ^{+}\leftarrow Y$ and the negative dependence due to $X \rightarrow^{-} Y$ cancel out. Thus, DEP(X,Y) and INDEP(X,Y|Z), though Z is a common effect of X and Y. Schurz/Gebharter (2015) call this situation one of *unfaithfulness due to cancelling paths*. Unfaithful independencies are not robust, since small changes of the involved conditional probabilities turn them into dependencies. An analogous alternative explanation can be given for the non-robust linking up case in fig. 4(b), in which the positive dependence due to $X \rightarrow^{+} Z \rightarrow^{+} Y$ and the negative dependence due to $X \rightarrow^{-} Y$ cancel out.

Fig. 4. Unfaithfulness due to cancelling paths: (a) explains non-robust screening off (DEP(X,Y) and INDEP(X,Y|Z)); (b) explains non-robust linking up (INDEP(X,Y) and DEP(X,Y|Z)).

In our view, the major advantage of the proposed justification lies in the fact that it neither presupposes advanced concepts of physics nor strong metaphysical assumptions. It rather justifies causality on the basis of ordinary phenomena in everyday life. In particular, we experience screening off and linking up in all kinds of *purposeful* actions. Here our actions (A) realize certain means (M) in order to produce certain purposes (P) (A → M → P); so DEP(A,P) holds, but M screens off A from P, i.e. our actions cannot reach their purposes without realizing certain means. Moreover, if a purpose P can be achieved by two independent means M_1 and M_2 (e.g. shooting an animal by two independent guns), then the achievement of the purpose P links up M_1 and M_2 (if gun M_1 missed the target, M_2 must have hit it successfully). These facts help to explain (i) why causality is an inborn reasoning mechanism of homo sapiens and (ii) why causality is so closely connected with interventions.

Let us finally compare our IBE justification strategy with the *fork asymmetry* argument, which goes back to Reichenbach (1956, 159-61) and has been elaborated by Papineau (1992). It runs as follows: Assume X and Y are two events, both correlated with a third event Z. Then either (a) X and Y are mutually correlated and Z screens them off, in which case Z is a common cause of X and Y; or (b) X and Y are uncorrelated, in which case Z is a common effect of X and Y. However, this justification strategy has gaps. As Papineau (1992, 240) observes, the argument doesn't work if X and Y can causally reach each other; Reichenbach excluded this case by assuming that X and Y are temporally simultaneous. Another gap is the third possible case (c) in which X and Y are correlated but Z doesn't screen them off, because X, Y, and Z are joint effects of a common cause C. In contrast, our proposed justification strategy doesn't suffer from these restrictions. It is not based on the fork asymmetry, but on the asymmetry between screening off and linking up, which is not considered by Reichenbach or Papineau.

2.4. Justification of causality in deterministic universes

We finally discuss the question of *determinism*. Our justification of directed causality presupposes that the probabilistic relations of screening off and linking up relations are asymmetric. Thus, in the screening off example (2) we assumed that Z screens off X and Y, but neither does X screen off Z from Y, nor does Y screen off Z from X. We made a similar assumption for the linking up relation in example (3). However, this assumption is only satisfied if the given causal model is *indeterministic*, that is, if the involved probabilities are different from 0 or 1.

We say that a sequence of variables U *determines* another variable Y if for every U-value u there exists one (and hence only one) Y-value y such that P(y|u) = 1. (Note: U can also be a single variable.) This definition covers not only the case in which an effect depends deterministically on its cause(s), but also the case in which a cause depends deterministically on its effect(s), or in which two effects depend deterministically on each other. In our example (2), if X determines Z, then it is impossible to vary Z's value when X's value is fixed, whence P(y|z,x) = P(y|x) = 1 or = 0 will hold for every X-value x. So INDEP(Z,Y|X) would hold, i.e. X would screen of Y from Z. Let us say that U screens off Y from Z *trivially* iff neither Y nor Z can varied when U is fixed: This implies that P(y|z,u) is either undefined (iff P(z|u) = 0) or P(y|z,u) = P(y|u) (iff P(z|u) = 1). In conclusion, if example (2) represents a situation of deterministic causality, then we get the result that X screens off Y from Z trivially.

The situation is illustrated in fig. 5. In the case of a merely *one-sided* determinism, where the cause determines its effect but *not* vice versa, the situation is still not symmetric, since Z's value is not determined by X and, thus, Z is a *non-trivial* off-screener of X from Y (see fig. 5(a)). So we can still identify the intermediate node Z by purely empirical (statistical) relations, and the explanation of these statistical relations by an intermediate or common cause structure is warranted. According to a full-blown deterministic world view, however, one assumes a *two-sided* determinism, in which not only the cause determines its effects, but the cause can also uniquely be recovered from its effects, or in other words, the cause is determined by its effects. Under this assumption (see fig. 5(b)) the empirical screening off relations are fully symmetric and the suggested justification of causality by an inference to the best explanation (IBE) *breaks down* completely. In fig. 5, "$\overset{1}{\rightarrow}$" abbreviates "P(effect|cause) = 1" and "$\overset{1:1}{\rightarrow}$" abbreviates "P(effect|cause) = P(cause|effect) = 1".

Fig. 5. Deterministic screening off. (a) One-sided determinism: X is a trivial off-screener, but Z is a non-trivial off-screener. *(b) Two-sided determinism:* X, Y and Z are trivial off-screeners; so the situation is completely symmetric. (The situation is exactly alike if Z is not an intermediate but a common cause of X and Y.)

Does this mean that our IBE-justification of directed causality doesn't apply to a fully deterministic world? The answer is: *No* –at least not in the circumstances that are *typical* for our world. It is indeed true that our IBE couldn't work if the considered structures were deterministically *complete*, in the sense that they would contain *all* variables which are needed to fully determine their value. However, the physical systems in our world are typically influenced by a multitude of minor 'disturbance' or *noise* factors, while the complete set of causes is usually unknown. In other words, *even if* our world is truly deterministic, our causal models are typically incomplete and hence indeterministic – a situation which is called *pseudo-indeterminism* (SGS, §2.5). A situation of this kind is illustrated in fig. 6: the part of the deterministic world which is encircled by the dashed line represents what we know about it; the variables D_i represent unknown disturbance or noise variables. Given that the noise variables are mutually independent and D_i is independent from X_{i-1} (what is usually assumed), all of our observations apply equally to pseudo-indeterministic structures of this sort: thus X_i screens off X_{i-1} from X_{i+1} but X_{i-1}, doesn't screen off X_i from X_{i+1}, etc.

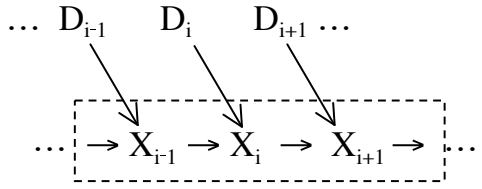

Fig. 6. Screening off in a pseudo-indeterministic substructure (encircled by the dashed line) which is part of a deterministic world; the variables D_i represent the unknown remainder causes of X_i. Because of $P(X_i | X_{i-1}) < 1$, X_i screens off X_{i-1} from X_{i+1} non-trivially.

One might object that our assumption that the noise variables D_i are independent is a "trick", because in a truly deterministic universe they are themselves determined and hence mutually dependent and dependent on the X_{i-1}. But the last part of this argument is wrong: Surely, the D_i are themselves determined in a deterministic universe, but all that one needs in order to allow for independent variations of the D_i and the X_i are *enough independent initial causes* at the start of the universe. In fig. 6, we need at least one starting cause X_0 for the X_i-chain and for each variable D_i one independent starting cause $D_{i,0}$, to produce these independencies at least approximately. To be sure, all these "initial causes" can be summarized by a one multi-dimensional *initial state variable* U_0 representing the universe at its beginning. What is important here is that sufficiently many different values of this initial state variable were realized in different parts (individuals or subsystems) at the beginning of the universe. This means, in other words, that the entropy of the universe at its beginning must have been sufficiently *low*. This observation brings our account, if applied to deterministic universes, in a close relationship to those accounts of deterministic causality that base causality on the continuous increase of entropy (cf. Frisch 2007).

3. *Structure and content of the theory of causality TC*

3.1. *Core axioms of TC: d-connection (Markov) and productivity (minimality)*

In this subsection we present the exact explication of the axioms of causal "d-" connection (C) and productivity (P) that have been justified by an IBE in sec. 2. (C) and (P) constitute TC's core and are traditionally expressed by the equivalent causal Markov condition (M) and the minimality condition (Min), respectively (SGS, §3.4.1-2). We prefer (C) and (P) over (M) and (Min), because (C) and (P) are philosophically more transparent and better suited for expressing TC's full content. Before we explicitly state (C) and (P), we have to introduce the following notions. A *causal graph* (or *structure*) is a pair (V,E), where V is a set of variables (the "vertices"), and $E \in V \times V$ is a set of directed arrows $X_i \rightarrow X_j$ (i.e. ordered pairs, the "edges"). A graph (V,E) together with a probability-distribution P over V is called a *causal model* (or *system*) (V,E,P). Causal structures and systems are parts of the world, while causal graphs (CGs) and models (CMs) are conceptual representations of causal structures and systems, respectively. Some further important notions:

> $X \rightarrow Y$: X is a *direct cause* of Y (Y is a *direct effect* of X).
> $X \rightarrow\!\rightarrow Y$: X is a (direct or indirect) *cause* of Y (Y is an *effect* of X), i.e. there is a directed path $X \rightarrow Z_1 \rightarrow \ldots \rightarrow Z_n \rightarrow Y$ from X to Y (for $n \geq 0$).
> $X - Y$: $X \rightarrow Y$ or $X \leftarrow Y$, i.e. X and Y are *adjacent*.
> $X_1 - - X_n$: a *path* $X_1 - \ldots - X_n$ between X_1 and X_n that *connects* X_1 and X_n; the variables X_i $(1 \leq i \leq n)$ are said to *lie* on this path.
> If X_i lies on a path π, then X_i is called (i) a common cause, (ii) an intermediate cause, or (iii) a common effect on π iff (i) $\leftarrow X \rightarrow$, (ii) $\leftarrow X \leftarrow$ or $\rightarrow X \rightarrow$, or (iii) $\rightarrow X \leftarrow$, respectively, is part of π.

A CG (or CM) is called *acyclic* iff it contains no cyclic paths $X \rightarrow\!\rightarrow X$. In this paper we concentrate on acyclic CMs, although the theory TC is not necessarily restricted to

them: it can also be applied to cyclic graphs, though some famous results (e.g. theorem 1 below) do not hold in this case (cf. Schurz 2013, 397f).

The principle of causal connection or "d-connection" says that every (conditional) probabilistic dependence between two variables X and Y is the result of some causal path connecting them. The correct formulation of this principle has to account for all possible combinations of screening off and linking up along all paths connecting X and Y. If a path π connects X and Y in a CG (V,E), then π can generate probabilistic dependence conditional on a (possibly empty) subset of variables $U \subseteq V - \{X,Y\}$ only if *no* common or intermediate cause on π is in U and *all* common effects on π are in U, or have an effect in U. The disjunctive clause "or have an effect in U" is needed because two independent variables X and Y may not only be linked up by common effects, but also by effects of common effects, as for example by the variable Z' in the frame consisting of $X \rightarrow Z \rightarrow Y$ and $Z \rightarrow Z'$.

If X and Y are connected in a graph (V,E) by *several* paths, then X and Y become dependent *iff at least one* of these paths generates a probabilistic dependence. These considerations are summarized in the following axiom of causal connection (C):

(4) *Axiom of causal connection (C):* Every physically possible CM (V,E,P) [in an intended domain] satisfies the *condition of causal connection (C)*, which is *defined* as follows:
For all $X,Y \in V$ and $U \subseteq V-\{X,Y\}$: *If* DEP(X,Y|U), *then* X and Y are *d-connected* given *U* in the following sense:
X and Y are connected by some path π such that no intermediate or common cause on π is in U, while every common effect on π is in U or has an effect in U. (In this case we say that X and Y are *d*-connected given U *by* path π.)

Condition (C) entails the following well-known principle:

(5) *Unconditional dependence:* If DEP(X,Y), then X and Y are connected by a directed or common cause path (i.e., are d-connected given $\varnothing$).

If X and Y are not d-connected given U, X and Y are said to be *d-separated* by U. The concepts of d-separation and d-connection were developed by Pearl (1988, 117). (C) asserts an *implication* from a (probabilistic) dependence to a d-connection – or, in contraposed form, an implication from a d-separation to an independence, which is the formulation used by Pearl (1988, 119; 2000, th. 1.2.4). If X and Y are connected by a path π but not d-connected by π given U, then U is said to *block* path π.

Condition (C) is equivalent with the famous causal Markov condition (cf. Pearl 2009, 16; SGS 29f):

(6) *Definition of the causal Markov condition (M)*:
(V,E,P) satisfies (M) *iff* every $X \in V$ is independent of its non-effects conditional on the set of its direct causes or "parents" par(X), i.e., INDEP(X,U|par(X)) holds for every subset $U \subseteq V$ of non-effects of X.

(7) *Theorem 1:* For every acyclic CM: (C) and (M) are equivalent.

SGS (sec. 3.4.1-2) prefer (M) as the core axiom of TC. The equivalence between (M) and (C) has been demonstrated by Verma and Pearl (s. Pearl 1988, 119f, th. 9, cor. 4), and by Lauritzen et al. (1990, 50), who call (C) the *global* and (M) the *local* Markov condition (see

also SGS 46, th. 3.3). Observe that in its contraposed form, (C) asserts a (conditional) independence for *all* d-separation relations of a causal graph, while (M) asserts such an independence only for the d-separation relations between a variable X and its non-effects conditional on its parents; the other independencies *follow* from these as probabilistic consequences. For this reason, (C) expresses the *full* content of the causal Markov condition in a much more direct way than (M).

In (4) we distinguished between the *definition* of the d-connection condition and the corresponding *axiom* which states that this condition holds for all physically possible causal models [in an intended domain]. SGS (29) speak likewise of "axioms", but present them as definitions; Pearl (2009) only states the definitions. The formulation as axioms impels us to critically reflect upon the *problem of generality*: do *all* correlations between analytically independent variables really result from causal connections? The justification of causality by an IBE in sec. 2 works only for correlations that participate in relations of screening off or linking up. We argued in sec. 2 that all correlations that can be utilized by means of interventions participate in screening off and/or linking up relations. However, not all correlated variables can be intervened on. Possible failures of the causal Markov condition have been discussed in the context of EPR-correlations in quantum mechanics, in which the correlated states of two entangled particles are not screened off by common causes (cf. Hausman 1998, 252; Healey 2009). Cartwright (2007, 122) argues that similar problems may even arise in ordinary (macroscopic) domains. Well-taken defences against these objections have been given by the proponents of TC (SGS 59-63; Pearl 2009, 62, Hitchcock 2010), though not all problems are solved by these defences and the debate is ongoing. In this paper we don't take a stance on whether (C) is strictly general or holds only in certain domains. To account for the latter possibility we inserted "[in an intended domain]" in axiom (C). In any case, we understand (C) as a synthetic (i.e. not analytically true) principle whose content can be true or false in the realistic sense, and we believe that (C) holds for most physically closed systems. Moreover, (C) provably holds for every *subsystem* of a (C)-satisfying causal system (V,E,P) that is *causally sufficient*, i.e., that doesn't omit any true and non-degenerate common cause of variables in V (cf. SGS, 22).

Axiom (C) asserts that a probabilistic dependence implies a causal connection. The second core axiom, (P), asserts that the other direction, from causal connection to probabilistic dependence, holds under *special* conditions: a *direct* causal connection implies *ceteris absentibus* a probabilistic dependence. (P) holds only under certain conditions because *unfaithful* structures due to *cancelling paths* (fig. 4, sec. 2.3) cannot be excluded a priori. We can isolate X's causal influence on Y from the influence of possibly cancelling paths in unfaithful causal models by conditionalization on all of Y's parents that are different from X (par(Y) – {X}). So the axiom of productivity is explicated as follows:

(8) *Axiom of productivity (P):* Every physically possible CM [in an intended domain] satisfies the *condition of productivity (P)*, which is defined as follows:
For all X, Y in the CM: If $X \rightarrow Y$, then DEP(X,Y|par(Y)–{X}) holds.

For a closer investigation of the condition of productivity see Schurz/Gebharter (2015, sec. 2.3). In distinction to the first core axiom, (P) is justified by a methodological requirement: in order to be empirically significant, causal arrows must be responsible for at least some (conditional) probabilistic dependence; causal arrows without empirical effects are

eliminated by Ockham's razor. Given (C), (P) is equivalent with the well-known minimality condition (Min) (SGS 31):

(9) *Definition:* A (C)-satisfying CM (V,E,P) is minimal (i.e. satisfies (Min)) *iff* no arrow can be omitted from E without violating condition (C), i.e. every submodel (V,E',P) with $E' \subset E$ violates (C).

(10) *Theorem 2:* For all (C)-satisfying and acyclic CMs, (Min) and (P) are equivalent.

Axiom (P) is new; a proof of theorem 2 is found in Schurz/Gebharter (2015, th. 2). The advantage of (P) over (Min) is twofold: First, (Min) tells us only that every arrow $X \rightarrow Y$ is needed to explain some dependence within the given CM, while (P) states this dependence explicitly. Second, (P) is independent of (C), while (Min) presupposes (C).

Based on the precise explications of TC's axioms, we are able to prove that the explanations of screening off and linking up given in sec. 2.2 and 2.3 are indeed entailed by (C) and (P), as follows:

(11) Assume a CM of the form (a) $X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \rightarrow Y$, or of the form (b) $X \rightarrow Z \leftarrow Y$, which satisfies conditions (C) and (P). Then:
(11.1) (C) entails INDEP(X,Y|Z) in case (a), and INDEP(X,Y) in case (b).
(11.2) Condition (DT) ($\exists x,y: \Sigma_{z \in Val(Z)} P(y|z) \times P(z|x) \neq \Sigma_{z \in Val(Z)} P(y|z) \times P(z)$) entails DEP(X,Y) in case (a), and condition (DO) ($\exists x,y,z: DEP(x,z) \wedge DEP(z,y)$) entails DEP(X,Y|Z) in case (b).

Proof of (11): (11.1) is obvious.

For (11.2): Case (a): By probability theory we have (a) $P(y|x) = \Sigma_z P(y|x,z) \times P(z|x)$ and (b) $P(y) = \Sigma_z P(y|z) \times P(z)$. The sum in (a) equals (c) $\Sigma_z P(y|z) \times P(z|x)$ by condition (C) of sec. 2.3, since Y is not d-connected with X given Z. It follows that $P(y|x) \neq P(y)$ holds exactly if the two sums in (c) and (b) are unequal, i.e., if (DT) holds.

Case (b): By probability theory, (a) $P(x|y) = P(x|y,z) \times P(z|y) + P(x|y, \neg z) \times P(\neg z|y)$ and (b) $P(x) = P(x|z) \times P(z) + P(x| \neg z) \times P(\neg z)$. By INDEP(X,Y) we have $P(y|x) = P(y)$. So the sums in (a) and (b) must be equal. These sums are weighted averages, with the weights in the sum in (a) being $P(z|y)$ and $P(\neg z|y)$, and the weights in the sum in (b) being $P(z)$ and $P(\neg z)$. By (DO) we have (i) $P(z|y) \neq P(z)$ and (ii) $P(x|z) \neq P(x| \neg z)$. It follows from (i), (ii), and the laws of weighted averages that the two sums in (a) and (b) would have to be different if INDEP(x,y|Z), i.e. $P(x|y,z) = P(x|z)$ and $P(x|y, \neg z) = P(x| \neg z)$, would hold. Thus either $P(x|y,z) \neq P(x|z)$ or $P(x|y, \neg z) \neq P(x| \neg z)$ must hold, which gives us DEP(X,Y|Z). Q.E.D.

3.2. The empirical content of TC

In this section 2 we showed that TC's core axioms offer the best (if not the only) available explanations of screening off and linking up. These explanations are highly unifying: In terms of the account of Schurz/Lambert (1994), a large number of statistical (in)dependence relations is reduced to a much smaller number of directed causal arrows plus a few general principles. Since the general axioms of TC are the same in all applications, their weight vanishes according to Schurz/Lambert's unification account. In the screening off example with three variables, six (in)dependence relations DEP(X,Z), DEP(X,Z|Y), DEP(Z,Y), DEP(Z,Y|X), DEP(X,Y), INDEP(X,Y|Z) are reduced to two causal arrows $X \rightarrow Z$ and

$Z \rightarrow Y$. More generally speaking, in a Bayes net with n variables, there are $n \times (n-1)/2$ pairs of variables that can be conditionalized to at most $2^{(n-2)}$ subsets of variables; thus $n \times (n-1) \times 2^{(n-3)}$ conditional (in)dependences are reduced to $n \times (n-1)/2$ causal connections, which is a reduction by a factor of $2^{(n-2)}$.

However, in order to be empirically significant, this is not enough. Causal explanations should not be entirely post facto, but should be able to generate *empirical content* by means of which TC is independently testable. To make this notion precise, we define it as follows: An *empirical* model is a pair (V,P) of a set of empirically measurable variables V together with a probability distribution P over V. An empirical model (V,P) is called an empirical *submodel* of a CM (V',E',P') iff $V \subseteq V'$ and $P = P' \upharpoonright V$ (the restriction of P' to V).[5] We also say that (V',E',P') *expands* (V,P). We define:

(12) *Empirical content of TC:* A version of TC has empirical content *iff* there exists a logically possible empirical model that cannot be expanded to a CM satisfying this version of TC.

If TC would *not* have empirical content, the content of all particular causal models would be entirely ex-post. For any empirical model (V,P) one could then invent a "causal explanation" in accordance with TC, i.e., a TC-model (V',E',P') that expands (V,P). If this were the case, TC would be exposed to the objection of being superfluous "causal metaphysics".

In investigating TC's empirical content we follow the analogy between causality in TC and force in classical physics mentioned in sec. 1. As the total force law (sum of forces = mass × acceleration) and the actio-equals-reactio law constitute the core of classical physics, (C) and (P) constitute the core of TC. But there are further general principles, such as faithfulness (F), external noise (EN), temporal forward-directedness (T), locality (L) and probabilistic freedom of interventions (Fr), which are introduced in sections 3.3. These principles constitute *extended versions* of TC, just like the laws of gravitational or frictional force constitute extended versions of Newtonian physics.

According to our knowledge, the first investigation of the question of TC's empirical content by logical means is undertaken in Schurz/Gebharter (2015). Their first result, stated in theorem 3, is negative: (C)+(P) alone don't have empirical content, not even when the condition of acyclicity is added:

(13) *Theorem 3:* Every analytically possible empirical model (V,P) can be expanded to an acyclic CM (V',E',P') satisfying (C) and (P). (Proof in Schurz/Gebharter 2015, th. 3.)

Technically, theorem 3 is unspectacular. Its philosophical consequences, however, deserve critical reflection. Proponents of TC have often argued that (C) or the equivalent causal Markov condition (M) is satisfied by all (or most of all) known empirical and/or technical systems (SGS 29; Pearl 2009, 62f; Hitchcock 2010, sec. 3.3). However, since TC's core axioms are empirically empty, it is impossible to confirm (C) without additional assumptions. However, no such additional assumptions are stated in the quoted passages. The same problem applies to critics of (C), such as Cartwright (2007, 122): to turn their examples into counterexamples to (C), they must make further assumptions about causality.

[5] Empirical submodels correspond to what is called "partial (potential) models" in structuralist philosophy of science (cf. Balzer et al. 1987; Sneed 1971, ch. 3).

Is it a problem that TC's core is empirically empty? Not necessarily. Sneed (1971, 126) has demonstrated with scrutiny that the core of classical physics, the total force law, is also empirically empty. For every system of (point) masses with given accelerations one can construct force functions that satisfy the total force law. However, it is well-known that the empirical content of general classical physics abruptly increases when special force laws (e.g. the law of gravitational force) are added (cf. Schurz 2013, ch. 5). Do we meet a similar situation in the case of TC? The answer to this question given in Schurz/Gebharter is "yo": Empirical content can indeed be added, but not as easily and as much as in the case of physics.

The axioms (C) and (P) are purely *structural* insofar as they do not make any assumptions about the "substance" or physical nature of the cause-effect relation. There is a further purely structural principle of causality, which is a strengthening of (P) and the exact inverse of (C), namely the condition of *faithfulness*, which is defined as follows (cf. SGS 31, Zhang and Spirtes 2008, 24):

(14) *Definition of the faithfulness condition (F):* (V,E,P) satisfies (F) *iff* (V,E,P) satisfies the converse of (C): *if* X and Y are d-connected given $U \subseteq V-\{X,Y\}$, *then* DEP(X,Y|U).

In other words, a CM is faithful iff P verifies *only* those probabilistic independence relations that are implied by (C). It is easily seen that:

(15) Faithfulness implies productivity.

However, faithfulness is logically stronger than productivity. Contrary to (P), (F) has various exceptions. The most important kind of an unfaithful causal model has been explained in sec. 2.3: here the probabilistic effects of several causal paths that d-connect two variables X and Y cancel each other out, so that INDEP(X,Y) results although X and Y are d-connected.

It is easy to see that axiom (C) plus condition (F) have empirical content. A result of this kind can be found in Zhang and Spirtes (2008, 253), though not in terms of content, but in terms of "detectable kinds of unfaithfulness". As explained in Schurz/Gebharter (2015, sec. 3.2), Zhang and Spirtes' theorems imply the following results for the empirical content of (C)+(F):

(16) *Theorem 4:* (C)+(F) have empirical content: No empirical model (V,P) with $\{X,Y,Z\} \subseteq V$ verifying the logically possible (in)dependence relations in (a) or (b) can be expanded to a CM (V',E',P') satisfying (C) + (F):

(a) $\forall U \subseteq (V-\{X,Y\})$: DEP(X,Y|U) $\wedge$ DEP(Y,Z|U), and there exist two distinct sets $\mathbb{W},\mathbb{W}' \subseteq V-\{X,Z\}$ with $Y \in \mathbb{W}$ and $Y \notin \mathbb{W}'$, both of which screen off X from Z

(b) INDEP(X,Y), INDEP(Y,Z), INDEP(X,Z), DEP(X,Y|Z), DEP(Y,Z|X), DEP(X,Z|Y).

Since unfaithful causal systems exist, (F) would be empirically false if it were formulated as a strictly general axiom. Proponents of TC argue that (F) is highly probable, i.e. satisfied by almost all empirical models. These arguments are based on the fact that unfaithful CMs are *parameter instable* in the following sense: their unfaithful inde-

pendencies can be destroyed by arbitrary small changes of the probability distributions P(X|par(X)) of the variables X conditional on the set of their parents. These probability distributions are called the causal model's *parameters*. The parameters of acyclic models can be varied independently from each other without destroying the independencies entailed by (C).

(17) *Lemma 1:* A (C)-satisfying acyclic CM (V,E,P) is faithful *iff* it is parameter-stable (cf. Pearl 2009, 48, def. 2.4.1)

Lemma 1 implies that for every probability measure over the set of parameters of a CM which is "smooth" (i.e. absolutely continuous with the Lebesgue measure over $[0,1]^P$), the probability of unfaithfulness is *zero* (cf. SGS 41f; Steel 2006, 313). This means more concretely that in every physical causal system whose causal parameters are underlying small variations by external disturbations, the faithfulness condition will hold with near certainty. In conclusion, the faithfulness condition has not strict-deductive but merely *probabilistic-inductive* content.

A further possibility of strengthening TC is the condition of temporal forward-directedness (T). To make this condition precise, we define a *causal event-model* (V,E,P,t) as a CM whose variables are event-variables, together with a time function t: V → Reals, where t(X) is the time point at which the possible X-values (events) x occur.

(18) *Axiom of temporal forward-directedness (T):* Every physically possible causal event model (V,E,P,t) [in an intended domain] satisfies *condition* (T), which is defined as follows: X → Y implies t(X) < t(Y).

This condition is no longer purely "structural" but implies something about the "physical nature" of the cause-effect relation. Among other things, Schurz/Gebharter (2015, sec. 3.3) prove the following result:

(19) *Theorem 5 (Impossibility of screening off by future events):* (C)+(F)+(T) entail that empirical event-models (V,P,t) featuring the (in)dependencies DEP(X,Z) and INDEP(X,Z|Y) with t(Y) > t(X), t(Z) are impossible.

A still stronger extension of TC (not mentioned in Schurz/Gebharter 2015) is possible by adding the condition of *locality*, which asserts that in an event model no causal influence is propagated with a speed greater than the velocity of light c. To express this condition for-

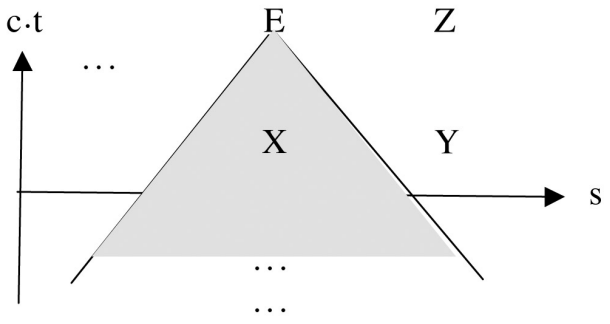

Fig. 7. Two-dimensional Minkowski diagram.
X but not Y or Z are in the past lightcone of E.

mally, we define a *spatial causal event model* as a quintuple (V,E,P,t,s) such that (V,E,P,t) is an event model and $s:V \rightarrow |R^3$ a position function assigning to each variable its position in (Euclidean) space ($|R^3$).

(20) *Definition of the condition of locality (L):* A spatial causal event model satisfies condition (L) *iff* $X \rightarrow Y$ implies $0 < |s(X)-s(Y)| / t(Y)-t(X) \leq c$.

(L) implies (T), since $|s(X)-s(Y)| / t(Y)-t(X)$ is negative if $t(Y) < t(X)$ and zero if $t(Y) = t(X)$. (L) entails that if an event X is not in the *past lightcone* of another event Y, X cannot causally influence Y; this is illustrated in fig. 7. Hence, from condition (L) the following counterpart of theorem 5 can be derived:

(21) *Theorem 6 (Impossibility of screening off by events outside the light cone):* (C)+(F)+(L) entail that empirical event-models (V,P,t) featuring the (in)dependencies DEP(X,Z), and INDEP(X,Z|Y) with $(|s(X)-s(Y)|/t(X)-t(Y)) \notin (0,c]$ and $(|s(Z)-s(Y)| / t(Z)-t(X)) \notin (0,c]$ are impossible.

A final strengthening of TC (that is also not treated in Schurz/Gebharter 2015) is possible by adding *intervention conditions*. An intervention on a variable X can be described as a value-instantiation of a so-called "intervention variable" I_X that causes X to take a specific value x. Usually (but not necessarily) it is assumed that intervention variables are under human control. We propose the following simple definition:

(22) *Definition*: I_X is an *intervention variable* for X in a CM (V,E,P) (with I_X,X,Y $\in$ V) *iff* the following *intervention condition (I_X)* is satisfied: (i) $I_X \rightarrow X$ is the only causal arrow connecting I_X with other variables in V, and (ii) X deterministically depends on all I_X-values i_x except on I_X's value "off".

This definition of an intervention variable comes close to the notion of "policy variables" in SGS (50).

All interventionist accounts (cf. Haussman 1998, Woodward 2003) take the following three features of interventions as the "key" to causality:

(23) *Three key features of interventions:*

(a) We can bring about the effect by producing its cause: DEP(Y,I_X) holds in $I_X \rightarrow X \rightarrow Y$.
(b) We cannot bring about the cause by producing its effect: INDEP(X,I_Y) holds in $X \rightarrow Y \leftarrow I_Y$.
(c) We cannot bring about the effect of a common cause by producing a correlated effect: INDEP(X,I_Y) holds in $X \leftarrow C \rightarrow Y \leftarrow I_Y$.

These three features have a straightforward explanation by TC's core axioms: (a) holds because of condition (22)(ii): I_X being on (i.e. $I_X \neq$ off) sets X to some value x, and because of $X \rightarrow Y$ and axiom (P), this must change Y's probability distribution. (b) and (c) hold because probabilistic dependencies are not propagated over common effects by axiom (C). Theorem 7 proves (by means of these three key effects) that the addition of intervention conditions adds empirical content to (C), even *without* the faithfulness assumption, under the mild additional assumption that an X-Y cycle is excluded:

(24) *Theorem 7:* Assume a CM (V,E,P) that satisfies (C) and contains two probabilistically dependent empirical variables X and Y together with intervention variables I_X and I_Y satisfying the intervention conditions (I_X)(i) and (I_Y)(i). Then: If E does not contain an X-Y cycle, then DEP(I_X,Y) and DEP(I_Y,X) cannot both hold.

Proof of theorem 7: From DEP(X,Y) together with (C), (I_X)(i), (I_Y)(i) and the exclusion of X-Y cycles it follows that the causal structure must be $I_X \rightarrow$ X---Y$\leftarrow I_Y$, with '---' for 'directed or common cause path'.

Assume DEP(I_X,Y). So by (C), some path $I_X \rightarrow$X---Y d-connects I_X and Y. Then X---Y can neither be X$\leftarrow\leftarrow$Y nor X$\leftarrow\leftarrow$C $\rightarrow\rightarrow$ Y. *For assume otherwise:* Then X would be a common effect of I_X and Y (or of I_X and C), and thus the path $I_X \rightarrow$ X---Y would be blocked. Therefore, X $\rightarrow\rightarrow$ Y must hold. From this it follows that the path X $\rightarrow\rightarrow$ Y$\leftarrow I_Y$ is blocked by Y. By the assumed acyclicity, this path is the only (type of) path connecting X with I_Y, from which it follows by condition (C) that INDEP(I_Y,X) must hold.

Thus, DEP(I_X,Y) implies INDEP(I_Y,X). In the same way it can be proved that DEP(I_Y,X) implies INDEP(I_X,Y). So DEP(I_X,Y) and DEP(I_Y,X) cannot both be true. Q.E.D.

Theorem 7 holds even without the intervention condition (22)(ii). If we drop this condition, we get what has been called a weak or "parametric" intervention variable (cf. Eberhardt and Scheines 2007, 986).

Intervention conditions are not general conditions, but assumptions about particular causal models. So they do not belong to the *general* theory of causality. However, intervention assumptions are usually justified by the following general assumption about the "relative" freedom of human actions:

(25) *Probabilistic freedom of interventions (Fr):* Most of the (possible) actions I of a person that manipulate the variables of a person-external causal system (V,E,P) are not probabilistically dependent on those variables in V that are not effects of I.

(C)+(Fr) make the truth of intervention conditions and therefore also the empirical consequences of theorem 7 highly probable, without making any ad-hoc assumptions. Thus the theory TC enriched with condition (Fr) achieves additional empirical content. The justification of intervention conditions by (Fr) makes clear why the practical meaning of interventions as controllable by human actions is so important: This is so because usually we know that our actions are free in regard to V. Of course, freedom in some non-relativistic sense is a matter of deep philosophical debate, but the notion of freedom expressed within (Fr) is a very weak one: It only requires that our actions are not the effects of any causes in V.

4. Conclusion: A Happy Marriage between Causality and Unification

We have argued that the two paradigms of scientific explanation, causality and unification, are not in opposition, but are mutually supporting. The arguments in section 2 and 3

have shown that causal explanations are preferable to non-causal explanations because the general theory of causality, TC, offers a higher-level unification of those regularities that are used as general premises in the explanations of single events. The unificatory gain provided by TC constitutes an important amendment to the account of unification developed in Schurz/Lambert (1994). In this account Schurz/Lambert (ibid., 74f) assume a "primitive" preference for causal as opposed to non-causal explanations, because they don't find a justification of this preference in terms of unification (the same holds is true for the accounts that are discussed in Schurz 1999, 2014 and Gijsbers 2007). This paper offers such a justification and, thus, supplies an important complement of the account of Schurz/Lambert (1994).

The demand of unification calls for an important clarification, namely: the demand applies only *ceteris paribus*. Out of two explanations of the same event E with *true* premises and comparable simplicity, the one with greater unification power is to be preferred. But of course, a true and less unificatory explanation is always preferable to a *false* although highly unificatory explanation.

Therefore, the cooperation between causality and unification does *not* exclude the possibility that the *true* causal explanation is not more but less unifying than a competing, highly unifying but *false* explanation. This helps to clarify a misunderstanding that may underlie an argument of Barnes against unification (1995, 265). He argued that it may well happen that three (kinds of) events E_i (i=1,2,3) are caused by three independent causes C_i (i = 1,2,3). Although the corresponding independent explanations do not produce unification, they are preferable to the attempt of explaining all three events in terms of one common cause C, because after all, they are the *true* explanations. What Barnes' example shows is that because not all events have a common cause, the request for "maximally unifying" (true) explanations cannot always be satisfied. Nevertheless, the theory of causality TC offers the following unifying explanation of this and all similar explanation situations at the higher level of explaining probabilistic regularities, as follows: *Either* (1.) the three (kinds of) events are probabilistically independent, in which case TC predicts and explains why they *cannot* have a common cause, or (2.) they are probabilistically dependent, in which case TC predicts and explains why they must be either related to each other in the form of a causal chain, or why they must be effects of a common cause. In case (2.) an explanation of the three events E_i by three distinct 'proximate' causes C_i (even if it is true) is clearly inferior, because it cannot explain the correlations between the C_i, which must obtain if the explanations are true. However, even in case (1.), TC achieves a unification surplus compared to a non-causal explanation. In conclusion, true causal explanations are more unified than true non-causal explanations, even if the causal structure of our world is not maximally unified.

Acknowledgements

This work has been supported by the CRC 991 of the DFG. For valuable help I am indebted to Markus Schrenk, Oliver Scholz, Victor Gijsbers, Theo Kuipers, Mark Siebel, Dennis Dieks, Jonah Schupbach, Ilkka Niiniluoto, Ioannis Votsis and an unknown referee.

REFERENCES

Armstrong, D. M. (1983): *What Is a Law of Nature?* Cambridge: Cambridge University Press.

Balzer, W., Moulines, C. U., and Sneed, J. D. (1987): *An Architectonic for Science*. Dordrecht: Reidel.

Barnes, E. (1995): "Inference to the Loveliest Explanation", *Synthese* 103, 251-277.

Beebee, H., Hitchcock, C., and Menzies, P. (Eds., 2009): *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Blalock, H. (1961): "Correlation and Causality: The Multivariate Case", *Social Forces 39*, 246-251.

Carnap, R. (1956): "The Methodological Character of Theoretical Concepts", in: H. Feigl and M. Scriven (Eds.), *The Foundations of Science* (pp. 38-76). Minneapolis: University of Minnesota Press.

Cartwright, N. (2007): *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

De Regt, H. (2006): "Wesley Salmon's Complementary Thesis: Causalism and Unificationism Reconciled?", *International Studies in the Philosophy of Science* 20(2=, 129-147.

Eberhardt, F., and Scheines, R. (2007): "Interventions and Causal Inference", *Philosophy of Science 74*, 981-995.

Fales, E. (1990): *Causation and Universals*. London: Routledge.

French, S. (2008): "The Structure of Theories", in: S. Psillos and M. Curd (Eds.), *The Routledge Companion to Philosophy of Science* (pp. 269-280). London: Routledge.

Friedman, M. (1974): "Explanation and Scientific Understanding", *Journal of Philosophy* 71, 5-19.

Frisch, M. (2007): "Causation, Counterfactuals and Entropy", in H. Price and R. Corry (eds.), *Russell's Republic,* Oxford University Press, Oxford, 351-395.

Gijsbers, V. (2007): "Why Unification is Neither Necessary nor Sufficient for Explanation", *Philosophy of Science* 74, 481-500.

Glymour, C. (2004): "Critical Notice", *British Journal for the Philosophy of Science 55*, 779-790.

Haussman, D. (1998): *Causal Asymmetries*, Cambridge University Press, Cambridge.

Healey, R. (2009): "Causation in Quantum Mechanics", in: Beebee et al. (Eds.), *The Oxford Handbook of Causation* (pp. 673-686). Oxford : Oxford University Press.

Hitchcock, C. (2010): "Probabilistic Causation", in E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), URL = <http://plato.stanford.edu/archives/win2011/entries/causation-probabilistic/>.

Kitcher, P. (1981): "Explanatory Unification", *Philosophy of Science* 48, 507-531.

Kitcher, P. (1989): "Explanatory Unification and the Causal Structure of the World", in: Kitcher, P., and Salmon, W. (eds.): *Scientific Explanation*, Univ. of Minnesota Press, Minneapolis, .410-505.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990): "Independence Properties of Directed Markov-Fields", *Networks 20*, 491-505.

Lewis, D. (1970): "How to Define Theoretical Terms", *Journal of Philosophy 67*, 427-446.

Mach, E. (1883/2009): *The Science of Mechanics*, BiblioBazaar, Charleston 2009 (German orig. 1883).

Norton, J.D. (2009): "Is There An Independent Principle of Causality In Physics", *British Journal for the Philosophy of Science 60*, 475-486.

Papineau, D. (1992): Can We Reduce Causal Direction to Probabilities? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Volume 1992, Volume Two: Symposia and Invited Papers*, 238-252.

Pearl, J. (1988, 1997): *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (2000, 2009): *Causality*. Cambridge: Cambridge University Press.

Psillos, S. (2009): "Regularity Theories", in Beebee et al. (Eds.), *The Oxford Handbook of Causation* (pp. 131-157). Oxford: Oxford University Press.

Railton, P. (1978): "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science* 45, 206-226.

Reichenbach, H. (1956): *The Direction of Time*. Berkeley: University of California Press.

Russell, B. (1912/13): "On the Notion of Cause", *Proc. Arist. Soc.*13, reprinted in: Russell, B., *On the Philosophy of Science*, Bobbs-Merrill Comp., Indianapolis, 1965, ch. 5.1.
Salmon, W. (1984): *Scientific Explanation and the Causal Structure of the World*, Princeton Univ. Press.
Salmon, W. (1989): *Four Decades of Scientific Explanation*, Univ. of Minnesota Press, Minneapolis.
Schurz, G. (1999): "Explanation as Unification", *Synthese* 120/1, 95 - 114.
Schurz, G. (2013): *Philosophy of Science: A Unified Approach*, Routledge, New York.
Schurz, G. (2014): "Unification and Explanation: Explanation as a Prototype Concept. A Reply to Weber and van Dyck, Gijsberg, and de Regt", *Theoria* 29(1), 57-70.
Schurz, G., and Lambert, K. (1994): "Outline of a Theory of Scientific Understanding", *Synthese* 101/1, 65-120.
Schurz, G., Gebharter, A. (2015): "Causality as a Theoretical Concept", to appear in *Synthese*.
Sneed, J. D. (1971): *The Logical Structure of Mathematical Physics*. Dordrecht: Reidel.
Sperber, D., Premack, D. und Premack, A. J. (eds., 1995): *Causal Cognition*. Oxford: Clarendon Press.
Spirtes, P., Glymour, C., and Scheines, R. (1993, 2000): *Causation, Prediction, and Search*. Cambridge: MIT Press.
Strevens, M. (2008): *Depth. An Account of Scientific Explanation*, Harvard Univ Press, Cambridge.
Suppes (1970): *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.
Tomasello, M. (1999): *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.
Whewell, W. (1837): *History of the Inductive Sciences*, John W. Parker, London.
Williamson. J. (2005): *Bayesian Nets and Causality*. Oxford: Oxford University Press.
Zhang, J., and Spirtes, P. (2008): "Detection of Unfaithfulness and Robust Causal Inference", *Minds and Machines 18*, 239-271.

Gerhard Schurz is full professor of philosophy at the University of Düsseldorf and director of the Düsseldorf Center for Logic and Philosophy of Science (DCLPS). He was associated professor at the University of Salzburg and visiting professor at the University of California at Irvine and at Yale University. His research areas cover philosophy of science, logic, epistemology and cognitive science, and meta-ethics. He is author of more than 200 publications in international Journals. Book publications among others: *The Is-Ought Problem* (Kluwer 1997), *Einführung in die Wissenschaftstheorie* (Wissenschaftliche Buchgesellschaft, 4th edition 2014), *Philosophy of Science: A Unified Approach* (Routledge 2013).

Address: Department of Philosophy, Heinrich Heine University Düsseldorf, Universitätsstraße 1, Geb. 24.52, 40225. Düsseldorf, Germany. E-mail: schurz@phil.uni-duesseldorf.de