teorema

Revista Internacional de Filosofía

Vol. XXXVIII/3 • 2019

KRK EDICIONES

teorema

Revista internacional de filosofía

teorema *Revista internacional de filosofía* es una publicación cuatrimestral que aparece en febrero, mayo y octubre. Anualmente edita el suplemento **limbo** *Boletín internacional de estudios sobre Santayana.* Aunque se tomarán en consideración artículos pertenecientes a cualquier disciplina filosófica, **teorema** presta una atención especial a aquéllos que, preferentemente en español e inglés, discutan temas de lógica, filosofía del lenguaje, filosofía de la lógica, filosofía de la mente, ciencia cognitiva, filosofía e historia de la ciencia, teoría del conocimiento, metafísica, y otras áreas relacionadas. Es propósito de **teorema** *d*edicar especial consideración al pensamiento español en todas sus épocas y facetas. **teorema** publica también, principalmente por invitación, notas críticas y recensiones; sin embargo, las propuestas de publicación en este ámbito son muy favorablemente acogidas. La revista **teorema** *s*igue el procedimiento de revisión externa y anónima por pares. Los contenidos de la revista están recogidos, entre otras, en las siguientes fuentes bibliográficas: Arts and Humanities Citation Index®, Current Contents©/Arts and Humanities, Carhus Plus+, Dialnet, Dice, Elsevier Bibliographic Databases (SCOPUS), Fuente académica, ISOC-CSIC, Jstor, Latindex, Periodicals Index Online, Répertoire bibliografique de la philosophie, RESH, Sumaris CBUC, The Philosopher's Index, y Ulrich's Periodicals Directory. **teorema** ha sido declarada "revista de excelencia" por FECYT (Fundación Española para la Ciencia y la Tecnología), organismo dependiente del Ministerio de Economía y Competitividad del Gobierno de España.

teorema *Revista internacional de filosofía* is a four-monthly journal (issues in February, May and October). **limbo** *Boletín internacional de estudios sobre Santayana is* included as an annual supplement. Although papers in any philosophical discipline will be considered, the main aim of the journal is to publish original articles either in Spanish or in English in Logic, Philosophy of Language, Philosophy of Mind, Cognitive Science, Philosophy and History of Science, Epistemology, Metaphysics and related areas. The study of Spanish Thought from any period or discipline will be given special consideration. Although critical notices and book reviews are usually invited, suggestions are welcome. **teorema** is a blind- and peer-reviewed journal. The contents of **teorema** *a*re indexed and collected in the following bibliographic sources: Arts and Humanities Citation Index®, Current Contents©/ Arts and Humanities, Carhus Plus+, Dialnet, Dice, Elsevier Bibliographic Databases (SCOPUS), Fuente académica, ISOC-CSIC, Jstor, Latindex, Periodicals Index Online, Répertoire bibliografique de la philosophie, RESH, Sumaris CBUC, The Philosopher's Index, and Ulrich's Periodicals Directory. **teorema** has been listed as a "journal of excellence" by the Spanish Government Agency FECYT (Foundation for Science and Technology).

Los contenidos de **teorema** desde 1971 están accesibles libremente en <www.dialnet.unirioja.es>.

For free access to back issues of **teorema** from 1971 up to last year go to <www.dialnet.unirioja.es>.

Redacción/Editorial Office: **teorema**. Universidad de Oviedo, Edificio de Servicios Múltiples, Campus de Humanidades, E-33071, Oviedo, Spain. **teorema**, apartado 702, E-33080, Oviedo, Spain.
Phone: (34) 98 5104378, fax: (34) 98 5104385, teorema@uniovi.es, www.unioviedo.es/Teorema, www.revistateorema.com
Suscripciones/Subscriptions: Ediciones Krk, Álvarez Lorenzana 27, E-33006 Oviedo, Spain;
phone & fax: (34) 98 5276501, correo@krkediciones.com, www.krkediciones.com. dl.:as-1736-2015

ÍNDICE/TABLE OF CONTENTS

SECCIÓN MONOGRÁFICA/SPECIAL SECTION

LA EXPLICACIÓN EN CIENCIA/EXPLANATION IN SCIENCE

GUEST EDITOR: VALERIANO IRANZO

NOTA CRÍTICA/CRITICAL NOTICE

OBITUARIO/OBITUARY

teorema
Vol. XXXVIII/3, 2019, pp. 5-23
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 5-23]

Introduction: Explanation in Science

Valeriano Iranzo

I. A LONG (AND WINDING) PHILOSOPHICAL DEBATE

There was a time when explaining was not considered a legitimate aim for science. Pierre Duhem and Ernst Mach, to name but two of the most representative authors, justified their scruples about explanation by invoking the autonomy of physics with respect to metaphysics and the economy of thought, respectively. The prevailing philosophical view on science at the turn of the nineteenth century was that science has to do primarily with "representing" (Duhem), "anticipating experiences" (Mach), … rather than to explaining. This may sound, indeed, a bit strange to us. After all, most scientists and philosophers of science nowadays admit that explanation is not only a legitimate aim for science, but also a valuable one. A Nobel Prize recipient in physics, Steven Weinberg, claimed that: "...the aim of physics at its most fundamental level is not just to describe the world but to explain why it is the way it is" [Weinberg (1994), p. 169]. Philosophers of science as different as Philip Kitcher and Bas van Fraassen, to mention just two examples, acknowledge that: "A crucial part of a scientist's practice consists in her commitment to ways of explaining the phenomena" [Kitcher (1993, p. 82]; "…the search for explanation is valued in science because it consists for the most part in the search for theories which are simpler, more unified, and more likely to be empirically adequate" [van Fraassen (1980), pp. 93-4]. However, it was not until almost the middle of the 20th century that explanation gained respectability thanks to Carl Gustav Hempel. His "covering law model", which can be found prefigured in other authors of the time (like Popper), became the background philosophical lore about explanation for several decades.

Explanation was understood by Hempel –in line with Logical Positivism's core assumptions– as a relationship between statements. Thus,

the statement that describes the event to be explained is the explanandum; the set of further statements required to explain it is the explanans. A fundamental constraint here is that, in addition to statements referring to initial conditions, the explanans must include at least one law, so that this particular sort of general statements –lawlike statements– are essential for doing the explanatory work. On the other hand, Hempel initially insisted that the inferential link between explanans and explanandum should be deductive –hence the so-called deductive-nomological (DN) model [Hempel and Oppenheim (1948); Hempel (1965b)]. Only logical and semantic properties of the statements are taken into account in the analysis of scientific explanation. Ontological concerns, those that could offend the empiricists' feelings of the time, were carefully avoided.

Nevertheless, and despite the subsequent modifications introduced by Hempel –allowing cases in which the explanandum is not deductively followed from the explanans–[1] his proposal soon came under devastating criticism. It can be said that in the late 1960s there was a widespread consensus that the Hempelian covering law model is untenable. Alternative approaches were developed. A standard classification distinguishes four subsets: probabilistic, unificationist, pragmatist, and causal-mechanical accounts. Until 1995 approximately this multiplicity of options coexisted, but from then on there was a noticeable change in that the causal approaches to the explanation in its different variants (interventionist, mechanistic, ...) were those clearly favoured by the academic community. Thus, even though few authors would claim that "asking for explanations" simply equates to "asking for causes", many of them would subscribe that any acceptable philosophical account of scientific explanation is forced to deal with *causal* explanations. That means that reflection on explanation involves also reflection on the notion of causal relation, if not also on the notion of cause itself –an item virtually absent in the Hempelian approach.[2] Wesley Salmon summarizes this change of mentality in the philosophical community as follows:

> There is a fundamental intuition –…– according to which causality is intimately involved in explanation. Those who are familiar with Hume's critique of causality may deny the validity of that intuition by constructing non-causal theories of scientific explanation. Others may skirt the issue by claiming that the concept of causality is clear enough already, and that further analysis is unnecessary. My own view is (i) that the intuition is valid –scientific explanations does involve causality in an extremely fundamental fashion– and (2) that causal concepts do stand in serious need of further analysis.[3]

This paragraph was firstly published in 1984, but Salmon's statement conveys a generalized attitude among philosophers of science at the early nineties. Here is a brief sketch of the story that led to Salmon's predicament.[4]

a) *Explanation as unification*

Michael Friedman and Philip Kitcher endorsed two different unificationist accounts of explanation. Friedman defended that explanation is tantamount to unification and the latter is understood as "reducing the total number of independent phenomena that we have to accept as ultimate or given" [Friedman (1974), p. 15)]. The law of ideal gases, for instance, is explained by the kinetic theory of gases insofar as a number of independently acceptable phenomena –unexplained phenomena, actually- are reduced to one. In the vein of the Hempelian account, the explanatory task is attached to laws, especially to the more comprehensive theoretical ones. Kitcher, in turn, underwrites that we "derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches how to reduce the number of types of fact that we accept as ultimate" [Kitcher (1989), p. 432]. Theories unify to the extent that they provide one pattern (or a few number of patterns) to derive the greatest number of sentences accepted by the scientific community. An *argument pattern* is an ordered triple composed by a schematic argument (a sequence of *schematic sentences*; i.e.: sentences in which some of the non-logical vocabulary has been replaced by dummy letters), *filling instructions* for completing the dummy letters in the schematic sentences, and *classifications* (they describe which sentences in schematic arguments are premises and conclusions). Here is an example:

QUESTION: Why do the members of G, G' share P?

ANSWER:

(1) G, G' are descended from a common ancestor G_0

(2) G_0 members had P.

(3) P is heritable.

(4) No factors intervened to modify P along the G_0-G, G_0-G' sequences.

Therefore, (5) Members of G and G' have P.

In this example there are five schematic sentences. Filling instructions require that *G*, *G'*, G_0 be replaced by names of groups of organisms, and that *P* be replaced by the name of a trait of organisms. Finally, the classification would state that (1)-(4) are the premises and that (5) is the conclusion deduced from them [Kitcher (1993), p. 83].

Generally speaking, the fewer the argument patterns employed and the larger the number of sentences derived, the better systematization we have. Particularly, the "explanatory store" over a corpus of statements *K* –all those currently accepted by the scientific community– is the best systematization of *K*, that is, the minimal set of explanatory patterns which allow the derivation of *K*.

The goal is unification, yes, but the explanatory import is attached to particular argument patterns –explanatory schemata–, and not to the most basic regularities found in nature (pace Friedman). We look, rather, for the minimal explanatory store for *K*. However, both authors agree on the idea that the explanatory relationship is a deductive relationship. Kitcher does not explicitly demand the necessity of laws for putative explanations, but he stills endorses the idea that explaining equates to giving an argument whose (deductive) conclusion is the explanandum.

How does Kitcher's unificationist approach tackle the problems previously raised against Hempel? The flagpole example is one of the most famous counterexamples against the Hempelian D-N model. A flagpole shadow is entailed by the height of the pole plus the angle of the sun above the horizon plus laws about the rectilinear propagation of light. Consequently, the flagpole shadow –the explanandum– is "D-N explained". But it is also true that we could change the argument so that the height of the pole is entailed by the flagpole shadow plus the remaining items. However, we would not say that the height of the pole is explained by its shadow (plus the other items). Unfortunately, the D-N model does not discriminate between cases where the explanatory relation is asymmetrical, even though the deductive constraint is fulfilled.[5]

Now, what is the answer provided by Kitcher to the flagpole counterexample? When confronted to those asymmetries, he resorts to our entrenched argument patterns. He argues that here we have two explanatory schemata: the "origin and development pattern" and the "shadow-pattern". The former appeals to the conditions under which the object originated and the subsequent changes it has suffered; the latter invokes the shadow of objects to derive their dimensions. The "origin and development pattern" should be favoured, according to him, because the "shadow pattern" does not allow us to derive the dimensions of those

objects which do not have shadows. Given that the number of sentences derived by the "shadow pattern" is less than those that can be derived from the alternative pattern, the latter is more unifying and should be preferred because of its higher explanatory value.

On account of this example, someone could think that the most explanatory patterns according to Kitcher are precisely those that fit with the causal order of the phenomena explained. But he insists that there is no objective causal structure in the world to ground the asymmetries of explanatory relationships: "one event is causally dependent on another just in case there is an explanation of the former that includes a description of the latter" [Kitcher (1989), p. 420]. Putting the matter in other words, our judgments/beliefs about causality just mirror our judgments/beliefs about explanatory relationships.

b) *Explanation as Statistical Relevance*

A further difficulty for Hempel's approach has to do with explanatorily irrelevant information. "Mr. Jones fails to get pregnant" –the alleged explanandum– is a deductive consequence from "All males who take birth control pills regularly fail to get pregnant" plus "Mr. Jones is a male" plus "Mr. Jones has been taking birth control pills regularly". Again, "Mr. Jones fails to get pregnant" is "explained" according to the D-N model, but we do not consider this is a putative explanation [Salmon (1971), p. 34]. Of course, taking birth control pills have no effect concerning pregnancy in males, so why should we consider it has any explanatory import for this particular explanandum?

The moral of the story is that only *relevant* information should be counted when explaining an event. Salmon's "Statistical Relevance" (S-R) model appeals to a probabilistic criterion. The idea is that a bit of information is explanatorily relevant if and only if it is statistically relevant, that is, if it affects the probability of what has to be explained. Since taking birth pills does not increase/decrease the probability of Mr. Jones getting pregnant, it has no epistemic import at all for it. Putting the matter in formal terms, if M=male, T=taking birth pills, and P=pregnancy, p (P|M & T) = p (P|M) = 0. However, being F=female, and taking for granted that the percentage of females who get pregnant after taking the pills is less than that of those females who do not take the pills, p (P|F & T) ≠ p (P|F). Therefore, T is explanatory relevant for F (regarding P), but completely irrelevant for M.

According to the S-R model, an explanation for a particular event is all the information statistically relevant to it, that is, the set of all factors that make any difference to the probability of the event. It's worth noticing here that both Hempel and the unificationists agreed on the idea that explaining an event is making it expected. Explanation demands a set of statements –laws, descriptions of initial conditions, explanatory schemata, …– that either entail or make highly probable the explanandum. But Salmon's S-R model departs from this assumption. Strictly speaking, to give an explanation equates to providing a probability distribution rather than providing an argument whose conclusion is the explanandum. Certainly, we must be careful to get the correct probability values and also not to overlook any statistically relevant factor involved. And this, and only this, is all we need to explain an event, regardless of its probability value. In fact, a highly improbable event may be explained by citing the relevant conditional probabilities. A consequence of this is that inconsistent explananda may be appropriately explained by the same corpus of information. If the aforementioned constraints are fulfilled, the explanation is fully satisfactory for both explananda. Here is an example:

> Two patients, x and y, are infected by streptococcus. Let V = recovery, T ($\neg$T) = 'treated (untreated) with penicillin', and R ($\neg$R) 'the strain is resistant (non-resistant)'. According to our medical statistics, p (V | T & $\neg$R) = 0.9; p (V | $\neg$T & $\neg$R) = 0.4; p (V | T & R) = 0.1; p (V | $\neg$T & R) = 0.1

Now, let's suppose that x has been infected by a resistant strain and y by a non-resistant one but, after receiving the treatment, both of them recover. The relevant information for explaining both events is the same, no matter that x's recovery is much more unlikely than y's recovery. Furthermore, the same information should be taken into account for explaining two inconsistent explananda (i.e.: x's recovery and x's non-recovery).[6]

This could be considered as a counterintuitive consequence of the S-R model. Notwithstanding, the main limitations for it have to do with the prospects to grasp causal links by means of statistical dependencies. Let's see what these are.

Science students are advised at introductory courses in scientific methodology not to confuse correlations with causes. If A is the cause and B is the effect, then presumably p (B | A) > p (B). Two events causally related are statistically dependent since the cause raises the probability

of the occurrence of the effect. But very often the way we proceed in science is, firstly, collecting data about a potential association/correlation between the variables (measuring frequencies, for instance), and secondly, inferring a causal relation from those data. But after detecting a statistical dependency between A and B four possibilities remain open: (i) A causes B; (ii) B causes A; (iii) A and B are effects of a common –and, perhaps, unknown– cause; (iv) A and B are associated by chance. Another famous example nicely shows how the S-R model can circumvent this difficulty. The reading of a barometer (B) and the occurrence of a storm (S) are highly correlated so that $p\,(B|S) > p\,(S)$ –and it is also the case that $p\,(S|B) > p\,(S)$. But we would hardly consider that B explains S (nor, alternatively, that S explains B) since both events are explained by a common cause, i.e.: the decrease of the atmospheric pressure (P).

The S-R model perfectly fits with our intuitions about this example. Since $p\,(S|P) = p\,(S|P\&B)$, B is statistically irrelevant to S given P. But P is statistically relevant to S given B, because $p\,(S|B) \neq p\,(S|P\&B)$. Analogously, $p\,(B|S) = p\,(B|P\&S)$ –so, S is irrelevant to B given P. And $p\,(B|S) \neq p\,(B|P\&S)$, so P is relevant to B given S. Shortly, P explains B and also S, but neither B explains S nor S explains B. In a situation like this it is said that B is *screened off* from S by P (and also that S is screened off from B by P).[7]

But winning a battle is not like winning the war. The point is that causal nets are not always statistically indistinguishable and different causal networks can accommodate the same class of probability distributions: "…, the resolving power of any possible method for inferring causal structure from statistical relationships is limited by statistical indistinguishability. If two causal structures can equally account for the same statistics, then no statistics can distinguish them" [Spirtes *et al.* (2000), p. 59].

Even imposing reasonable constraints on those probability distributions in principle intended to infer causal relationships –the Causal Markov Condition and the Minimality Requirement–, statistical indistinguishability cannot be avoided. Causal nets may be underdetermined by conditional probabilities. And we should not think that this is a problem just for very complex causal structures. An example of "strongly statistical indistinguishability" is:[8]

G_1 = A causes C; D causes B; B causes C; A causes D.

G_2 = A causes C; D causes B; B causes C; D causes A.

It's worth adding that this is a serious objection against the S-R model only insofar as it is taken for granted that explaining an event is closely related to locating it in *a causal network*, to making explicit its *causal history*, and so on. But the fact is that after several decades of debate "the great majority of philosophers is convinced that an account of explanation must provide a starring, if not exclusive role for causation" [Strevens (2014), p. 48)]. Woodward (2003), where it is defended a causal-interventionist interpretation of explanation, was probably a definitive turning point in this direction even though further versions within the causal framework –not necessarily interventionist– has been subsequently developed. The causal-mechanistic approach, which traces back to the eighties [Salmon (1984)], for instance, reemerged with strength –expurgated from its strongest physicalist commitments– at the turning of the century [Machamer, Darden & Craver (2000)]. Right now, it is surely the most discussed option in the literature.[9]

The list of papers included below reflects this state of the matter. Three of those four related to the analysis of explanation (those of J. Reiss, S. Psillos & S. Ioannidis, and S. Pérez-González) discuss problems internal to the causal tradition –two of them are particularly concerned with the explanatory import of mechanisms.

However, despite the widely predominance enjoyed nowadays by the causal tradition, that's not the full story. The consensus around the centrality of the notion of cause in order to explicate explanation does not entail assuming that there are no exceptions. Some authors have pointed at the limits of causal explanation through particular examples mainly –but not always– taken from physics [Lange (2016)]. The contribution of J. Suárez & R. Deulofeu, see below, goes along the non-causalist path but appeals to an episode of biology. Equilibrium explanations –a sort of ubiquitous explanation in biology and economics– are those favoured examples that, supposedly, cannot be reduced to the causalist-mechanical framework.[10]

This point raises some doubts about the prospects for giving an all-encompassing analysis for scientific explanation. Given the huge variety that can be found among different scientific fields, is it reasonable to look for "explanatory monism", so to say? It has been maintained that laws do not play a basic role in biology, for instance, in contrast to what happens in physics. Granted that, an account of explanation which exploits the explanatory import of laws is handicapped when dealing with biomedical sciences –conversely, physics would, in principle, be a more comfortable place for unificationist accounts. Analogously, mechanisms seem specially fitted for explanation in medicine, bio-chemistry, geology, … But, what

about social sciences? Even though talking about "social mechanisms" may be perfectly sound, it is debatable to what extent the sort of mechanistic explanation for fluctuations in the financial markets are similar to that invoked concerning the DNA replication in meiosis, for instance. Comparative detailed research, focused on specific scientific episodes, is required here. Even though addressing this issue is beyond the scope of this introduction, we will have a brief look at those views which highlight the contextualist constraints –not necessarily related to the peculiarities of the scientific fields– operating on explanation.

c) *The Contextual/Pragmatic Dimension of Explanation: Is That All?*

Given the difficulties to provide a general characterization of explanation, some authors have insisted that explanation is irremediably *contextual*. These approaches, labelled as "pragmatic" accounts" of explanation, highlight the relation between the explainer and her audience.[11] They are focused on questions as the assumptions required in the act of explaining to get some understanding for the audience, the role played by the agents' beliefs and interests concerning what counts as a correct explanation, the peculiarities of explanations related to idiosyncratic domains, …

Pragmatic approaches are intended to cast doubt on the philosophical task of giving a general or "structural" definition of explanation, like all those aforementioned. However, it is debatable to what extent the issues raised by pragmatic accounts cannot be accommodated in those standard approaches. The contextual relativity of explanation could be restricted, perhaps, to accepting that an amount of information related to the local context where the explanatory demand arises may be highly relevant. But this does not mean that contextual factors turn explanation into a purely psychological or subjectivist affair.[12]

Putting at the forefront the pragmatic dimension of explanation introduces a further topic deserving of attention. At the outset of this introduction we pointed out that philosophers and scientists nowadays agree that explanation is a matter of concern in scientific research. Theory-building, in particular, is driven –albeit, non-exclusively– by this concern. And, in principle, scientists prefer *theories* that unify different phenomena or domains, …, that have diverse empirical consequences (and some of them at least, about novel phenomena), that can be embedded in our background scientific knowledge, that are simple, … It could be said, then, that generally speaking scientists prefer good explanations to theories that score badly in those factors –commonly called

'theoretical virtues'. The debatable issue here, however, is whether these explanatory advantages have any confirmational import. When confronted with two alternative explanations for the same explanandum, should we consider that the best one qua explanation is also more confirmed than the other? Alternatively, if we confer more credibility to the best explanation of both, precisely because it shows better explanatory credentials, are we also favouring the most confirmed option of both?

Thinking that explanatory goodness increases the plausibility of the *explanans* is a key idea for partisans of "inference to the best explanation" (IBE, hereafter).[13] A standard way of introducing this inferential pattern is:

(P_1) F is some fact or collection of facts.

(P_2) Hypothesis H_1, if true, would explain F.

(P_3) No competing explanations (H_2, H_3, ..., H_n) would explain F better than H_1.

(Conclusion) One is justified in believing that H_1 is true.

The peculiarity of IBE is that the conclusion –the *explanans;* H_1 in this example– is inferred because of its explanatory yieldings about a particular explanandum. However, this is somewhat ambiguous. Thus, those who subscribe the importance of IBE do not entirely agree about its role. While some authors think it is primarily related to the context of discovery (IBE understood as a heuristical strategy), other authors insist that it has full epistemic import (see Iranzo (2007) for further discussion). There are still those overtly sceptics about IBE who do not consider that IBE refers to a specific inferential pattern whose reliability must be taken for granted. Bas van Fraassen, for instance, claims that the explanatory appeal of a hypothesis, however great, does not provide any *confirmational* advantage for the explanatory hypothesis. Rather, that feature is just an *informational* virtue –to use van Fraassen's words– that can be justified by pragmatic reasons alone [van Fraassen (1980), p. 87 and ff.].

It could be argued that differences between good and bad scientific explanations could hardly be qualified unless a consensus on what is explanation is reached. But the fact is that both debates –the nature of explanation and the significance and the epistemic value of IBE– have been developed separately for decades. Whatever it is, current discussion on this issue has evolved along two main paths.[14] Firstly, elaborating a precise characterization of the various virtues encompassed under the

generic label of 'explanatoriness'; secondly, forging a conceptual link between IBE and Bayesianism, which is the most well-established theory of confirmation at present.[15] It is expected then, that a careful scrutiny of those properties that qualify a hypothesis as a good explanation are somehow positively connected to its probability or degree of confirmation. Admittedly, the results obtained do not always play in favour, far from it, of the explanationists, that is, in favour of those who attach an epistemic (confirmational) import to "explanatoriness". W. Roche & E. Sober argue head-on against this view, while J. Schupbach defends IBE against a popular, potential criticism (see both papers below).

II. The Papers

Four of the six contributions included in this monographic section –those of Reiss, Psillos & Ioannidis, Pérez-González and Suárez & Deulofeu– are devoted to the analysis of explanation itself: what it is and how could we understand it, if possible, in terms of a more fundamental or pristine notion (causation, mechanism, ….). It should be added that Reiss and Psillos & Ioannidis address this question from a general perspective, while Pérez-González and Suárez & Deulofeu are focused on particular scientific disciplines (economics and biology, respectively). There are two more contributions, those of Roche & Sober and Schupbach, that are devoted to "inference to the best explanation" (IBE). The general concern here is whether the empirical assessment of hypotheses should be constrained by their respective explanatory merits. While Roche & Sober defend a skeptical argument against this possibility, Schupbach offers an interpretation of IBE that allows it to sidestep the so-called challenge of conjunctive explanations. Let's pause on all this.

According to explanatory causalism explaining an event has to do with ascertaining the causes that provoke it so that causality is the grounding notion for explanation. A basic associated insight is that scientific explanation is objective insofar as it reveals the framework of causal relationships actually operating in a particular context. In "Causal Explanation: Is All There Is to Causation?", Julian Reiss argues that absence causation is a challenge not only for physicalist and realist theories of causation but also for counterfactual and difference-making ones. He suggests an anti-objectivist account of causation —he explicitly acknowledges its Humean flavour— in order to cope with this problem: causes are in-

ferred from explanatorily successful stories. They are picked out by virtue of explanatory considerations since there is no objective causal structure in the world which legitimate causal explanations should reflect. His slogan is: "Explanation comes first; causation, second". Reiss defends that explanations are a kind of speech acts, i.e.: "transfers of understanding" between agents. Causal explanations, in particular, are those explanations which enable agents to make plausible causal inferences. But they are not considered "causal" to the extent that the explanans provides information about the causal history of the explanandum. Rather, what counts as causal explanation is established according to "social norms for causal inference" or "intersubjective facts about inferential practice". Among those norms Reiss mentions the evidential standards to trade-off between Type-I and Type-II errors in statistics or the injunction to discard alternative causal hypotheses before asserting a causal claim. Rules like these are, indeed, the effective constraints on causal explanation.

In "Mechanistic Causation: Difference-Making is Enough", Stathis Psillos and Stavros Ioannidis assume that causal explanation is crucial in scientific practice. Although they agree with Reiss on this point, they focus on an influential way to understand causal explanations, that is, on mechanistic accounts of it. In contrast to Reiss's approach, however, they think that causation "through mechanisms" comes first and explanation, second. According to them mechanisms are: (i) what turn a relation between A and B into a causal relation and (ii) what give causes their explanatory import. Shortly, mechanisms are necessary to causation and also to scientific explanation. They criticize, however, the prevailing account about mechanisms, according to which mechanisms essentially involve activities (in addition to entities, properties and relations). Psillos and Ioannidis think, rather, that "difference making is prior to production". Mechanisms are "networks of difference-making relations" –the latter usually understood in terms of counterfactual dependence– for them. Admittedly, activities are implemented to account for the productive dimension of mechanisms: a mechanism produces a result that can be properly considered as its effect. But Psillos and Ioannidis argue that establishing causality necessarily involves contrary-to-fact commitments.

Nevertheless, even taking for granted that understanding causality in terms of production cannot avoid difference making (since A cannot be the putative cause of B, unless A makes some difference to the occurrence of B), we could still think that that is not enough. In response to this Psillos and Ioannidis insist that mechanism is a concept effectively used in scientific practice. They resort to an episode in the history of

medicine –i.e.: the discovery of deficiency in vitamin C as the cause of scurvy– which is actually an example of absence causation (recall that this was the leading issue in Reiss's paper). They maintain that scientific practice demands reconstruction of "stable causal pathways", certainly, but identifying and detailing them equates to detecting the factors which make differences concerning the disease. The sort of evidence invoked here is not some sort of "mechanistic evidence" qualitatively distinct from evidence about difference-making relations. They conclude, then, that no metaphysical baggage related to activities, powers or capacities is required to understand the causal/explanatory role played by mechanisms in science.

Advocates of the mechanistic standpoint on explanation think that mechanisms play a substantial role in nearly all scientific domains. Besides, most of them think that an appropriate notion of mechanism should be suitable for all those domains. In "The Search for Generality in the Notion of Mechanism", Saúl Pérez-González discusses the prospects for such project. According to him, the development of an all-encompassing notion of mechanism is pursued through two different and alternative strategies. The "extrapolation strategy" tries to articulate a notion of mechanism taking one or a few fields of science as reference, and then applies that notion to the remaining fields. The "across-the-sciences" strategy consists of thinking about how mechanisms are understood across all the sciences and elaborates a notion of mechanism that includes just the shared features. After analysing paradigmatic examples of both strategies, Pérez-González argues that both face outstanding difficulties. The extrapolation strategy leads to notions unable to account for the varieties of mechanisms, while the across-the-sciences strategy leads to vacuous characterizations of mechanisms. He concludes that the search for generality does not look promising and suggests that it would be preferable to develop field-specific notions of mechanism.

A different approach is endorsed in "Equilibrium explanation as structural non-mechanistic explanations: The case of long-term bacterial persistence in human hosts". Javier Suárez and Roger Deulofeu depart from the widespread acceptance of the "New Mechanism" standpoint with the aim of questioning its universality. In contrast to the causal-mechanistic framework, they appeal to "structural explanations", that is, explanations that account for the phenomenon to be explained in virtue of the mathematical properties of the system where the phenomenon obtains, rather than in terms of the mechanisms that causally produce

the phenomenon. Structural explanations are very diverse in kind depending on the relevant structural properties invoked (bowtie structures, topological properties of the system, equilibrium constraints). Suárez and Deulofeu focus on a particular biological model, i.e., Blaser and Kirschner's nested equilibrium model of the stability of persistent long-term human-microbe associations. After investigating the role played by the mathematical properties of this model, they consider that it has fully explanatory import since: (i) it provides a set of differential equations —a mathematical structure— that satisfies an evolutionarily stable strategy (ESS); (ii) the explanation of host-microbe persistent associations is robust to any perturbation due to the nested nature of the ESSs; and more importantly for their case, (iii) this is so because the properties of the ESS directly mirror the properties of the biological system *in a non-causal way*. They conclude that this example vindicates the claim that equilibrium explanations look more similar to structural explanations than to causal-mechanistic ones.

Two further papers cope with the alleged link between explanatory value and inference.

In "Inference to the Best Explanation and the Screening-Off Challenge" Roche & Sober argue that "explanatoriness" is evidentially irrelevant. The "screening-off" thesis (SOT) affirms that the statement 'H would explain O if H and O were true' adds nothing at all to the empirical support that O by itself gives to H. The formal rendition of this is: p (H|O&EXPL) = p (H|O), where EXPL is the proposition that if H and O were true, then H would explain O. The main example for them is an extrapolation from a frequency estimate found in a sample to a particular member of the population. Thus, if *freq* (heavy smoking before age 50 | lung cancer after age 50) = α, and Joe —a random member of the population not included in the sample— got lung cancer after fifty, the probability that Joe was a heavy smoker before age 50 given that he got lung cancer after fifty —that is, p (H|O)— equates to α. Now, if we add EXPL —i.e.: the proposition that if H and O were true, H would explain O—, then p (H|O) = p (H|O&EXPL) = α. Consequently, EXPL is evidentially irrelevant to H.

Roche & Sober qualify the scope of SOT to examples in which the background information includes frequency data. However, they claim that there are realistic cases, similar to the aforementioned example, which fulfil this condition. Furthermore, they think that these cases go against IBE. They discuss two versions of IBE according to which inferring (=believing) H is licensed when H is the best potential explanation and also when H's

overall score regarding the explanatory virtues usually invoked in this debate (explanatory power, fertility, parsimony, ….) is high. Roche & Sober argue that even for these strengthened versions of IBE there are realistic counterexamples where all those explanatory considerations are screened-off by O. From this they conclude that there are corresponding versions of SOT –logically stronger than it, indeed– that undermine IBE.

Jonah Schupbach's paper ("Conjunctive Explanations and Inference to the Best Explanation") starts with an observation that is hardly disputable, i.e.: that sometimes there are different potential explanations for the same explanandum. This may occur both in everyday and scientific contexts. In case that accepting them all (or, at least, two of those explanations) provides us with a richer explanation, we have a "conjunctive explanation". At first sight, however, IBE urges us to infer the best option among *competing* explanatory hypotheses. But, if competition occurs just when hypotheses are incompatible (either because they are directly inconsistent by themselves or because the available evidence renders them incompatible), conjunctive explanations are straightforwardly excluded from the domain of applicability of IBE. Hence, a weaker notion of competition is required. His proposal here –jointly developed in a previous paper with D.H. Glass– is to define competition between hypotheses in terms of their (dis)confirmatory relations. He suggests a measure for the "net" degree of competition, based on the log-likelihood measure of confirmation, which contains two addends. One of them is related to the "direct competition" between H_1 and H_2 – the reciprocal disconfirmational effect without taking into account the evidence E –i.e.: the explanandum. The other addend alludes to the "indirect competition" since H_1 and H_2 could be competitors relative to some explanandum E even though they are entirely compatible (because, for instance, only one of the hypotheses is needed to explain E). Particularly, direct competition takes into account conditional probabilities between H_1 and H_2 –that is, $p(H_1|H_2)$, $p(\neg H_1|H_2)$ $p(H_1|\neg H_2)$ and $p(\neg H_1|\neg H_2)$–, while indirect competition considers the likelihoods of the conjoined hypothesis and its negations with respect to E –i.e.: $p(E|H_1\&H_2)$, $p(E|\neg H_1\&H_2)$, $p(E|H_1\&\neg H_2)$, $p(E|\neg H_1\&\neg H_2)$.

Nonetheless, even though Schupbach and Glass's probabilistic explication of competition plausibly widens the domain of applicability for IBE, there are problematic cases. Schupbach discusses an example where the conjunctive explanation is the best explanation but it includes competing hypotheses (on Schupbach and Glass's weak reading of competi-

tion). Thus, we should embrace the conjunctive explanation (H_1&H_2) since: (i) H_1 and H_2 together account for the evidence better than either does individually –that is, $p(E|H_1\&H_2) > p(E|\neg H_1\&H_2)$ and also $p(E|H_1\&H_2) > p(E|H_1\&\neg H_2)$, and (ii) the available evidence separately supports both hypotheses, *even though they disconfirm one another unconditionally and conditional on E*. According to this, the core prescription of IBE – "choose the best explanation among competing hypotheses"– is challenged. Schupbach's final considerations minimize the importance of competition as a necessary requirement to apply IBE. Accordingly, after pointing at the difference between "the *single* most explanatory hypothesis" and "the most explanatory conclusion", he recommends that IBE should be interpreted as inference to the most explanatory conclusion (regardless of that conclusion's logical form) as opposed to inference to the most explanatory single hypothesis.

Departament de Filosofia
Universitat de València
Avda. Blasco Ibáñez, 30
46010 València
E-mail: iranzov@uv.es

ACKNOWLEDGMENTS

Research leading to this monographic section has been supported by the Ministry of Economy and Competitiveness (MINECO), Spain, project FFI2016-76799-P

NOTES

[1] The "inductive-statistical" explanation (I-S model) [Hempel (1965b), pp. 381 and ff.].

[2] Hempel did not completely withdraw the notion of "causal explanation". See below, footnote 5.

[3] Salmon (1997), p. 323. See also, Cartwright (2004).

[4] For a detailed story, see Salmon (1989).

[5] Hempel distinguished between "laws of coexistence" and "laws of succession" [Hempel (1965b), p. 352]. The main difference between them is that the latter ineluctably refer to time order. Usually they describe changes in a physical, biological, …, system, through differential equations. Causal explanations are, according to Hempel, a subset of D-N explanations which include laws of succession. Then, his reply to the flagpole counterexample is that the

laws involved in it are laws of coexistence, so *they are not causal laws*. Therefore, even if we have two alternative D-N explanations when we interchange explanans and explanandum, the charge cannot be that the D-N model fails because it does not adequately discriminate the causal order of events. This reply, however, is hardly convincing (see the illuminating discussion in Psillos (2002a), sect. 8.5).

[6] Incidentally, $p\,(\neg V \mid T \,\&\, R) = 1 - p\,(V \mid T \,\&\, R) = 0.9$. Hence, if the same explanans is appropriate for those inconsistent explananda, then that very same explanans is appropriate for both an expected and an unexpected event.

[7] Common causes are good examples of screening-off relations, but they are not the only ones. See below the paper from W. Roche & E. Sober for a discussion in a different context.

[8] Spirtes *et al.* (2000), p. 60. Causal structures use to be represented by means of directed acyclic graphs. An introductory discussion of Bayesian nets can be found in Illari and Russo (2014), chap. 7. For more details, see Spirtes *et al.*

[9] Actually, some authors allude to the "New mechanistic" philosophy, which expands the scope of the notion of mechanism beyond philosophy of science. For a comprehensive view of the current debate on the notion of mechanism –and mechanistic explanation–, see Glennan and Illari (2018).

[10] See Reutlinger and Saatsi (2018) for a state of the art of non-causalist approaches to explanation. By the way, there are neo-Hempelian proposals still in play. An example is Diez (2014).

[11] Van Fraassen (1980), chap. 5, and Achinstein (1983) are the most refined proposals to date.

[12] See Woodward (2014) for this suggestion. The paper of Julian Reiss included below could also be seen as a compatibilist proposal between causalism and pragmatism.

[13] Presumably, the expression "inference to the best explanation" was coined by Gilbert Harman [Harman (1965)]. A historical antecedent related to IBE is Charles Peirce's term '*abduction*', a specific mode of reasoning irreducible to deduction and induction [see Campos (2011) and Psillos (2002b)].

[14] And there may be good reasons to remain so. In Cabrera (2018) it is argued that both issues should be kept separated.

[15] Some recent works on theoretical virtues are: Sober (2015), Keas (2018) and Schindler (2018). On the alleged connection between Bayesianism and IBE, see Lipton (2004) and Psillos (2007) for a positive and a negative assessment, respectively. Glymour (2015) is a critical perspective on probabilistic measures – not necessarily related to the Bayesian Criterion of Relevance to incremental confirmation, see Schupbach and Sprenger (2011)– for explanatory virtues.

REFERENCES

ACHINSTEIN, P. (1983), *The Nature of Explanation*; New York, Oxford University Press.

CABRERA, F. (2018), 'Does IBE Require a 'Model' of Explanation?'; *British Journal for the Philosophy of Science* <*https://doi.org/10.1093/bjps/axy010*>.
CAMPOS, D. (2011), 'On the Distinction between Peirce's Abduction and Lipton's Inference to the Best Explanation'; *Synthese*, vol. 180, pp. 419-42.
CARTWRIGHT, N. (2004), 'From Causation to Explanation and Back'; in B. Leiter (ed.), *The Future for Philosophy*, Oxford: Oxford University Press.
DÍEZ, J. (2014), 'Scientific W-Explanation as Specialised Embedding: a Neo-Hempelian Approach'; *Erkenntnis,* vol. 79, pp. 1413-43.
FRIEDMAN, M. (1974), 'Explanation and Scientific Understanding'; *Journal of Philosophy*, vol. 71, pp. 5–19.
GLENNAN, S. and ILLARI, P. M. (eds.) (2018), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*; Abingdon, Routledge.
GLYMOUR, C. (2015), 'Probability and the Explanatory Virtues'; *British Journal for the Philosophy of Science*, vol. 66, pp. 591-604.
HARMAN, G. (1965), 'The Inference to the Best Explanation'; *Philosophical Review*, vol. 74, pp. 88–95.
HEMPEL, C. (1965a), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*; New York, Free Press.
— (1965b), 'Aspects of Scientific Explanation'; in Hempel 1965a, pp. 331–496.
HEMPEL, C. and OPPENHEIM, P. (1948), 'Studies in the Logic of Explanation'; *Philosophy of Science*, 15, pp. 135–175. Reprinted in Hempel (1965a), pp. 245–290.
ILLARI, P. and RUSSO, F. (2014), *Causality – Philosophical Theory Meets Scientific Practice*; Oxford, Oxford University Press.
IRANZO, V. (2007), 'Abduction and Inference to the Best Explanation'; *Theoria*, vol. 22, pp. 339-46.
KEAS, M. N. (2018), 'Systematizing the Theoretical Virtues'; *Synthese* vol. 195, pp. 2761–93.
KITCHER, P. (1989), 'Explanatory Unification and the Causal Structure of the World'; in *Scientific Explanation*, P. Kitcher and W. Salmon (eds.), Minneapolis, University of Minnesota Press, pp. 410–505.
— (1993), *The Advancement of Science. Science Without Legend, Objectivity Without Illusions*; New York, Oxford University Press.
LANGE, M. (2016), *Because Without Cause. Non-Causal Explanations in Science and Mathematics*, New York, Oxford University Press.
LIPTON, P. (2004), *Inference to the Best Explanation* (2nd edition), London, Routledge.
MACHAMER, P., DARDEN L. and CRAVER, C. F. (2000), 'Thinking about Mechanisms'; *Philosophy of Science*, vol. 67, pp. 1-25.
PSILLOS, S. (2002a) *Causation and Explanation*; Chesham, Acumen.
— (2002b), 'Simply the Best: A Case for Abduction'; in A. C. Kakas and F. Sadri (eds.), *Computational Logic: Logic Programming and Beyond*, Berlin, Springer-Verlag, pp. 605–26.
— (2007), 'The Fine Structure of Inference to the Best Explanation'; *Philosophy and Phenomenological Research*, vol. 74, pp. 441-48.

REUTLINGER, A. and SAATSI, J. (eds.) (2018), *Explanation without Causation – Philosophical perspectives on Non-Causal Explanations*; Oxford, Oxford University Press.
SALMON, W. (1970), 'Statistical Explanation'; in R. G. Colodny (ed.), *The Nature and Function of Scientific Theories*, Pittsburgh, University of Pittsburgh Press, pp. 173-231.
— (1984), *Scientific Explanation and the Causal Structure of the World*; Princeton, Princeton University Press.
— (1989), *Four Decades of Scientific Explanation*. Minneapolis, University of Minnesota Press.
— (1998), *Causality and Explanation*; New York, Oxford University Press.
SCHINDLER, S. (2018), *Theoretical Virtues in Science: Uncovering Reality Through Theory*; Cambridge, Cambridge University Press.
SCHUPBACH, J. N. and SPRENGER, J. (2011), 'The Logic of Explanatory Power', *Philosophy of Science*; vol. 78, pp. 105–127.
SOBER, E. (2015), *Ockham's Razor: A User's Manual*; New York, Cambridge University Press.
SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000), *Causation, Prediction and Search*; Cambride, Mass., The MIT Press.
STREVENS, M. (2014), 'Probabilistic Causality'; in L. Sklar (ed.), *Physical Theory – Method and Interpretation*; New York, Oxford University Press, pp. 40-62.
VAN FRAASSEN, B. (1980), *The Scientific Image*; Oxford, Oxford University Press.
WEINBERG, S. (1994), *Dreams of a Final Theory*; New York, Vintage Books.
WOODWARD, J. (2003), *Making Things Happen: A Theory of Causal Explanation*; New York, Oxford University Press.
— (2017), 'Scientific Explanation'; *The Stanford Encyclopaedia of Philosophy*, E. N. Zalta (ed.), URL =<*https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/*>.

RESUMEN

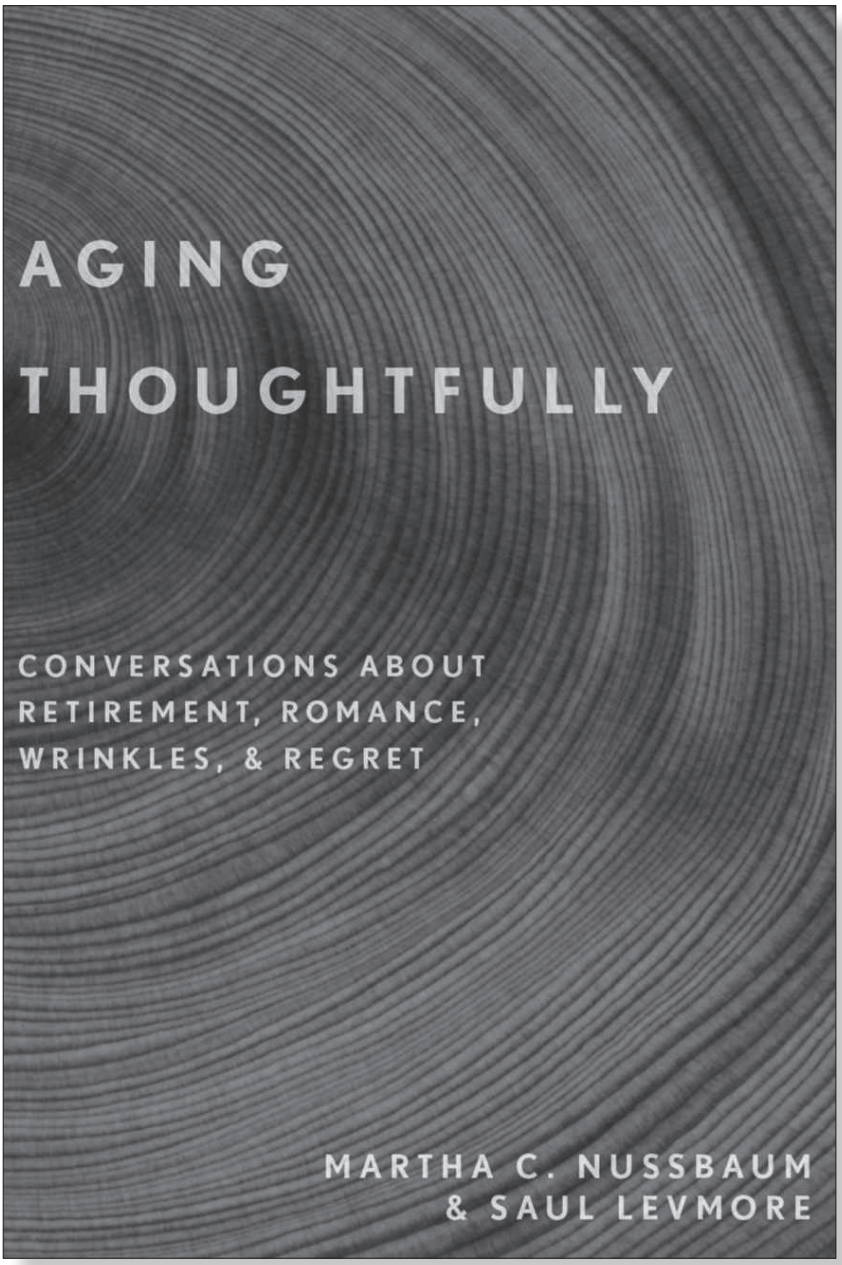
La presente introducción contiene dos partes. En la primera se ofrece una visión general de las principales posiciones defendidas en el debate filosófico sobre la explicación científica. En la segunda se resumen y comparan los seis artículos incluidos en la sección monográfica.

PALABRAS CLAVE: *explicación, explicación científica, inferencia hacia la mejor explicación.*

ABSTRACT

This introduction contains two parts. The first part offers an overview of the main positions developed in the philosophical debate about scientific explanation since Hempel's covering-law model. The second part summarizes and compares the six papers included in the monographic section.

KEYWORDS: *Explanation, Scientific Explanation, Inference to the Best Explanation.*


AGING
THOUGHTFULLY
CONVERSATIONS ABOUT
RETIREMENT, ROMANCE,
WRINKLES, & REGRET
MARTHA C. NUSSBAUM
& SAUL LEVMORE

teorema
Vol. XXXVIII/3, 2019, pp. 25-52
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 25-52]

Causal Explanation Is All There Is to Causation

Julian Reiss

RESUMEN

La ciencia trata los factores ausentes como si pudieran incluirse en relaciones causales. Los filósofos discrepan sobre problema de la "causación por ausencia" [*absence causation*]. Quienes entienden las causas como aquello que establece diferencias tienden a aceptar tal tipo de causación; quienes defienden perspectivas realistas o basadas en procesos tienden a rechazarla. En este artículo, defiendo que ninguno de los enfoques actualmente existentes tiene éxito. Ofrezco entonces una alternativa que entiende la explicación causal como conceptualmente prioritaria respecto a la causación y también un enfoque inferencialista de la explicación. Finalmente, muestro cómo mi propuesta sobre la causación se aplica a la causación por ausencia.

PALABRAS CLAVE: *causación, explicación, explicación causal, inferencialismo, causación por ausencia*.

ABSTRACT

Science treats absences as though they can stand in causal relationships. Philosophers disagree on the issue of absence causation. Proponents of difference-making accounts of causation tend to accept; proponents of process or realist accounts to reject it. I argue in this paper that no existing treatment is successful. I then offer an alternative that understands causal explanation as conceptually prior to causation and an inferentialist account of explanation. Finally, I show how my account of causation applies to causation by absences.

KEYWORDS: *Causation, Explanation, Causal Explanation, Inferentialism, Causation by Absences*.

I. INTRODUCTION

Can absences be causes? Scientific practice suggests they can. Here are examples from a variety of scientific disciplines:

> • *Physics*: Bombarding a diamond with high-energy particles such as photons can cause electrons to be ejected from the bond between the carbon atoms, forming a 'hole', which is positively charged. If

an electric field is applied to the crystal, the freed electrons will tend to drift in the direction of the applied field, producing a current. The holes or absent electrons will flow the opposite way, contributing to the current [Shockley (1950), pp. 9-10)].

• *Marine Geology*: Wherever the absence of oxygen causes anaerobic conditions, iron sulfide may form [Kuenen (1950), p. 218].

• *Biology*: As part of the mechanism of lactose regulation in *E.Coli*, the absence of lactose causes the Lac repressor to bind to the lac operator site and prevent the transcription of the lac operon [Griffiths et al. (1999)].

• *Nutritional science/physiology*: Prolonged starvation causes the body to fuel the brain with β- hydroxybutyrate instead of glucose [Cahill (2006)].

• *Psychology*: The absence of a noisy background makes trace discrimination so easy that genuine trace decay is masked by a ceiling effect [Baddeley and Scott (1971), p. 276].

• *Economics*: The absence in Islamic law of the concept of a corporation contributes to economic underdevelopment in the Middle East [Kuran (2004)].

• *Sociology*: Father absence negatively affects children's social-emotional development [McLanahan et al. (2013)].

• *International relations*: The absence of territorial threat causes a reduction in the likelihood of conflict in a dyad and is necessary for a dyadic democratic peace [Gibler and Tir (2010)].

• *Development studies*: 'The causal connection between democracy and the nonoccurrence of famines is not hard to seek [i.e., democracy causes the absence of famines/democracy prevents famines.]' [Sen (1999)].

• *World history*: In 17th century Asia Minor, the absence of strong government opposition together with the cooperation of local magnates, religious students, and corrupt officials, caused unemployed mercenary soldiers and provincial magnates to become leaders of semiautonomous regional power centres [Goldstone (2016), p. 385].

Examples like these can relatively easily be multiplied. I take it as my starting point that the sciences treat absences as if they can stand in

causal relationships.[1] An account of causation in the sciences should be able to make sense of this fact about scientific practice.

Philosophers of causation are divided on the issue of absence causation. Advocates of counterfactual or difference-making accounts of causation tend to accept it. Since causation consists in the whether or not a difference between a positive event and its absence makes a difference to an outcome, it does not matter whether the positive event is actual and the absence merely possible or vice versa. There is no structural difference between 'My drinking of cheap wine caused my hangover the next morning' and 'My abstaining from drinking of cheap wine prevented me from getting a hangover the next morning'. By contrast, proponents of physicalist and realist theories of causation tend to reject causation by absences. They argue that *ex nihilo nihil fit* (nothing comes of nothing). David Armstrong, for instance, writes: 'Omissions and so forth are not part of the real driving force in nature. Every causal situation develops as it does as a result of the presence of positive factors alone' [Armstrong (1999), p. 177].

I argue in this paper that no existing account of causation that offers a treatment of absence causation is successful. Difference-making approaches tend to multiply causation beyond the acceptable. In other words, they encounter what I call the 'problem of proliferation of causes'. Physicalist and realist approaches drive a wedge between positive causation and causation by absences that is solicited neither by ordinary language nor by scientific practice. I argue that the key to solving the problem of absence causation is to notice that it is explanatory considerations that enable us to judge which of a number of potentially relevant factors is a cause. Taking this idea as the starting point, I will argue that there is nothing beyond causal explanation in the concept of cause.

Proponents of causal explanation maintain that the explanans in a causal explanation provides information about the causal history of the event described in the explanandum [Lewis (1986)]. I argue that they have the conceptual order precisely upside down. Explanation comes first, causation second. There is no objective causal structure of the world, information about which is gathered and employed in causal explanations only at a later stage. In other words, the network model, according to which [Beebee (2004), p. 291]:

> [t]he complete causal history of the universe can be represented by a sort of vast and mind-bogglingly complex "neuron diagram" of the kind com-

> monly found in discussions of David Lewis, where the nodes represent events and the arrows between them represent causal relations…

is mistaken. Instead, I argue that causation and explanation are established jointly in a complex inquiry that does not neatly separate into a 'causal inference' and an 'explanation' stage.

II. Existing Work on Absence Causation

This section reviews and criticises existing stances on causation by absences. To make things easy, and because most of the discussion is framed in terms of singular or token-level rather generic or type-level causation, let me introduce a toy example that has, nevertheless, some scientific content. Meet Hamlin, the heedless hermit. Hamlin lives reclusively in a little hut in a faraway forest. Hamlin is not too fond of people and leaves his hut only to replenish the pantry. A bit on the paranoid side too, he locks the only door to his hut at night. It is a long trek to the village stockist and so one summer Hamlin digs out a two-storey cellar under his hut to allow him to survive longer periods without going out. One day the next winter, Hamlin intends to go out to buy goods to fill the last morsel of space in his cellar but he finds that he cannot locate the key to his door. 'I might as well', he thinks to himself, and spends the next 24 years living off his inventory until a group of scouts note a strange smell emanating from the hut and alarm the authorities. Taken to a hospital, he is given a full medical check-up. His state of health is determined to be surprisingly good under the circumstances but he is extremely pale and appears to suffer from a softening of his bones.

This case illustrates diverse kinds of causation by and of absences, including the hermit's heedlessness that causes a *key to be absent*, an *absent key* that causes Hamlin's complete seclusion, the *deprivation* of sunlight, which causes his vitamin-D *deficiency*, which in turn may cause all sorts of afflictions such as osteomalacia, osteoporosis, rickets, and depression [Gillie (2004)].

II.1 *David Lewis*

David Lewis defends a difference-making theory of causation according to which, roughly, C causes E if E counterfactually depends on C, i.e., if it is true that had C not been the case, E would not have been the case (either).[2] Lewis accepts that absences can be causes [e.g., Lewis (2004) [2000])].[3] But he immediately notes that doing so is not innocuous:

'One reason for an aversion to causation by absences is that if there is any of it at all, there is a lot of it — far more of it than we would normally want to mention. At this very moment, we are being kept alive by an absence of nerve gas in the air we are breathing' [*ibid.* p. 100]. Hamlin's lack of exposure to sunlight caused his vitamin-D deficiency. But under a counterfactual account of causation, so did the fact that earlier groups of Scouts did not find him or the village stockist's failure to carry vitamin-D supplements,[4] and a zillion other people's failure to do something that would have prevented Hamlin's vitamin-D deficiency. Let us call claims such as 'The village stockist's failure to carry vitamin-D supplements caused Hamlin's vitamin-D deficiency' 'irrelevant absence causation claims'. Irrelevant absence causation claims are intuitively false, and I will argue below that there are good reasons for maintaining that they are false indeed. I call the problem posed by theories of causation that deem irrelevant absence causal claims true the 'problem of proliferation of causes'.

Lewis's solution to the problem of proliferation of causes is (a) to bite the bullet and accept that irrelevant absences are in fact causes; but (b) to argue that there are Gricean pragmatic reasons for not mentioning them in a conversation [*ibid.* p. 101]: 'There are ever so many reasons why it might be inappropriate to say something true. It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned, and so on [Grice 1975]'. Thus, while it is true, according to this account, that the village stockist's failure to provide vitamin-D supplements caused Hamlin's deficiency, we don't normally mention this because it would be inappropriate to do so, as it would be to mention to one's partner, 'You look fat!' even though, indeed, they look fat. In the case of the grocer's neglect an argument could be made that mentioning it in a conversation violates Grice's maxim of relation as it is, while true, irrelevant in the context at hand.

The problem with Lewis's suggestion is that we don't just *fail to assert* irrelevant absence causation claims, we positively *deny* them [Beebee (2004), McGrath (2005)]. I certainly wouldn't causally attribute Hamlin's state to the grocer's neglect, and there is some empirical evidence that indicates that 'ordinary folk' (i.e., students at elite universities) are largely in agreement about analogous cases [Livengood and Machery (2007)]. What makes matters worse is that pointing out to irrelevant absence causation deniers that under a counterfactual account of causation irrelevant absence causation claims are true does not appear to make them change

their judgement. They instead take this as a reason to doubt the counterfactual theory [McGrath (2005)].

Another reason for thinking that irrelevant absence causation claims are not merely inappropriate to make but false is that they do not have the usual connotations of causal claims. Causal claims normally support claims about predictions. But I will not, when notified of the village stockist's continued 'negligence' (and not much else), predict that other individuals in his trading area will develop vitamin-D deficiency. Causal claims normally support claims about interventions. But I will not ever propose a policy that mandates grocers to supply vitamin-D to hermits. Causal claims normally support claims about the attribution of blame and praise. But I will not travel to the village, enter the shop and reprimand the owner for his negligence. And if I get asked why Hamlin came down with vitamin-D deficiency, I will be met with incredulity if I answer 'The village stockist didn't give him food supplements. *That* absence does not explain.

II.2 *Contrastive Causation*

Jonathan Schaffer works largely in the Lewis tradition but maintains that causation is contrastive, that is, the prototypical form of a causal claim is 'C rather than C* caused E rather than E*', where C* and E* are alternative events [Schaffer (2004b), (2005). Schaffer, like Lewis, accepts causation by absences [see in particular Schaffer (2004a)]. He gives four reasons in favour of doing so [Schaffer (2005), pp. 300-1]:

(1) Absence causation is intuitive: intuition accepts some absences as causal.

(2) Absences play the predictive and explanatory roles of causes and effects.

(3) Absences play the moral and legal roles of causes and effects.

(4) Absences mediate causation by disconnection.

I have already given examples that illustrate (1) and (3). All scientific examples given at the beginning of this paper are examples for (2). Schaffer gives a gory example for (4): decapitation causes death by preventing oxygenated blood from preventing brain starvation. Thus, the absence of blood mediates decapitation and death.

Schaffer, too, notes the problem of proliferation of causes. And he gives exactly Lewis's response [*ibid.* p. 302]:

> The one aspect of the paradox of absences that the contrastive strategy does not directly resolve is… the problem of counterintuitive causal claims. That is, contrastivity allows that the queen's reigning on her throne

> rather than watering my flowers causes my flowers to wilt rather than blossom. But perhaps this remaining implausibility can be explained away pragmatically. Perhaps the reason it sounds wrong to say that the queen's not watering my flowers causes them to wilt is that we never supposed that the queen would deign to water my flowers. Contrastivity helps explain why this affects the acceptability of the absence claim. We resist taking such an unrealistic supposition as a contrast. The queen's watering my flowers is not easily swallowed as a relevant alternative. At c* sits an irrelevance. The contrasts trigger the pragmatics.

But how can we explain the making of false assertions on the basis of pragmatics in this case? We often make false claims that can be justified pragmatically. 'No, you don't look fat!' is a case in point. Apart from being hurtful, the truth may be too complex or irrelevant. A truth may not speak to the intended audience while the uttered falsehood does. None of these reasons apply with respect to irrelevant absence causation claims. It's certainly not hurtful to say that the grocer's failure to supply supplements caused the hermit's vitamin-D deficiency or that the Queen of the United Kingdom sitting on her throne caused Schaffer's flowers to wilt. It's not complex, at least not any more than the intuitively true causal claims about Hamlin's forgetfulness and lifestyle. As their name suggests, irrelevant absence causation claims are irrelevant, but the response to making one is not, 'That is irrelevant', but rather: 'That is false'. Pointing to the irrelevance of the contrast events therefore does not solve the problem.

Does the claim speak to the audience? I maintain that causal claims are not established, asserted, or defended for their own sake (Reiss 2015). Scientist don't pursue causal inquiries in order to add to our knowledge of the causal structure of the world. First and foremost, causal claims are useful claims. Correlatively, acceptability of a causal claim stands and falls with its usefulness. Causal claims are useful because they support predictions and explanations, interventions and the attribution of blame and praise. Not all causal claims are good at all these functions. The sentence 'Gravity causes stars to collapse' is not helpful to attribute blame or praise. Many causal relations are fragile and subject to interferences. Therefore, the corresponding claims are often not useful for predictions. If a causal claim mentions a factor on which we cannot intervene, we cannot exploit the relation to bring about a desired effect. An irrelevant absence causation claim is not useful for any of these purposes.

So here is a possible defence of the Lewis/Schaffer approach based on pragmatics. An irrelevant absence causation claim is true, but denied because ordinary folk and, in particular, scientists (as well as legal theorists, historians and so on) expect causal claims to be useful and, since it is not, it does not speak to them. When amongst each other, metaphysicians in the Lewis tradition make free use of such claims.

Of course, this won't work. When a teacher is explaining to a student that humans descended from apes, she is strictly speaking uttering a falsehood. But this falsehood might speak better to the student than the truer claim that human beings and the other great apes descended from a common hominid ancestor who was not, strictly speaking, an ape [this example is due to Elgin (2007)]. But the teacher would normally know that the simple claim is false and use it deliberately in order to enhance understanding or retaining. When we deny that the grocer caused the hermit's vitamin-D deficiency, we do not have such objectives in mind. We're convinced of the falsehood of the irrelevant absence causation claim ourselves.

I conclude that the Lewisian two-stage picture of (1) there is a plethora of true claims of causation by absence, given by the appropriate relations of counterfactual dependence; and (2) only some of these are assertible, pragmatics determines which, is mistaken.

II.3 *Physical Connection*

One of David Hume's criteria for causation was that a cause and effect must be contiguous. That is, there must not be spatio-temporal gaps between the cause and the onset of the effect. There are various theories of causation building on this idea [e.g., Aronson (1971), Ehring (1998), Fair (1979), Russell (1948), Salmon (1984), (1994)]. These accounts maintain, essentially, that for C to cause E C and E must be connected by a causal process of the right kind. The main difference between different physical connection accounts lies in their understanding of the notion of a 'causal process'.

Phil Dowe has addressed absence causation explicitly, and developed an account of absence causation within the framework of a causal process theory [Dowe (2004), (2007)]. According to Dowe (2007), p. 167:

> C causes E iff
>
> 1. there is a set of causal processes and interactions… between C and E, and
>
> 2. $ch_{C\varrho}(E) > ch_{\_C\varrho}(E)$, where ϱ is an actual causal process linking C with E,

where [*ibid.* p. 90]:

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

Absences are not physically connected to the events we sometimes speak of as their effects. Whatever Hamlin did when he forgot where he put his key did not issue in a causal process that interacted with lock on the door responsible for his captivity. Dowe consequently rejects causation by absences. What he offers instead is a novel concept, called 'quasi-causation' [Dowe (2004)] or causation* [Dowe (2007)], to characterise these kinds of cases. Dowe calls causation by absence 'omission'[5] and defines it as follows [Dowe (2007), p. 136]:

> Omission: not-A caused* B if
>
> (O1) B occurred and A did not, and there occurred an *x* such that
>
> (O2) *x* caused B, and
>
> (O3) if A had occurred then B would not have occurred, and there would have been a causal relation between A and the process due to *x*, such that either
>
> (i) A is a causal interaction involving the causal process *x*, or
>
> (ii) A causes *y*, a causal interaction involving the causal process *x*,
>
> where A and B name positive events, and *x* and *y* are variables ranging over facts or events.

Cases of causation by absence are thus termed cases of causation*. Lack of sunlight caused* Hamlin's vitamin-D deficiency. Vitamin D that is absorbed from food or supplements or synthesised in the skin after exposure to sunlight is converted by the liver into calcifediol. Calcifediol is then converted in the kidneys into calcitriol, the active form of vitamin D in the body and a secosteroid hormone. Calcitriol increases the uptake of calcium from the gut into the blood. When the blood serum level of calcium is low, calcium will leave the bones and if the vitamin-D deficiency is prolonged, this process leads to rickets and osteoporosis. Supposing that Hamlin did develop osteoporosis (B), the just mentioned process (*x*) caused it, and (presumably) it is true that if he had been ex-

posed to sunlight (A), then osteoporosis would not have occurred. A would have interacted with *x*.

Dowe's account does not, however, solve the problem of proliferation. Anyone's providing the hermit with vitamin D would interrupt the decalcification process and thus 'The village stockist's failure to provide vitamin-D supplements to Hamlin caused* his deficiency' (or any other irrelevant absence causation claim) is true. But irrelevant absence causation claims are false.

There is another problem with Dowe's account. In ordinary English there is no distinction between 'cause' and 'cause*' or 'quasi-cause'. This does not immediately imply that there is no corresponding difference in nature. Ordinary language glosses over many important differences, and it evolves in response to changes in culture, the environment, and our knowledge of the world. Among Francis Bacon's 'Idols of the Mind' were the 'Idols of the Market Place', which concerned exactly the potential lack of correspondence between ordinary language concepts and the structure of the world [Urbach and Gibson (1994) Book I, Aphorism 43]:

> There are also Idols formed by the intercourse and association of men with each other, which I call Idols of the Market Place, on account of the commerce and consort of men there. For it is by discourse that men associate, and words are imposed according to the apprehension of the vulgar. And therefore, the ill and unfit choice of words wonderfully obstructs the understanding. Nor do the definitions or explanations wherewith in some things learned men are wont to guard and defend themselves, by any means set the matter right. But words plainly force and overrule the understanding, and throw all into confusion, and lead men away into numberless empty controversies and idle fancies.

If there is a significant lack of correspondence between language and world, scientific investigation can reveal this and introduce more precise and accurate concepts. Thus, modern physics distinguishes instantaneous from average velocity [Kuhn (1981)/(1963)], modern biology between biospecies, ecospecies, and phylospecies, modern psychology between working memory, short term memory, iconic memory, and long term memory (for the second and third example, see Taylor and Vickers (2017). There is no analogue with respect to absence causation. It is well understood that decapitation causes death by preventing oxygenised blood from flowing to the brain. No new concepts have been introduced in science to describe this fact.

And this is odd since there are thousands of concepts to describe acts of causing in ordinary and scientific language: smoking *kills*, increases in the money stock *inflate* the price level, Suzy *shoved* Billy, the storm *delayed* the plane, enzymes *phosphorylate* proteins. These are all causal relations, and the specific causative verb used provides more information about the kind of causal relation than would 'cause'. 'Kill' provides information about the effect (death); 'inflate' about the direction (bigger); 'shove' about action (push) and the manner (forcefully); 'delay' about the timing (later); 'phosphorylate' about the mechanism (phosphorylation). There is no causative verb that expresses 'causation by absence' that would be more accurate to use than 'cause' or whichever causative verb that is uses and that does not distinguish between positive and negative causation.[6]

Absence causation does not raise a scientific puzzle that scientists could solve by splitting the concept into two or more. Absence causation is a well-known phenomenon that does not seem to require that kind of conceptual manifestation. If it did, scientists would have long introduced novel terminology that works better for their purposes. Absence causation poses at best a metaphysical problem. But it does so only if one presupposes that causation must be a relation or for some other reason must originate in an event or some other metaphysical entity. Starting instead, as I do, with the view that philosophy should be continuous with scientific practice, certain metaphysical principles shouldn't override well established knowledge and custom.[7]

II.4 *Causation vs explanation*

Helen Beebee agrees with Phil Dowe and many others who argue that effects must emanate from something real [e.g., Anjum and Mumford (2018), Armstrong (1999), Moore (2009), Mumford and Anjum (2011)] that there is no causation by absence [Beebee (2004), p. 291]. Unlike Dowe, however, Beebee recognises the problem of proliferation of causes. She therefore proposes to amend the definition of causation by absence with a clause stating that only those absences count as causes that deviate from the normal course of affairs [*ibid.* p. 296]:

> (I) The absence of an A-type event caused b if and only if b counterfactually depends on the absence: Had an A-type event occurred, b
>
> (i) would not have occurred; and

> (ii) the absence of an A-type event is either abnormal or violates some moral, legal, epistemic, or other norm.

Hamlin's losing the key to his hut comes out as a cause of his vitamin-D deficiency because clause (ii) is satisfied: it is abnormal to misplace the key to one's house for 24 years, especially if that means that one cannot get out. At the same time, the grocer's failure to provide food supplements is not a cause as his behaviour is not abnormal.

Beebee then goes on to argue that this definition is fine as far as the ordinary concept of causation is concerned, but it is unsatisfactory as an account of the metaphysics of causation. Human-made norms should not be thought to affect what there is by way of causal facts.

Her account of the metaphysics of causation builds on a distinction between causation and causal explanation. In what she thinks of as ordinary cases of causation, causal explanation and causation go together. Why did the match light? Because it was struck. The striking of the match caused it to light. But in cases of causation by absence, no causal relation corresponds to the explanatory claim. We may answer the question, 'Why did Hamlin have vitamin-D deficiency?' by saying, 'Because of the lack of sunlight', but lack of sunlight did not cause the deficiency.

How can we make sense of the idea that causal explanations do not always describe causal relations, i.e., that it is not always the case that the explanans of a causal explanation describes a cause and the explanandum an effect? Beebee invokes David Lewis' account of causal explanation, according to which, 'to explain an event is to provide some information about its causal history' [Lewis (1986), p. 217] in support. In her view [*ibid.* p. 302]:

> One can give information about an event's causal history in all sorts of other ways—by saying, for instance, that certain events or kinds of event do not figure in its causal history, or by saying that an event of such-and-such kind occurred, rather than that some particular event occurred.

According to the Lewisian account, 'JFK died because someone shot him' is a causal explanation in that it provides some information about JFK's death, but it does not describe a causal relation as 'someone shot JFK' is not an event — it is at best a disjunction of particular events. Similarly, citing that something that would have caused one outcome did not happen explains the occurrence of the alternative outcome because we learn that a particular event was not in the effect's causal history and we learn about the causal structure of a nearby world in which Hamlin was exposed to sunlight. Common sense is mistaken when it judges that some absence

caused an outcome. But that is understandable as causation and causal explanation are very similar and do overlap to a considerable extent.

There are various issues with Beebee's account. Let me focus on the main problem here: an appeal to Lewis' theory of causal explanation invites some classical counterexamples to older theories of scientific explanation.

Causes provide information about the occurrence of their effects; but effects also provide information about the occurrence of their causes. Take a standard counterexample to the deductive-nomological model of explanation [Hempel and Oppenheim (1948)]: We can infer the height of the flagpole from the length of the shadow (provided we have information about the position of the sun), but we'd be hard pressed to accept the length of the shadow as explaining the height of the flagpole. Now, as we have seen, Lewis explicitly allows causal explanations to be existential in character (e.g., 'There exists an individual who shot JFK' explains that JFK died). But as the length of the shadow *provides information about* height of the flagpole, the length of the shadow also provides information about the existence of *causes* of the height of the flagpole: *viz.*, that the causes of the height of the flagpole must have been exactly such that it could cast the shadow we have observed. Similarly in common-cause structures: the drop of the barometer reading provides information about the causes of the storm — but the barometer reading does not explain the storm [Hartsock (2010)].

So, we can't be quite as permissive as Lewis and, by extension, Beebee. Unless Beebee (or anyone else) succeeds in providing an account of causal explanation that allows non-causes to explain outcomes causally without running into counterexamples, we will have to come to the conclusion that her attempt to distinguish between positive and negative causation by declaring the latter to be non-causation but causal explanation fails because causal explanations need to cite causes.

III. What Is a Scientific Explanation?

I agree with Beebee that (some) absences causally explain outcomes. I also agree that (some) events causally explain outcomes. What I deny is that this explanatory equivalence between negative and positive causation, as well as the linguistic equivalence discussed in Section II.3 translate into a significant causal difference.

When no existing account can handle certain kinds of causal claim that are important to the sciences it is time to look for something new. I

do so in this section and the next, leaving my own treatment of absence causation to Section 5. To motivate my account, note that what's wrong with Dowe's and Beebee's accounts of absence causation is that they attempt to dichotomise causal statements into statements of causation proper and statements of second-class causation, be it quasi- or causation* or causal explanation. I do not deny that there are important differences among causal relations. The following:

(a) The father burped his child.

(b) The father caused his child to burp.

(c) The father made his child burp.

(d) The father got his child to burp.

(e) The father let his child burp.

are all expressions of causal relations (or of causings if one does not believe that causation is a relation) but they all provide different information about what precisely happened. (a) expresses a direct involvement; (b) is indirect; (c) expresses intentionality on the father's and some degree of resistance on the child's part; (d) expresses successful encouragement; and (e) permission. There is a difference between (a) burping and (e) letting burp but no more of a difference than there is between (b) causing to burp and (c) making burp.

What all these sentences have in common is that they explain the outcome. I suggest that this is all they have in common. Traditional accounts have the order of conceptual priority wrong. They maintain that causal concepts represent aspects of an objective causal structure of the world and that scientific explanations are successful to the extent that they cite information about this objective causal structure of the world. I maintain instead that scientific inquiry aims to establish explanations of phenomena of interest. A good explanation is one that serves its purpose (see below for an account of the purposes of explanations). Causal claims are articulations of science's inventory of explanatory knowledge.

Following Douglas Walton [e.g., Walton (2004)] I maintain that an explanation is a transfer of understanding from an explainer to an explainee, following a request. Explanations are thus certain kind of speech act [see also Achinstein (1983), Achinstein (2010), Donato Rodriguez and Zamora Bonilla (2012), Faye (2007)]. The explainee (who may be a single person or a group such as a scientific community) initiates a dialogue by asking a why-question. Such a request is based on an assump-

tion of a partially shared understanding or starting point [Walton (2004), p. 83]. For example, if we ask a physician why Hamlin was vitamin-D deficient, we share the starting point that only naturalistic explanations are admissible. Absent such starting points there is little chance that the dialogue will be successful. Starting points may include [see van Eemeren and Grootendorst (1992): Ch. 14): particular facts ('Hamlin lost his key'), suppositions ('Hamlin would have continued to go out occasionally and would not have covered up completely had he not locked himself in'), generalisations ('Individuals who live at latitudes not too close to the polar regions, who follow a healthy diet and do not cover up fully whenever they are outside do not normally develop vitamin-D deficiency'), values ('it's a good thing to live healthily') and norms ('people normally leave their house at least occasionally').

By asking a why-question, the explainee indicates a gap in understanding it requests to be filled in by the explainer. A gap in understanding is often an inconsistency or incoherence between existing commitments.[8] If an explainee holds all of the commitments mentioned in the previous paragraph, she will expect the Hamlin to be healthy. But he has vitamin-D deficiency and osteomalacia or osteoporosis. She asks why he has these afflictions because her commitments entailed that Hamlin would be fine. More generally, the explainee is justified in asking, 'Why P?' if (a) both explainee and explainer are committed to P; and (b) some of the explainee's other commitments (most of which are shared with the explainer) entitle the explainee to expect not-P [*cf.* Donato Rodriguez and Zamora Bonilla (2012), p. 36). The explanation is successful if and only if the contradiction or incoherence is resolved.

Once the contradiction or incoherence is resolved, the explainee has an improved ability to make new inferences. The following are some of the purposes a successful explanation can serve [Keil 2006]:

- to predict a similar event in the future (starving a person of sunlight will make her vitamin-D deficient);
- to diagnose the reason for failure in order to fix the system (providing large amounts of vitamin-D will help if vitamin-D deficiency is the reason for osteomalacia but not, or not alone, if it is due to kidney failure);

- to attribute praise or blame even when the outcome is singular (Hamlin, or Hamlin's forgetfulness, can be blamed for his poor health condition);
- to justify or rationalise an action (if Hamlin were to take action against forgetfulness this would be justified and rationalised by pointing to the harm he caused himself);
- to serve aesthetic pleasure (this does not apply in the hermit case; but: 'One can explain a work of art, a mystery of cosmology, or the intricacies of a poem with the sole goal of increasing appreciation in another, providing that person with a better polished lens through which to view the explanandum'; *ibid.* p. 234).

Understanding is simply the ability to make inferences of this kind [Newman (2012), (2013), (2017)]; there is a large literature in cognitive and development psychology on understanding and inference-making ability, for example: Cain et al. (2001), Oakhill (1984). Inferences include both formal (ones that are valid in virtue of their form such as *modus ponens*) as well as material inferences (ones that are 'valid' in virtue of the content of the concepts involved such as causal and other inductive inferences). In science and everyday life, the latter type is dominant. To understand better means to be able to make more useful inferences. To transfer understanding from explainer to explainee by addressing a gap in his understanding is to improve the explainee's ability to make useful inferences.

Causal explanations are transfers of understanding that not only make P expectable to the explainee, they also improve his capacity for causal inference. Causal inferences follow specific norms. For example, in order for an explainer to be entitled to assert the lack of sunlight as a cause of Hamlin's vitamin-D deficiency, he would need to be in the possession of some evidence that it is actually true that he was not exposed to sunlight while locked into his hut, and that, at least casually, rules out alternative explanations of the vitamin-D deficiency (such as malnutrition, obesity or short bowel syndrome). The norms characterising causal inference are context-dependent and therefore in part dependent on the situation in which the causal explanation is sought. There are many variables that affect these norms (for a more detailed treatment: see Reiss 2015). To cite just one: stakes. To give someone a life sentence requires higher evidential standards than blaming the neighbour for killing one's plants by starving them of sunlight.

Causal explanations can thus quite easily be distinguished from other kinds of explanations. Here is an example, due to Mark Lange, of a why-question that seeks a *mathematical* explanation [Lange (2016), p. 7; quoted from Khalifa et al. (2018)]:

> Consider the fact that at every moment that Earth exists, on the equator (or on any other great circle) there exist two points having the same temperature that are located antipodally (i.e., exactly opposite each other in that the line between them passes through the Earth's center). Why is that?

To establish an explanation of this phenomenon, we do not engage in causal reasoning. Instead we construct a mathematical proof, in this case one based on the intermediate value theorem. Deriving a theorem follows norms different from those of causal inference.

The counterexamples that plague other views sometimes referred to as 'explanation-as-inference' do not affect the account presented here because of the nature of causal reasoning. In order for a description of an event or factor to come out as a causal explanation of some phenomenon, its assertibility has to be established by the norms for causal inference. These norms include the precept to rule out alternative (causal and non-causal) explanations of the phenomenon of interest, and 'there is reverse causation from putative effect to putative cause' and 'there is a common cause' is on any list of standard alternative explanations for an association. In the stock examples of the shadow and the flagpole, and the barometer and the storm, we cannot rule out reverse causation and a common cause, respectively. Thus, if a speaker offers the length of the shadow as an explanation of the height of the flagpole or a drop in the barometer reading as an explanation of the storm, he would make utterances to which he is not entitled. My interest here is primarily in causation and causal explanation, which is why I offered a solution to the counterexamples to 'explanation-as-inference' accounts in terms of causal explanation. Khalifa et al. (2018) have shown that the asymmetry problem can also be solved within an inferentialist account of explanation without appeal to *causal* asymmetry.

IV. A Neo-Humean Account of Causation

David Hume is usually credited with the regularity account of causation, according to which C causes E if and only if C and E regularly co-occur, E temporally follows C, and C and E are spatio-temporally

contiguous [e.g., Psillos (2002)]. According to this account, (a) causation is a relation in the world; and (b) this relationship is one of regular association. There is nothing beyond regular association 'in the objects'.

It has been argued that Hume also maintained an alternative account (or that his writings can be interpreted as defending such an account) according to which causal claims are expressions of our habits of inference. Observing C, we infer that E will happen, and that inference is projected onto the world. It is that inference that is the source of the idea of a necessary connection. This account has therefore also been called the 'necessitarian' or 'projectivist' account [Beauchamp and Rosenberg (1981), Beebee (2007)].

According to this view, then, causation is a property of the mind, a kind of reasoning. Causal claims do not refer to any objective relations (or other things) in the world. My own account is very similar to Hume's in this respect — albeit different in its understanding of the reasoning involved. Causal claims are inter-subjective in that their assertibility depends on beliefs, values, and norms of reasoning that are shared among the members of a community and thus not entirely subjective or arbitrary.

To help build my account, let me invoke Peter Achinstein's notion of an *epistemic situation*. According to Achinstein, an epistemic situation 'is an abstract type of situation in which, among other things, one knows or believes that certain propositions are true, one is not in a position to know or believe that others are, and one knows (or does not know) how to reason from the former to the hypothesis of interest, even if such a situation does not in fact obtain for any person' [Achinstein (2001), p. 20]. For an agent to be in an epistemic situation ES is to share certain beliefs, values, and norms of the kind referred to above as 'starting points'. Among the norms particularly noteworthy are norms of causal reasoning, which, among other things include the injunction to rule out alternative causal hypotheses before asserting a causal claim, evidential standards that allow the agent to trade off type-I and type-II errors and so on.

> **Causation**. For any two distinct agents in an epistemic situation ES, a causal claim that relates cause C and effect E is assertible if and only if one agent's citing C in ES successfully causally explains E to the other.[9]

Let me add two qualifications to this definition. First, I am not fully committed to a definition of causation in terms of causal explanation. In other work I have defended an account that invokes inferential relationships directly, without going through causal explanation [Reiss (2015)]. A

disadvantage of invoking explanation is that doing so might open a Pandora's box of issues related to explanation such as whether all explanations are contrastive, what to make of the difference between explanations-how, explanations-that and explanations-how possibly, how to deal with the goal- and/or context-relativity of explanations and so on. The development of answers to these potential problems will have to wait for another paper. However, going through causal explanation allows me to offer necessary and sufficient assertibility conditions which the inferentialist account prevents. The inferential networks that are associated with causal claims are far too varied to allow the formulation of such conditions. The account proposed here shifts that variability to the notion of causal explanation. Causal claims have very little in common, but, I suggest, they all have in common that they causally explain. On this point I am in full agreement with Michael Scriven who argued a very long time ago that [Scriven 1966), p. 256)]:

> When we are looking for causes, we are looking for explanations in terms of a few factors or a single factor; and what counts as an explanation is whatever fills in the gap in the inquirer's or reader's understanding.

My account can be understood as an elaboration of this idea of Scriven's.

The second qualification is that I only formulate assertibility conditions, not truth conditions. The assertibility conditions laid out above are implausible as truth conditions. A scientist living in the first half of the 18th century will have been entitled to assert causal claims involving phlogiston in the explanation of combustion. But we don't want to say that such claims are true. My hunch is to define the truth conditions in terms of an ideal epistemic situation in which all knowable facts are actually known, and all agents agree on values and norms of reasoning. The full development of this idea too will have to wait for another occasion.

The assertibility condition for a sentence such as 'The father caused his child to burp' in some epistemic situation ES is that if in ES an explainee asks 'Why did the child burp?' by stating 'The father pressed her lightly on the belly', the explainer would resolve a tension in the explainee's reasoning and improve his inferential abilities.

The account offered here is similar to Hume's but very differently motivated. Hume did not think we could have knowledge of or speak meaningfully about causation in the objects or 'objective causation' because of his associationism. With no sense impression to be associated

with the word 'cause', there was no place for objective causation in our image of the world.

This motivation has lost much of its pull. Today it is at least controversial to claim that causal relations are never directly observable (for positions against this claim, see for instance Anscombe (1971), Beebee (2009), Cartwright (2000), Ducasse (1926)[1993]]; Beebee cites some evidence about the observability of causation from psychology). And I don't think there's anyone left who thinks that we can't meaningfully talk about something we can't see (the death blow to this idea may have been Quine (1953) but I won't argue).

My own motivation for developing an inferentialist account of causation derives from the inability of representationalist accounts — accounts maintaining that 'cause' refers to some objective feature of world — to come to grips with the way in which causal language is used in science, legal, historical and clinical practice, and in everyday life (for some arguments to that effect and a review of the literature, see Reiss (2015), Ch. 1]. There simply doesn't seem to be any single property all causal relations 'in the objects' share, and disjunctive theories (which define causation as a disjunction of properties) don't seem to fare much better. It is therefore that I believe we should try something new.

Apart from solving the problem of proliferation of causes (see next section), the account I favour has a number of other desirable properties. One is that it can provide a situation-specific account of the difference between causes and conditions. We would not normally cite the presence of oxygen in the air as a cause of the forest fire. This is a problem for difference-making accounts of causation such as Lewis' because the presence of oxygen in the air certainly makes a difference to whether or not the fire occurs. But in most epistemic situations citing oxygen explains nothing and thus, on the view of causation presented here, it does not cause the fire. By contrast, if there is an epistemic situation where, say, the absence of oxygen is a condition for the proper working of some production process, a leak in a pipe and ensuing presence of oxygen does explain and therefore does cause the fire (to cite an example due to Mackie (1980).

We can define a causal condition as follows:

> **Causal Condition**. It is assertible that C is a causal condition for E if and only if there exists an (actual) epistemic situation in which an agent's citing C successfully causally explains E to another agent.

It is thus easy to see why some speakers might confuse causes and causal conditions.

More generally, because of the way it is constructed, it is quite impossible for any counterexample to affect the account. For any causal claim, if the claim is assertible, there will be norms determining that this is so. As the explanatory account of causation makes use of just these norms, it will not count a genuine case of causation as non-causation and vice versa. Of course, it may be the case that any given speaker is unaware of certain norms or misapplies them, that there is disagreement about what are the correct norms or how to apply a norm, that a norm does not completely determine correct usage, that competing norms provide different answers to a causal question and that norms evolve over time. Two things follow from this. First, it is possible for a speaker to make false causal claims. There are inter-subjective facts about inferential practice a speaker can ignore or misapply. 'The village stockist caused Hamlin's vitamin-D deficiency' is false in world that shares our inferential norms. Second, the boundaries of the concept of cause are blurry. I don't think, for instance, that the norms describing ordinary language use are able to decide whether in cases of symmetric overdetermination (in which two factors C1 and C2 are able to bring about an effect E and both come to completion) each factor should be called a cause. This is different in legal practice where when two persons are equally causally involved in a third person's death, the actions of either will be regarded as a separate cause of the death, even if the death would have occurred without the action of either (but not without the action of both). The lesson here is: at any point in time there will be indeterminate cases but they will be resolved over time or, when a resolution is required immediately, we can (and will) plump for one.

V. Absence Causation on the Explanatory Account

On the shared understanding that (a) 'Hamlin lost his key'; (b) 'Hamlin would have continued to go out occasionally and would not have covered up completely had he not locked himself in'; (c) 'Individuals who live at latitudes not too close to the polar regions, who follow a healthy diet and do not cover up fully whenever they are outside do not normally develop vitamin-D deficiency'; (e) 'it's a good thing to live healthily'; and (f) 'people normally go out occasionally', 'lack of sunlight causes vitamin-D deficiency' the explainee could expect Hamlin not to be vitamin-D deficient on the basis of the shared understanding. He is

therefore justified in asking why Hamlin did get sick. Citing the causal claim resolves that tension. It also allows further inferences, for instance about the attribution of blame or the justification of an utterance of disapproval. Thus, we can blame Hamlin's condition on his forgetfulness, as the latter caused him to be locked in, and being locked in caused him not to be exposed to sunlight.

The explanation 'Hamlin developed vitamin-D deficiency because of lack of sunlight' is a causal explanation in part because in order to be entitled to making the explanation, the explainer must be in the possession of evidence that no other risk factors such as malnutrition or obesity explains the deficiency. The inferences the explanation permits are also typical of causal inferences. In other words, the inferences that permit the explanation and that are licensed by the explanation are causal inferences.

Contrarily, the village stockist's failure to provide vitamin-D supplements does not explain the outcome. There is no shared understanding for instance of his having made a promise to provide the vitamin or there being a general norm to that effect. Suppose instead that we lived in a world in which everyone covered up completely and so in order to receive sufficient amounts of vitamin-D they buy supplements. If in that world the village stockist failed to supply the vitamin to Hamlin, his failure and not Hamlin's forgetfulness or the lack of sunlight would explain, and therefore cause, the outcome.

Importantly, the account presented here does not drive a conceptual wedge between positive and negative causation. All causal claims are true in virtue of the explanations in which they are used. There is no 'real' connectedness in some cases and no or 'pseudo' connectedness in others. This does not mean that there are no differences. Via explanations, different causal claims are related to different kinds of inferences. To use an example introduced above, 'The father *made* the child burp' entails intention on the father's part and resistance on the child's whereas 'The father *let* the child burp' entails permission, i.e., the removal of (or refusal to introduce) an obstacle. Similarly, we can make different inferences when we hear that someone lets a pet die by neglect than when we hear that someone killed his pet by direct involvement. But there is no dichotomy such that all cases of positive causation fall on one side of some border and all cases of negative causation on the other.

VI. CONCLUSIONS

Let me concluding by way of offering some responses to possible objections. One objection might be that, against what was argued in Section 2.4, there are causal explanations that don't cite causes after all. Might Lewis be correct in saying that 'JFK died because someone shot him' is a causal explanation but 'Someone shooting him caused JFK to die' is a false causal claim? The account of causal explanation offered here agrees with Lewis that the former claim is a causal explanation. The account of causation described in Section 4 entails that the associated causal claim is true (both judgements presuppose that there are situations in which the claim 'Someone shot JFK' is offered as an explanation of JFK's death, but this is of course not hard to imagine). Is this a counter-example to the proposed account?

No. It is mere metaphysical prejudice that leads to refusing 'Someone shot JFK' to figure as a cause in causal claims. Lewis and his followers accept only events as causes. Natural and scientific language is a lot more flexible than that. Causes can be events, states, factors, variables, substances, processes, agents and probably a host of other things I cannot think of at present. My account does not place any restrictions on what kinds of entities can figure in causal claims as any restriction would lead to cases that look and waddle and quack like causation but would not come out as cases of causation on the account. As far as I can see, there is no problem in accepting 'gunshot' as a cause of death, and forensic and medical practice agrees.

A more serious objection is that reasoning and inference are not something *in* the world but rather *about* the world. In Jonathan Bennett's words, reasoning cannot play 'the role of a puller and shover and twister and bender' [Bennett (1988), p. 22]. My answer to this worry is to ask what difference it would make if, for each and every true causal claim, there was 'a thing' (an event, a property, a state of affairs…) in the world that would make the claim true? It obviously wouldn't make a difference to our inferential practices. Scientific, legal, clinical and historical practice as well as everyday discourse would proceed in the exact same manner. 'But these practices must be grounded in something — in the causal structure of the world!', the objection might continue. To which I'd respond: yes, inferential practices are grounded in something. But this something is not the causal structure of the world. It is inferential success. As we have seen above, there are a variety of more ultimate pur-

poses for offering causal explanations. To the extent that existing practices are successful at achieving these purposes, they are justified. If specific norms fail to advance our purposes, they will be changed over time. 'But how do you explain their success?', the objection goes on.

Well, that is asking one question too many.

Institute of Philosophy and Scientific Method
Johannes Kepler University
Altenberger Str. 50, 4040 Linz, Austria
E-mail: julian.reiss@jku.at

ACKNOWLEDGMENTS

I wish to thank the CHESS research group and Wiebke Szymczak for valuable comments. Errors and omissions remain, of course, my responsibility. This research was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 667526 K4U). The content reflects only the authors' views, and the ERC is not responsible for any use that may be made of the information it contains. I also acknowledge funding from the Spanish Ministry of Science and Innovation for the research project 'Laws, explanation, and realism in physical and biomedical sciences' (FFI2016-76799-P).

NOTES

[1] Most of the examples mentioned above involve causation *by absences*. There is also causation *of absences* (or prevention: see the example from development studies) and causation by absences of absences (see the example from biology). Since it makes no difference to anything I am going to say in this paper, the focus will be primarily on causation by absences.

[2] For simplicity I only provide a sufficient condition. The necessary condition is harder to formulate because of redundant causation: if events C', C" etc. compete with C to bring about E, that is, presuming C and E are actual events and C caused E, if E would have obtained in the absence of C because of any of the other events, then E is not counterfactually dependent on C. I do not think that the problem of redundant causation is solvable within a counterfactual framework. To avoid having to deal with the complexities redundant causation bring with it, I omit the necessary condition here.

[3] Strictly speaking, it is propositions that enter relations of counterfactual dependence, not events. The proposition 'Hamlin developed vitamin-D deficiency' is counterfactually dependent on the proposition 'Hamlin lacked exposure to sunlight for 24 years'. But that doesn't mean that an absence literally does the causing [Lewis (2004) [2000], p. 100]: 'So I have to say that when an

absence is a cause or an effect, there is strictly speaking nothing at all that is a cause or effect. Sometimes causation is not a relation, because a relation needs relata and sometimes the causal relata go missing'.

[4] I'm supposing here, not unreasonably I hope, that village stockists don't have a legal or moral duty to carry vitamin-D supplements.

[5] Dowe distinguishes cases of omission, which have the absence on the side of the cause from cases of preventions, which have the absence on the effect side, and from cases of prevention by omission, which have absences on both sides. Since the philosophical worries are exactly the same between all three kinds of case, throughout the paper I focus on omissions.

[6] One might argue that ordinary (and legal) language sometimes does draw important distinctions between positive and negative causation. The difference between killing and letting die is of course very important, in legal practice and elsewhere. Distinguishing killing from letting die won't solve the problem, however, since 'letting die' is still a (periphrastic) causative verb expressing causal sufficiency [Lauer (2010)]. That is to say, 'letting die' is causing, not quasi-causing.

[7] My own account does not in fact violate ancient metaphysical principles such as *ex nihilo nihil fit*. As long as one does not presuppose that 'cause' always represents some real entity, activity, power or relation, absence causation does not pose any metaphysical conundrum either.

[8] I say 'contradiction or incoherence' because the tension between existing commitments and P is often not as strong as a contradiction in the logical sense.

[9] I use the locution 'causal claim that relates cause C and effect E' rather than 'C causes E' in order to allow for causative verbs other than 'cause' to figure in causal claims.

REFERENCES

ACHINSTEIN, P. (1983). *The Nature of Explanation*; Oxford, Oxford University Press.
— (2001), *The Book of Evidence*; Oxford, Oxford University Press.
— (2010), *Evidence, Explanation, and Realism. Essays in the Philosophy of Science*; New York, Oxford University Press.
ANJUM, R. L. and S. MUMFORD (2018), *Causation in Science and the Methods of Scientific Discovery*; Oxford, Oxford University Press.
ANSCOMBE, E. (1971), *Causality and Determination: An Inaugural Lecture*; Cambridge, Cambridge University Press.
ARMSTRONG, D. (1999), "The Open Door"; in *Causation and Laws of Nature*, Howard Sankey (ed.); Dordrecht, Kluwer, pp. 175-185.
ARONSON, J. (1971), "On the Grammar of 'Cause'"; *Synthese* 22, pp. 414-30.
BADDELEY, A. D. and D. SCOTT 1971, "Short Term Forgetting in the Absence of Proactive Interfering"; *Quarterly Journal of Experimental Psychology* 23(3): 275-283.
BEAUCHAMP, T. L. and A. ROSENBERG (1981), *Hume and the Problem of Causation*; Oxford, Oxford University Press.

BEEBEE, H. (2004). "Causation and Nothingness", in *Causation and Counterfactuals*. J. Collins, N. Hall and L. Paul (eds.); Cambridge (MA), MIT Press**:** pp. 291-309.
— (2007), "Hume on Causation: The Projectivist Interpretation"; in *Causation, Physics and the Constitution of Reality*, H. Price and R. Corry (eds.), Oxford, Oxford University Press, pp. 224-249.
— (2009), "Causation and Observation"; in *The Oxford Handbook of Causation*. H. Beebee, C. Hitchcock and P. Menzies (eds.), Oxford, Oxford University Press, pp. 471-497.
BENNETT, J. (1988), *Events and Their Names*; Indianapolis (IN), Hackett Publishers.
CAHILL, JR., G. (2006), "Fuel Metabolism in Starvation"; *Annual Review of Nutrition* 26, pp. 1-22.
CAIN, K., J. OAKHILL, M. BARNES and P. BRYANT (2001), "Comprehension Skill, Inference-Making Ability, and Their Relation to Knowledge"; *Memory & Cognition* 29(6), pp. 850-859.
CARTWRIGHT, N. (2000), "An Empiricist Defence of Singular Causes"; in *Logic, Cause and Action: Essays in Honour of Elisabeth Anscombe*, Roger Teichmann. Cambridge, Cambridge University Press, pp. 47-58.
DONATO RODRIGUEZ, X. and J. ZAMORA BONILLA (2012), "Explanation and Modelization in a Comprehensive Inferentialist Account"; in *Epsa Philosophy of Science: Amsterdam 2009*, H. de Regt, S. Hartmann and S. Okasha (eds.), Dordrecht, Springer, pp. 33-42.
DOWE, P. (2004) "Causes Are Physically Connected to Their Effects: Why Preventers and Omissions Are Not Causes"; *Contemporary Debates in Philosophy of Science*. Christopher Hitchcock. Oxford, Blackwell**:** 187-196.
— 2007. *Physical Causation*; Oxford, Oxford University Press.
DUCASSE, C. J. (1926) [1993], "On the Nature and the Observability of the Causal Relation"; in *Causation*, E. Sosa and M. Tooley (eds.), Oxford, Oxford University Press, pp. 125-136.
EHRING, D. (1998), *Causation and Persistence*. Oxford, Oxford University Press.
ELGIN, C. Z. (2007), "Understanding and the Facts"; *Philosophical Studies* 132, pp. 33-42.
FAIR, D. (1979), "Causation and the Flow of Energy"; *Erkenntnis* 14(3), pp. 219-250.
FAYE, J. (2007), "The Pragmatic-Rhetorical Theory of Explanation"; in *Rethinking Explanation*, J. Persson and P. Ylikoski (eds.), Dordrecht, Springer, pp. 43-68.
GIBLER, D. and J. TIR (2010), "Settled Borders and Regime Type: Democratic Transitions as Consequences of Peaceful Territorial Transfers"; *American Journal of Political Science* 54(4), pp. 951-968.
GILLIE, O. (2004), "Sunlight Robbery: Health Benefits Are Denied by Current Public Health Policy in the Uk"; London, Health Research Forum.
GOLDSTONE, J. (2016.), *Revolution and Rebellion in the Early Modern World: Population Change and State Breakdown in England, France, Turkey, and China, 1600 – 1850*; New York (NY), Routledge.
GRICE, P. (1975). "Logic and Conversation"; in *Syntax and Semantics*. Vol. 3. P. Cole and J.L. Morgan (eds.); New York (NY), Academic Press.

GRIFFITHS, A., GELBART W., J. MILLER and R. LEWONTIN (1999), *Modern Genetic Analysis*, New York (NY), Freeman.

HARTSOCK, M. (2010), *Absences as Causes: A Defense of Negative Causation*; PhD, University of Missouri-Columbia.

HEMPEL, C. and P. OPPENHEIM (1948), "Studies in the Logic of Explanation", *Philosophy of Science* 15, pp. 135-175.

KEIL, F. (2006), "Explanation and Understanding"; *Annual Review of Psychology* 57, pp. 227-264.

KHALIFA, K., J. MILLSON and M. RISJORD (2018), "Inference, Explanation, and Asymmetry", *Synthese* (online first.)

KUENEN, Ph. H. (1950), *Marine Geology*. New York (NY), John Wiley & Sons.

Kuhn, T. (1981)/(1963), "A Function for Thought Experiments"; in *Scientific Revolutions*, I. Hacking (ed.) Oxford, Oxford University Press, pp. 6-27.

KURAN, T. (2004), *Why the Middle East Is Economically Underdeveloped: Historical Mechanisms of Institutional Stagnation*; University of Southern California, Los Angeles (CA). *http://eppam.weebly.com/uploads/5/5/6/2/5562069/kuran.0130.pdf*

LANGE, M. (2016) *Because without Cause: Non-Causal Explanations in Science and Mathematics*; New York (NY), Oxford University Press.

LAUER, S. (2010), "Periphrastic Causative Verbs in English: What Do They Mean?"; Stanford University Department of Linguistics. http://www.sven-lauer.net/output/Lauer-QP-causatives.pdf

LEWIS, D. (1986), "Causal Explanation"; *Philosophical Papers*. II. Oxford, Oxford University Press**:** 214-240.

— (2004) [2000], "Causation as Influence"; in *Causation and Counterfactuals*. J. Collins, N. Hall and L. A. Paul (eds.), Cambridge (MA), MIT Press: pp. 5-106.

LIVENGOOD, J. and E. MACHERY (2007), "The Folk Probably Don't Think What You Think They Think: Experiments on Causation by Absence"; *Midwest Studies in Philosophy* XXXI, pp. 107-127.

MACKIE, J. (1980), *The Cement of the Universe: A Study of Causation*; Oxford, Oxford University Press.

MCGRATH, S. (2005) "Causation by Omission"; *Philosophical Studies* 123, pp. 125-148.

MCLANAHAN, S., L. TACH and D. SCHNEIDER (2013). "The Causal Effects of Father Absence"; *Annual Review of Sociology* 39, pp. 399-427.

MOORE, M. (2009), *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*; Oxford, Oxford University Press.

MUMFORD, S. and R. L. ANJUM (2011), *Getting Causes from Powers*; Oxford, Oxford University Press.

NEWMAN, M. (2012), "An Inferential Model of Scientific Understanding"; in *International Studies in the Philosophy of Science* 26(1), pp. 1-26.

— (2013). "Refining the Inferential Model of Scientific Understanding", *International Studies in the Philosophy of Science* 27(2), pp. 173-197.

— (2017), "Theoretical Understanding in Science"; *British Journal for Philosophy of Science* 68(2), pp. 571-595.

OAKHILL, J. (1984), "Inferential and Memory Skills in Children's Comprehension of Stories"; *British Journal of Educational Psychology* 54(1), pp. 31-39.
PSILLOS, S. (2002), *Causation and Explanation*; Stocksfield, Acumen.
QUINE, W. v. O. (1953), "Two Dogmas of Empiricism"; in *From a Logical Point of View*, W. v. O. Quine. Cambridge (MA), Harvard University Press, pp. 20-46.
REISS, J. (2015), *Causation, Evidence, and Inference*; New York (NY), Routledge.
RUSSELL, B. (1948), *Human Knowledge: Its Scope and Limits*; New York (NY), Simon & Schuster.
SALMON, W. (1984). *Scientific Explanation and the Causal Structure of the World*; Princeton, Princeton University Press.
— (1994), "Causality without Counterfactuals"; *Philosophy of Science* 61, pp. 297-312.
SCHAFFER, J. (2004a), "Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation"; in *Contemporary Debates in Philosophy of Science*, C. Hitchcock (ed.), Oxford, Blackwell, pp. 197-216.
— (2004b), "From Contextualism to Contrastivism"; *Philosophical Studies* 119(1), pp. 73-103.
— (2005), "Contrastive Causation", *Philosophical Review* 114(3), pp. 327-358.
SCRIVEN, M. (1966), "Causes, Connections and Conditions in History"; in *Philosophical Analysis and History*, W. Dray (ed.), New York (NY), Harper and Row, pp. 238-264.
SEN, A. (1999); *Development as Freedom*, Oxford, OUP.
SHOCKLEY, W. (1950), *Electrons and Holes in Semiconductors (with Applications to Transistor Electronics)*; Princeton (NJ), van Nostrand.
TAYLOR, H. and P. VICKERS (2017), "Conceptual Fragmentation and the Rise of Eliminativism"; *European Journal for Philosophy of Science* 7, pp. 17-40.
URBACH, P. and J. GIBSON, Eds. (1994), *Francis Bacon: Novum Organum*; Chicago and La Salle (IL), Open Court.
VAN EEMEREN, F. and R. GROOTENDORST (1992) *Argumentation, Communcation, and Fallacies*; London, Routledge.
WALTON, D. (2004), "A New Dialectical Theory of Explanation"; *Philosophical Explorations* 7(1), pp. 71-89.

teorema
Vol. XXXVIII/3, 2019, pp. 53-75
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 53-75]

Mechanistic Causation: Difference-Making is Enough

Stathis Psillos and Stavros Ioannidis

RESUMEN

En este artículo defendemos el punto de vista de que los mecanismos están respaldados por redes de relaciones que establecen diferencias. En primer lugar, distinguimos y criticamos dos tipos diferentes de argumentos a favor de entender los mecanismos a partir de la noción de actividad: un enfoque que prioriza metafísica (Glennan) y otro que prioriza la ciencia (Illari y Williamson). En segundo lugar, presentamos un punto de vista alternativo de los mecanismos entendiéndolos en términos del establecimiento de diferencias y lo ilustramos examinando un caso histórico: la prevención del escorbuto. Usamos este ejemplo para argumentar que la evidencia a favor de un mecanismo no algo distinto a la evidencia a favor de relaciones que establecen diferencias.

PALABRAS CLAVE; *mecanismo, causación, producción, actividades, diferenciación, escorbuto.*

ABSTRACT

In this paper we defend the view that mechanisms are underpinned by networks of difference-making relations. First, we distinguish and criticise two different kinds of arguments in favour of an activity-based understanding of mechanism: Glennan's metaphysics-first approach and Illari and Williamson's science-first approach. Second, we present an alternative difference-making view of mechanism and illustrate it by looking at the history of the case of scurvy prevention. We use the case of scurvy to argue that evidence for a mechanism just is evidence for difference-making relations.

KEYWORDS: *Mechanism, Causation, Production, Activities, Difference-Making, Scurvy.*

I. INTRODUCTION

Causal relations are explanatory. If C causes E then C explains the occurrence of E. Mechanisms are widely taken to be both what makes a relation causal and what makes causes explanatory. So, typically, if one explains the occurrence of event E by citing its cause C, i.e., if one asserts that C brings about E or that E occurs because of C, one is ex-

pected to cite the mechanism that links the cause and the effect: it is in virtue of the intervening mechanism that C causes E and hence that C causally explains E. On this account of causation, it is not enough to show that E depends on C — where dependence should be taken to be robust, e.g., a difference-making relation. Unless there is a mechanism, there is no causation. Difference-making is taken to be enough for prediction and control but not enough for explanation [cf. Williamson (2011)].

Now, when it comes to causation there are two competing views available: production and dependence [cf. Psillos (2004)]. On the production account, C causes E iff C produces E. 'Production' is a term of art, of course, with heavily causal connotations. The typical way to account for 'production' is by means of mechanism. So, C produces E iff there is a mechanism that links C and E. On the dependence account, C causes E iff C makes a difference to E. This difference-making is typically seen as counterfactual dependence, viz., if C hadn't happened, then E wouldn't have occurred. As is well-known, both views face problems and counterexamples. For instance, the production account cannot accommodate causation by absences. The lack of water caused the plant to die, but there is no mechanism linking the *absence* of water with death. The difference-making account cannot accommodate cases of overdetermination and pre-emption. For instance, suppose that two causes act independently of each other to produce an effect. There is certainly causation, but no difference-making since the effect would be produced even in the absence of each one of the causes [cf. Williamson (2011)].

The key aim of the present paper is to defend the view that difference-making is more fundamental than production in understanding mechanistic causation. In particular, we shall argue that mechanisms are best understood as networks of difference-making relations. To do this, we shall criticise the popular idea that the productivity of mechanisms requires commitments to activities, qua a sui generis ontic category. There are two routes to this popular view, one top-down and another bottom up. The top-down approach, most ably defended by Stuart Glennan (2017), is the metaphysics-first approach. On this view, in order to account for what mechanisms are *as things in the world*, activities must be posited as a distinctive metaphysical item. Activities are taken to be components of mechanisms, distinct from entities and their properties, and are supposed to account for what makes a mechanism productive. The bottom-up approach, recently defended by Phyllis Illari

and John Williamson (2011), is the science-first approach. On this view in order to account for the pervasive role of mechanisms in science, and in particular, for the fact that mechanisms are (spatio-temporally) localised, we have to think of mechanisms as embodying activities. Sections II and III respectively will argue against both approaches to activities. Section II will show that there is no need to hypostatise activities over and above the properties and relations of things that make up causal pathways; section III will show that the 'local' argument for activities does not make a case for an activities-based understanding of mechanisms.

Section IV will revisit activities, this time as part of a productive account of causation. It will be argued that the very idea of production requires difference-making relations. Finally, in section V we will look in some detail at the history of the case of scurvy prevention. This case will drive home the point that it is enough to understand mechanisms as underpinned by relations of difference-making.

II. Against Activities 1

What is a mechanism? Glennan puts forward what he calls Minimal Mechanism: "a mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organised in such a way that they are responsible for the phenomenon" [Glennan (2017), p. 13]. Though minimal, this account is "an expansive conception of what a mechanism is" [Ibid. p. 106], mostly because it involves commitment to activities as a novel ontological category. "Activities" Glennan claims "(…) cannot naturally be reduced to properties of or relations between entities" [Ibid. p. 50].[1]

Given that activities play a key role in the mechanistic accounts of causation, it's important to be clear on what they are supposed to be. Here then are some characteristics of activities, according to Glennan.

Activities are concrete: "they are fully determinate particulars located somewhere in space and time; they are part of the causal structure of the world [Ibid. p. 20]. Activities are the ontic correlate of verbs. They include anything from walking, to pushing, to bonding (chemically or romantically) to infecting. Given this, activities "are a kind of process — essentially involving change through time" [Ibid. p. 20]. Some activities are non-

relational (unary activities) since they involve just one entity, e.g., a solitary walk. But some activities involve interactions: they are non-unary activities, viz., activities which implicate more than one entity [Ibid. p. 21].

Most activities, Glennan says, "just are mechanistic processes", i.e., spatio-temporally extended processes which "bring about changes in the entities involved in them" [Ibid. p. 29]. What, then, is a *mechanistic* process? According to Glennan, "To call a process mechanistic is to emphasise how the outcome of that process depends upon the timing and organisation of the activities and interactions of the entities that make up the process" [Ibid. p. 26].

Now, it appears that there is a rather tight circle here. A process is mechanistic when the entities that make it up engage in *activities*. But if activities just are *mechanistic processes*, then a process is mechanistic when the entities that make it up engage in mechanistic processes. Not much illumination is achieved. Perhaps, however, Glennan's point is that activities and processes are so tightly linked that they cannot be understood independently of each other. Yet, there seems to be a difference—activities (are meant to) imply action. To describe something as an activity is to imply that something acts or that an action takes place. A process need not involve action. It can be seen as a (temporal or causal) sequence of events. In fact, it might be straightforward to just equate the mechanism with the process, viz., the causal pathway that brings about an effect. In the sciences all kinds of processes are characterised as mechanistic irrespective of whether they are 'active' or not. Let us illustrate this point by a brief discussion of the case of active vs passive membrane transport, which are the two mechanisms of transporting molecules across the cell membrane. The transportation of the molecules takes place across a semi-permeable phospholipid bilayer and is determined by it. Some molecules (small monosaccharides, lipids, oxygen, carbon dioxide) pass freely the membrane through a concentration gradient whereas other molecules (ions, large proteins) pass the membrane against the concentration gradient and use cellular energy. The main difference between active and passive transport is precisely that in active transport the molecules are pumped using ATP energy whereas in the passive transport the molecules pass through the gradient by diffusion or osmosis. These different mechanisms play different roles. Active transport is required for the entrance of large, insoluble molecules into the cell, whereas passive transport allows the maintenance of a homeostasis between the cytosol and extracellular fluid. But they are both causal processes or pathways, even though only one of them is 'active'.

Glennan (2017), p. 32, takes it that "the most important feature of activities" is that most or all activities are mechanism-dependent. This, he thinks, suggests that "the productive character of activities comes from the productive relations between intermediates in the process, and that the causal powers of interactors derive from the productive relations between the parts of those interactors".

But this is not particularly illuminating. Apart from the fact that production is itself an activity, to explain the productive character of activities by reference to the productive activity of intermediaries, or of the constituent parts of the mechanism, just pushes the issue of the productivity of an activity A to the productivity of the constituent activities A1,…, An of the mechanism that realises A. Far from explaining how activities are productive, it merely assumes it. Now, Glennan takes an extra step. He takes it that some producings are explained "in terms of other producings, not in terms of some non-causal features such as regularity, or counterfactual dependence" [Ibid. p. 33]. In the context in which we are supposed to try to understand what distinguishes activities from non-activities, this kind of argument is simply question-begging.

If what makes entities engage in activities are their properties and relations to other entities in what sense are activities things distinct from them? In what sense are activities "a novel ontological category"? Here, we find Glennan's argument perplexing. His chief point is that thinking of activities as fixed by the properties and relations of things "reduces doing to having; it takes the activity out of activities" [Ibid. p. 50]. The language of relations "is a static language" [Ibid.]. But activities, we are told, are "dynamic" [Ibid. p. 51].

Let us set aside this figurative distinction between doing and having. After all, it is in virtue of having mass that bodies gravitationally attract each other, according to Newton's theory of gravity. More generally, it is by virtue of having properties that things stand in relations to each other, some of which are 'static' e.g., being taller than, while others are 'dynamic', e.g., being attracted by. To see why activities do not add something novel to ontology, let us stress that for Glennan activities are fully concrete particulars: "Any particular activity in the world will be fully concrete, though our representations of that activity may be more or less abstract" [Ibid. pp. 95-96]. Now, if activities are always particular, and if they are always specific, like pushings, pullings, bondings, infectings, dissolvings, diffusings, pumpings etc. there is no need to think of them as comprising a novel ontic category. For each fully concrete activi-

ty, there will be some account in terms of entities, their properties and relations. A pushing is an event (or a process) which consists in an object changing its position (over time) due to the impact by another body. Indeed, the very event itself *consists* in a change of the properties of a thing (or of its relations to other things). Similarly, for other concrete activities: there will always be some description of the event or the process involved by reference to the changes of the properties of a thing (that engage in the 'activity') or to the relations with other things.

Take the case of a mechanism such as the formation of a chemical bond. Chemical bonding refers to the attraction between atoms. It allows the formation of substances with more than one atomic component and is the result of the electromagnetic force between opposing charges. Atoms are involved in the formation of chemical bonds in virtue of their valence electrons. There are mainly two types of chemical bonds: ionic and covalent. Ionic bonds are formed between two oppositely charged ions by the complete transfer of electrons. The covalent bond is formed by the complete transfer of valence electrons between bonded atoms. Such type of bond is formed by the equal sharing of electrons between two bonded atoms. These atoms have equal contribution to the formation of the covalent bond. On the basis of the polarity of a covalent bond, it can be classified as a polar or non-polar covalent bond. Electronegativity is the property of an atom in virtue of which it can attract the shared electrons in a covalent bond. In nonpolar covalent bonds, the atoms have similar EN. Differences in EN yield bond polarity.

In *describing* this mechanism, there was no need to think of particular activities as anything other than events (sharing of electrons) or processes (transfer of valence electrons) that are fixed by the properties of atoms (their valence electrons; Electronegativity) and the relations they stand to each other (similar or different EN).

Glennan, however, takes it that "processes are collections of entities acting and interacting through time" [Ibid. p. 57]. Elsewhere [Ibid. p. 83], he notes that a mechanism is a "sequence of events (which will typically be entities acting and interacting)". If we were to follow Bishop Berkeley's advice to '*think* with the *learned* and *speak with the vulgar*' we could grant this talk in terms of activities, without hypostatising activities over and above the properties and relations by virtue of which entities 'act and interact'. We conclude that 'activity' is an abstraction without ontological correlate.

When he talks about entities, Glennan takes it that a general characteristic of entities is this: "The causal powers or capacities of entities are

what allow them to engage in activities and thereby produce change" [Ibid. p. 33]. What produces the change? It seems Glennan's dualism requires that there are causal powers *and* activities and that the former enable the entities that possess them to engage in activities, thereby producing changes (to other entities). It's as if the activities exist out there ready to be engaged with by entities having suitable causal powers. Glennan is adamant: "activities are not properties or relations; they are things that an entity or entities do over some period of time" [Ibid. p. 96].

But this cannot be right. The activities cannot exist independently of the entities and their properties (whether we conceive of them as powers or not). What activities an entity can 'engage with' depends on the properties of this entity. Water can dissolve salt but not iron, to offer a trivial example. The 'activities' an entity can engage in are none others than those that result from the kind of entity it is. If you assume powers, as Glennan does, then the activities of an entity are fixed by the manifestation of its powers (given suitable circumstances). Given a power ontology, the powers are the producers of change; the activities are merely the manifestation of powers.

As Glennan admits: "The central difference between activities and powers is that activities are actual doings, while powers express capacities or dispositions not yet manifested" [Ibid. p. 32]. As just noted, assuming particulars with powers, activities are the manifestation/exercising of these powers. When a cube of salt is put in water, it dissolves. The dissolving is the manifestation (assuming a power-ontology) of the active power of water to dissolve (water-soluble) materials and the passive power of the salt to get dissolved. The dissolving takes time (and hence it is a process); but it is not acting in any sense; it does not produce any changes in the salt; it *consists* in the changes in the salt. The 'scraping of the skin off the carrot' (Glennan's example) *is* the removal of the skin of the carrot (at least on this particular occasion) and hence it does not cause (or produce) the removal. Activities do not produce anything; they *are* the productions (of effects).

III. Against Activities 2

While Glennan's motivation for activities comes from the metaphysics of mechanisms, other philosophers vouch for activities on the grounds that science requires them. The general motivation appears to

be that science must constrain metaphysics. Not only is it the case that what there is has to be compatible with what science describes, but also the best route to the fundamental structure of the world should be the descriptions that science offers. Thus, proponents of activities have argued that if we take seriously the descriptions offered in such fields as molecular biology or neurobiology, we find that activities are central in these descriptions [Machamer, Darden & Craver (2000); Illari & Williamson (2013)]. lllari & Williamson, in particular, think that "[t]here is a good argument from the successful practice of the biological sciences for the appeal to activities in the characterisation of a mechanism" [Illari & Williamson (2013), p. 71].

Illari & Williamson (2011) offer a bottom-up argument in favour of what they call an 'active metaphysics' for the workings of mechanisms, by which they mean a metaphysics in terms of capacities [cf. Cartwright (1989)] or of powers [cf. Gillett (2006)] or of activities [cf. Machamer, Darden & Craver (2000)]. They contrast active metaphysics with 'passive' metaphysics, which characterises the working of mechanisms in terms of laws or counterfactuals. In what follows we are going to examine this kind of bottom-up argument, which we are going to call the 'local argument'.

Although we are here treating the local argument as an argument in favour of activities, Illari & Williamson take the argument to be more general, as it does not differentiate between activity-based and power-based views. In fact, in their (2013) Illari & Williamson offer reasons to prefer an ontology based on entities and activities over an ontology based on entities and capacities, a main reason being that an ontology of activities is more parsimonious. But since these arguments are largely metaphysical, and we are here focusing on bottom-up arguments, we are going to examine the local argument in its general form.

Illari & Williamson argue that biological practice, and in particular the fact that mechanisms are taken to be explanatory, constrains the ontology of mechanisms. More specifically, they think that a metaphysics of mechanisms that views within-mechanism interactions in terms of laws or counterfactuals, is "in tension with the actual practice of mechanistic explanation in the sciences, which examines only local regions of spacetime in constructing mechanistic explanations." So, passive approaches do not "allow mechanisms to be real and local (…) only active approaches give a local characterisation of a mechanism" [Illari & Williamson (2013), p. 835]. They think then that the local argument establishes that a characterisation of mechanism has to be given in terms of an

active metaphysics and not in terms of "counterfactual notions grounded in laws or other possible worlds" [Ibid. p. 838].

The local argument can be reconstructed as follows:

> The practice of mechanistic explanation requires that mechanisms be local (1). This in turn implies that a characterisation of mechanism has to be local (2).
>
> But only a metaphysics of powers or activities is a local metaphysics (3).
>
> So, a local characterisation of mechanism requires a metaphysics based on powers or activities (4) (2013, 834-838).

In response to this argument for an 'active' metaphysics of mechanisms, it seems to us that 'local' cannot have the same meaning in premises (1) and (2), on the one hand, and in premise (3), on the other: we can have local mechanisms without a local metaphysics. There are three points to note here.

First, it is certainly true that mechanisms are local to the phenomena they produce. In this context, 'local' means that mechanistic explanation involves the localisation of the parts into which the mechanism is decomposed, the operations of which produce the phenomenon for which the mechanism is responsible. Indeed, as Bechtel & Richardson (2010) have argued, localisation is a central strategy in constructing a mechanistic explanation: scientists decompose the phenomenon under study into component operations, and "localise them within the parts of the mechanism" [Ibid Introduction, p. xxx]. But then, localisation of parts can fully capture the sense in which mechanisms are 'local', without entailing a 'local' metaphysics, which is supposed to underlie a characterisation of the interactions among components, and not only the components themselves. Even if we accept a metaphysics of laws, within-mechanism interactions are interactions between 'local' components.

Second, it is not at all easy to account for within-mechanism interactions in terms of a 'local' metaphysics. Energy transformations in biological systems obey the laws of thermodynamics. But it is very difficult to reconcile a power ontology with what it seems to be a global principle, like the law of conservation of energy. This is something that friends of powers themselves have recognised [cf. Ellis (2001)]. So, contra Illari & Williamson, a focus on practice seems in fact to imply the opposite conclusion: global principles like the laws of thermodynamics are needed for

accounting for within-mechanism interactions (e.g. as studied by bioenergetics, cf. Nelson et al (2008), p. 489); but only a metaphysics in terms of laws seems to offer an adequate account of such global principles; so, a metaphysics of laws is required for a characterisation of the metaphysics of mechanisms. Again, the point here is that 'local' decompositions of mechanistic parts must be kept distinct from 'global' or 'local' ways to characterise interactions.

Third, there is a historical point to be made against the argument that mechanistic explanation is not compatible with a metaphysics of laws. This combination ('local' mechanisms that produce phenomena plus laws of nature) was a dominant view in 17th century mechanical philosophy. Contemporary mechanistic explanations, of course, are very different from their 17th century counterparts, which in many cases just involved parts of matter in motion. But the general pattern of explanation is similar: in giving a mechanistic explanation, one shows how the particular properties of the parts, their organisation and their interactions (which can be captured in terms of the laws that govern them), produce the phenomena.

In view of the previous points, premise (3) above can only be accepted if the meaning of 'local' is disambiguated. An option here is to say that mechanisms have to be local, in the sense that within-mechanism interactions have to be grounded in facts in the vicinity of the mechanism. So, one can think of causation as a local matter, i.e. as a relation between the two events that are causally connected, and not as a global matter, i.e. as involving a regularity. But note that so-called singular causation is compatible with a metaphysics of laws. One can view causation as a relation between 'local' events, but at the same time adopt an ontology of laws, where laws could be, for example, necessitating relations between universals, or humean regularities, i.e. 'global' facts about the universe [cf. Ioannidis & Psillos (2018)].

Note that Illari & Williamson themselves seem to recognise that in understanding scientific practice one need not talk about metaphysics, for they say: "Understanding the metaphysics of mechanisms on this level is now a philosophical problem with no immediate bearing on scientific method, of course" [Illari & Williamson (2011), p. 834]. But they add: "It does, however, bear on our understanding of science" [Ibid. p. 834]. While we agree with the first sentence, we believe (and we shall argue below) that an understanding of mechanism as causal pathways, underpinned by difference-making relations is all one needs in order to understand scientific practice. We conclude, then, that there is no reason coming from

scientific practice for accepting a power-based or an activities-based account of mechanism.

IV. Causation as Production

This last point, viz., that difference-making relations are enough to understand mechanisms and hence mechanistic causation and explanations, is contested. Many philosophers take it that causation is production. Glennan, for instance, is one of the defenders of this view. According to him, mechanisms, qua productive, are the truth-makers of causal claims:

> (MC) A statement of the form 'Event c causes event e' will be true just in case there exists a mechanism by which c contributes to the production of e [Glennan (2017), p. 156].

Actually, there are as many causal relations as there are activities. As he puts it: "There is on this view [the new mechanist view] no one thing which is interacting or causing, and when we characterise something as a cause, we are not attributing to it a particular role in a particular relation, but only saying that there is some productive mechanism, consisting of a variety of concrete activities and interactions among entities" [Ibid. p. 148]. This pluralist view leads him to the radical conclusion that "There is (…) no such thing as THE ontology or THE epistemology of THE causal relation, but only more localised accounts connected with the particular kinds of producing" [Ibid. p. 33].

MC tallies with Glennan's singularism about causation. All causings are singular and in fact fully distinct from each other. Singularism is committed to the view that causation is internal (intrinsic, as Glennan puts it) to its relata. Glennan shares this intuition. He says: "Productive causal relationships are singular and intrinsic. They involve continuity from cause to effect by means of causal processes" [Ibid. p. 154].

But is causation a relation, after all? And if yes, what are the *relata*? Events, is the answer that springs to mind. Glennan agrees but takes events to involve activities: "Events are particulars — happenings with definite locations and durations in space and time. They involve specific individuals engaging in particular activities and interactions" [Ibid. p. 149]. Or as he put it elsewhere: "an event is just one or more entities engaging in an activity or interaction" [Ibid. p. 177].

We have already argued in section 2 that activity is far from being a sui generis ontic category. Besides, there is the received account of events as property-exemplifications: events are exemplifications of properties (or relations) by an object (or set of objects) at a time (or a period of time). As Glennan admits: "If exemplifying a property were the same as engaging in an activity, then the two views would coincide". However, he takes it that "there are important differences between exemplifying properties and engaging in activities" [Ibid. p. 177].

The chief difference between property-exemplification and engaging in activities is, Glennan says, that "properties are paradigmatically synchronic states of an entity that belong to that entity for some time." Unlike activities, properties "do not involve change". Events, Glennan argues, "involve changes". It is indeed true that events involve change. The collision of the Titanic with the iceberg took time and during it, both the Titanic and the iceberg suffered changes in their properties, which resulted in another event, viz. the sinking of the Titanic. It is true that to account for this we have to introduce relations: the collision is between the Titanic and the iceberg. But relations, we are told, are not "activity-like". Glennan insists: "only events (which involve activities) can be causally productive". Properties, he says, "cannot produce anything" [Ibid. p. 178].

When all is said and done, the key question is: is causation production? Or is it difference-making? Glennan is clear: "While I grant that production and relevance are two different concepts of cause, I will argue that production is fundamental" [Ibid. p. 156].

Descriptively, Glennan distinguishes between three kinds of productive relations:

- Constitutive production: An event produces changes in the entities that are engaging in the activities and interactions that constitute the event.
- Precipitating production: An event contributes to the production of a different event by bringing about changes to its entities that precipitate a new event.
- Chained production: An event contributes to the production of another event via a chain of precipitatively productive events [Ibid. p. 179].

All this is fine but what is the chief argument for causation being *production*?

It seems to be this: "Mechanisms provide the ontological grounding that allows causes to make a difference" [Glennan (2017), p. 165].

Glennan's problem with the claim that mechanism is itself a network of relations of difference-making between events is that on the difference-making account "the causal claim depends upon the truth of a counterfactual, whereas on the mechanist account the truth depends upon the existence of an actual mechanism" [Ibid. p. 167]. Furthermore, it is claimed that the truth of the counterfactual requires contrasting an actual situation — where the cause occurs — and a non-actual but possible situation in which the cause does not occur.

Does the production account avoid counterfactuals? Glennan acknowledges that causation as production relies on some notion of relevance but takes this to require actual difference-makers. He takes it that actual difference-makers are "features of the actual entities and their activities upon which outcome depends" [Ibid. p. 203].

What is an *actual* difference-maker? A factor such that had it not happened, the effect would not have followed. But a) in an actual concrete sequence of events which brought about an effect x, all events were necessary in the circumstances; all were difference-makers. If any of them were absent, the effect, in its full concrete individuality, would not follow. A different effect would have followed. But b) what makes true the counterfactual that 'had x not actually happened, y would not have followed'? To 'delete' x from the actual sequence is to envisage a counterfactual sequence (that is, a distinct sequence of events) without x. It is then to compare two sequences: the actual and the counterfactual. This requires thinking in terms of counterfactual difference-making. What makes the counterfactual true is not the actual sequence of events but the fact, if it is a fact, that xs are followed by ys, which is a causal law.

Take the example of a ball striking a window while a canary nearby sings. The actual causal situation — the mechanism in all its particularity — includes the process of the acoustic waves of the canary's singing striking the window (say, for convenience, at the moment when the ball strikes the window) as well as the kinetic energy of the ball (which was a red cricket ball) etc. Despite the fact that the acoustic waves are part of the actual concrete mechanism and clearly contributed to the actual breaking (no matter how little), we would not say that it was the singing that caused the window-breaking. It clearly didn't make a substantial contribution to the breaking. Had it not been there, the window would still have shattered. How can *this* counterfactual be made true by the actual situation? In the actual situation, the singing was a difference-maker since it was part of the mechanism that made the difference. To show

that it did not make a difference (better put, that it made a difference without a difference) we have to compare the actual situation in which the singing took place and a non-actual but possible situation in which the singing did not happen. Whatever makes this counterfactual true, it is not the actual situation, in and of itself.

Not only production does not avoid counterfactuals (if actual difference-makers are to be shown that did not make a difference) but it seems that the very idea of production requires difference-making relations if the producer of change is nothing more specific than everything that happened before the effect took place.

V. Causation as Difference-Making: the Case of Scurvy

Given the difficulties with activities and the mechanistic production outlined above, it seems more promising to start with difference-making and give an account of mechanisms in terms of it. Such difference-making accounts of mechanism have been offered by various authors. For James Woodward (2002), difference-making is required to account for within-mechanism interactions. As he puts it, "components of mechanisms should behave in accord with regularities that are invariant under interventions and support counterfactuals about what would happen in hypothetical experiments" [Woodward (2002), p. 374]. Peter Menzies (2012) uses the interventionist approach to causation to give an account of the causal structure of mechanisms.

More recently, Gillies (2017) and Ioannidis & Psillos (2017; 2018) have offered difference-making accounts of mechanism by discussing particular case-studies. Common to both of these more recent accounts is the thought that a mechanism in life sciences should be viewed as a causal pathway connecting a cause with a particular effect. Gillies sums up his account as follows: "Basic mechanisms in medicine are defined as finite linear sequences of causes (C1 → C2 → C3 → … → Cn), which describe biochemical/ physiological processes in the body. This definition corresponds closely to the term 'pathway' often used by medical researchers. Such basic mechanisms can be fitted together to produce more complicated mechanisms which are represented by networks" [Gilles (2017), p. 633].

In our (2017; 2018) we have argued that when scientists talk about a 'mechanism', what they try to capture is the way (i.e. the causal pathway) a certain result is produced. Suppose, for instance, that pathologists want to find out how a certain disease is brought about. They look for a

specific mechanism, i.e. a causal pathway that involves various causal links between, for example, a virus and changes in properties of the organism that ultimately lead to the disease. In pathology, such causal pathways constitute the pathogenesis of a disease, and when pathologists talk about the mechanisms of a disease, it is such pathways that they have in mind [cf. Lakhani et al (2009)]. This leads to the following view: "[t]o identify a mechanism … is to identify a specific causal pathway that connects an initial 'cause' (the causal agent) with a specific result" [Psillos and Ioannidis (2017), p. 604]. So, mechanisms in biomedicine are "stable causal pathways, described in the language of theory" [Ibid. (2018), p. 1181], where to identify a *causal* pathway is to identify difference-making relations among its components.

Moreover, we have argued that in giving a characterisation of mechanism as a concept of scientific practice, one need not be committed to a specific view on the metaphysics of mechanisms: mechanism in our sense is a concept used in scientific practice and as such it is primarily a methodological concept. An important point here is that if we take this truly minimal account of mechanisms, then the burden is on the defender of a particular metaphysical characterisation of mechanism to say why such a methodological account is not enough and why it should be inflated with metaphysical categories (such as entities and activities).

To motivate further this difference-making account of mechanism, as well as the view that difference-making is prior to production, let us look briefly at the case of scurvy. This, we now know, is a disease resulting from a lack of vitamin C (ascorbic acid). If you think of it, the *absence* of vitamin C in an organism causes scurvy, which starts with relatively mild symptoms (weakness, feeling tired, and sore arms and legs) and if it remains untreated it may lead to death. If we take seriously the thought that absences, qua causes, are counterexamples to mechanistic causation, we should conclude that there is no mechanistic explanation of scurvy. But this would be clearly wrong. What is correct to say is that the lack of vitamin C disrupts various biosynthetic causal pathways, that is, mechanisms, e.g., the synthesis of collagen. In the latter process, ascorbic acid is required as a cofactor for two enzymes (prolyl hydroxylase and lysyl hydroxylase) which are responsible for the hydroxylation of collagen. Some tissues such as skin, gums, and bones contain a greater concentration of collagen and thus are more susceptible to deficiencies. But ascorbic acid is also required in the enzymatic synthesis of dopamine, epinephrine, and carnitine. Now, humans are unable to synthesise ascorbic acid, the reason

being that humans possess only 3 of the 4 enzymes needed to synthesise it; (the fourth enzyme seems to be defective). Hence humans have to take vitamin C through their diet [for a useful survey cf. Magiorkinis et al. (2011)].

The disrupted causal pathways that prevent scurvy can be easily accommodated within the difference-making account of causation. Had vitamin C been present in the organism x, x wouldn't have developed scurvy. In fact, the very causal pathway can be seen as a network of relations of dependence (or difference-making). Abstractly put, had vitamin C been present in human organism x, x's lack of working GULO enzyme would not have mattered; enzymes prolyl hydroxylase and lysyl hydroxylase would have been produced etc. and scurvy would have been prevented. [For a description of the causal pathways of the synthesis of vitamin C in the mammals that can synthesise it, see Linster & Van Schaftingen (2007)].

The history of scurvy is really interesting. During the Age of Exploration (between 1500 and 1800), it has been estimated that scurvy killed at least two million seamen. Although there were hints that scurvy is due to dietary deficiencies, it was not until 1747 that it was shown that scurvy could be treated by supplementing the diet with citrus fruits. In what is taken as the first controlled clinical trial reported in the history of medicine, James Lind, naval surgeon on HMS *Salisbury*, took 12 patients with scurvy "on board the Salisbury at sea" [Lind (1753), p. 149]. As he reported, "Their cases were as similar as I could have them". The patients were kept together "in one place, being a proper apartment for the sick" and had "one diet in common to all". He then divided them to 6 groups of 2 patients and each of which was allocated to 6 different daily treatments for a period of 14 days. One group was administered 2 oranges and 1 lemon per day for 6 days only, "having consumed the quantity that could be spared" [Ibid. p.150]. The other groups were administered cyder, elixir vitriol, vinegar, sea-water, and a concoction of various herbs, all of which were supposed to be anti-scurvy remedies. As Lind put it: "The consequence was that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit four duty, (…) (t)he other was the best recovered of any in his condition" [Ibid.]. Lind's experiments provided evidence that citrus fruits could cure scurvy. He said that oranges and lemons are "the most effectual and experienced remedies to remove and prevent this fatal calamity" [Ibid. p. 157].

Though Lind had identified a difference-maker, he was sidetracked by looking for the cause of scurvy, which he found in the moisture in the air, though he did admit that that diet may be a secondary cause of scur-

vy [cf. Bartholomew (2002); Carpenter (2012)]. But in 1793 his follower, Sir Gilbert Blane, who was the personal physician to the admiral of the British fleet, persuaded the captain of HMS *Suffolk* to administer a mixture of two-thirds of an ounce of lemon juice with two ounces of sugar poured to each sailor on board. As Blane reported the warship "was twenty-three weeks and one day on the passage, without having any communication with the land (…) without losing a man" [quoted by Brown (2003), p. 222]. To be sure, scurvy did appear, but it was quickly relieved by an increase in the lemon juice ration. When in 1795 Blane was appointed a commissioner to the *Sick and Hurt Board*, he persuaded the Admiralty to issue lemon juice as a daily ration aboard all Royal Navy ships. He wrote: "The power [lemon juice] possesses over this disease is peculiar and exclusive, when compared to all the other alleged remedies" [cf. op.cit.]. But even when it was more generally accepted that citrus fruits prevent scurvy, it was the acid that was believed to cure scurvy.

The first breakthrough took place in 1907 when two Norwegian physicians, Axel Holst and Theodor Frølich, looked for an animal model of beriberi disease. They fed guinea pigs with a diet of grains and flour and found out, to their surprise, that they developed scurvy. They found a way to cure scurvy by feeding the guinea pigs with a diet of fresh foods. This was a serendipitous event. Most animals are able to synthesise vitamin C; but not guinea pigs. In 1912, in a study of the etiology of deficiency diseases, Casimir Funk suggested that deficiency diseases (such as beriberi and scurvy) "can be prevented and cured by the addition of certain preventive substances". He added that "the deficient substances, which are of the nature of organic bases, we will call 'vitamines'; and we will speak of a beriberi or scurvy vitamine, which means, a substance preventing the special disease" [Funk (1912), p. 342]. By the 1920s, the 'anti-scurvy vitamine' was known as 'C factor' or 'anti-scorbutic substance' [cf. Hughes (1983)]. In 1927, Hungarian biochemist Szent-Györgyi isolated a sugar-like molecule from adrenals and citrus fruits, which he called 'hexuronic acid'. Later on, Szent-Györgyi showed that the hexuronic acid was the sought-after anti-scorbutic agent. The substance was renamed 'ascorbic acid'. In parallel with Szent-Györgyi's work, Charles King and W. A. Waugh identified, in 1932, vitamin C. The suggestion that hexuronic acid is identical with vitamin C was made in 1932, in papers by King and Waugh and by J. Tillmans and P. Hirsch [cf. Hughes (1983)].

The breakthrough in scurvy prevention occurred when scientists started to look for what has been called 'the mediator', which is a code-

word for the 'mechanism', which "transmits the effect of the treatment to the outcome" [Pearl & Mackenzie (2018), p. 270)]. As Baron and Kenny put it, mediation "represents the generative mechanism through which the focal independent variable is able to influence the dependent variable of interest" [Baron and Kenny (1986), p. 1173]. This mechanism, however, is nothing over and above a network of difference-makers: Citrus Fruits → Vitamin C → Scurvy. One such difference-maker, citrus fruits, was identified by Lind and later on by Blane. This explains the success in preventing scurvy after citrus fruits were administered as part of the diet of sailors. It is noteworthy, however, that Lind and the early physicians did not look for the mediating factor in the case of scurvy. As Bartholomew (2002), p. 696, notes, Lind did not try to isolate a single common constituent in citrus fruits in particular and in fruit in general which makes a difference to the incidence of scurvy. Instead he was trying to find out the contribution of different sorts of vegetable to the relief from scurvy. Still, even without knowing the mediating variable (vitamin C), the intake of citrus in a diet did make a difference to scurvy relief.

In order to find the difference-maker in the case of vitamin C deficiency it was necessary to find a model (animals) that does not synthesise its own vitamin C. In the late 1920a, Szent-Györgyi and his collaborator J. L. Svirbely used the recently isolated by Szent-Györgyi hexuronic acid to treat the animals in controlled experiments with guinea pigs. They divided the animals into two groups. In one the animals were fed with food enriched with hexuronic acid, while in the other the animals received boiled food. The first group flourished while the other developed scurvy. Svirbely and Szent-Györgyi decided that hexuronic acid was the cause of scurvy relief and they renamed it ascorbic acid. Ascorbic acid was the sought-after mediating variable: the difference-maker [cf. Schultz (2002).

It is useful to discuss the case of scurvy in relation to what has become known in the recent philosophical literature on mechanisms as the Russo-Williamson thesis (RWT) ([Russo & Williamson (2007)], i.e. that in the health sciences, in order to establish a causal connection between A and B, one needs evidence both for the existence of a difference-making relation between A and B *and* of a mechanism linking A to B. Williamson (2011) relies on this thesis to raise a problem for mechanistic and difference-making theories of causation. The problem is supposed to be that these theories, taken on their own, are not compatible with the causal epistemology adopted in biomedicine and other scientific fields, which conforms to RWT.

This argument seems to raise a problem for the difference-making account of mechanism presented in the beginning of this section. If A causes B in virtue of a mechanism linking A to B, where a mechanism involves a chain of events linked by difference-making relations, it seems that evidence of difference-making is enough to establish a causal claim, contrary to what RWT asserts. In other words, 'mechanistic' evidence need not be different in kind from difference-making evidence. However, Williamson & Wilde (2016) assume that there is a distinction between these two kinds of evidence. They think that "in order to establish that *A* is a cause of *B* there would normally have to be evidence both that (i) there is an appropriate sort of difference-making relationship (or chain of difference-making relationships) between *A* and *B* — for example, that *A* and *B* are probabilistically dependent, conditional on *B*'s other causes —, and that (ii) there is an appropriate mechanistic connection (or chain of mechanisms) between *A* and *B* — so that instances of *B* can be explained by a mechanism which involves *A*" [Ibid. p. 38].

In contrast to this, the case of scurvy shows that looking for mechanistic evidence is just looking for a special kind of 'difference-making' evidence and not for a different kind of evidence. This special difference-making evidence involves looking for the 'mediator'. As we have seen, Lind's experiments provided evidence for a difference-making relationship between Citrus Fruits and Scurvy, but no evidence about how exactly Citrus Fruits acted so as to prevent scurvy. When it was realised by Funk that scurvy is a 'deficiency disease', i.e. it was produced because of the lack of a particular substance, it became obvious that Citrus Fruits acted to prevent Scurvy by providing that preventive substance. So, scientists started looking for this preventive substance that was the mediating factor between Citrus Fruits and Scurvy. As we have already seen, however, what was required for finding the mediator and establishing the pathway Citrus Fruits → Vitamin C → Scurvy, was the isolation of a substance (hexuronic acid) from citrus fruits that was such as to prevent scurvy in controlled experiments with guinea pigs by Svirbely and Szent-Györgyi. So, the evidence for identifying the mediator was not evidence about particular entities engaging in activities, or some sui generis type of mechanistic evidence, as one would have believed if the activities-based account of mechanism were true; it was evidence about more difference-making relations, this time between the two initial variables (Citrus Fruits and Scurvy) and the mediating variable Vitamin C.

The case of scurvy thus shows that RWT can be accepted, without being committed to the existence of a special type of 'mechanistic' evidence over and above difference-making relations. Moreover, acceptance of RWT does not automatically lead to a rejection of a difference-making account of causation. Given a difference-making account of mechanisms, RWT can be understood as follows: typically, to establish a causal connection between A and B, we have to have both evidence for a difference-making relation between A and B, and evidence for one or more mediators; but all this evidence is, ultimately, evidence for difference-making relations. In his (2011), Gillies offers a similar formulation for RWT. He suggests: "In order to establish that A causes B, observational statistical evidence does not suffice. Such evidence needs to be supplemented by interventional evidence, which can take the form of showing that there is a plausible mechanism linking A to B" [Gillies (2011), p. 116].[2]

VI. Conclusions

In this paper we have defended the view that mechanisms are underpinned by networks of difference-making relations and have shown that difference-making is more fundamental than production in understanding mechanistic causation. Our argument was two-fold. First, we have argued against the view that the productivity of mechanisms requires thinking of them as involving activities, qua a different ontic category. We have criticised two different routes to activities: Glennan's top-down metaphysics-first approach and Illari and Williamson's bottom-up science-first approach. Second, we have looked in some detail at the history of the case of scurvy prevention, in order to illustrate the difference-making account of mechanisms and to argue that mechanistic evidence in science is evidence about difference-making relations. The search for mechanisms is clearly a pervasive feature of science; but it is nothing else than the search for stable causal pathways.

Department of History and Philosophy of Science
University of Athens
University Campus, 15771 Athens, Greece
E-mail: psillos@phs.uoa.gr
E-mail: sioannidis@phs.uoa.gr

ACKNOWLEDGMENTS

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 1968.

NOTES

[1] This section is an expanded and reworked version of Psillos (2018).

[2] Hill's influential (1965) has been viewed as offering a version of RWT [cf. Russo & Williamson (2007); Clarke et al. (2014)]. Note, however, that he does not talk explicitly about mechanisms in his paper. He offers 'plausibility' as a criterion for establishing causal claims, which can be understood as the existence of a biologically plausible mechanism; but he does not regard it as particularly important, since "[w]hat is biologically plausible depends upon the biological knowledge of the day" [Hill (1965), p. 298]. As 'strongest support' for causation he takes experimental evidence, e.g. whether some preventive action does in fact prevent the appearance of a disease. Lastly, his 'Coherence' criterion involves, among others, establishing a mediator; his example is "histopathological evidence from the bronchial epithelium of smokers and the isolation from cigarette smoke of factors carcinogenic for the skin of laboratory animals" [Ibid.], which was important in establishing a causal connection between smoking and lung cancer.

REFERENCES

BARON, R. M. and KENNY, D. A. (1986), 'The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations; *Journal of Personality and Social Psychology*, vol. 51, pp. 1173-1182.

BARTHOLOMEW M. (2002), 'James Lind's Treatise of the Scurvy (1753)'; *Postgrad Med J*, vol. 78, pp. 695-6.

BECHTEL, W. and R.C. Richardson (2010) [1993], *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*; 2nd edn.; Cambridge, MA, MIT Press/Bradford Books.

BROWN, S. R. (2003), *Scurvy How a Surgeon, a Mariner, and a Gentleman Solved the Greatest Medical Mystery of the Age of Sail*; Chichester, Summersdale Publishers Ltd.

CARPENTER, K. J. (2012), 'The Discovery of Vitamin C'; *Annals of Nutrition and Metabolism*, vol. 61, pp. 259–264.

CARTWRIGHT, N. D. (1989), *Nature's Capacities and their Measurement*; Clarendon Press, Oxford.

CLARKE, B., GILLIES, D., ILLARI, P., RUSSO, F. and WILLIAMSON, J. (2014), 'Mechanisms and the Evidence Hierarchy'; *Topoi*, vol. 33, pp. 339-360.

ELLIS, B. (2001), *Scientific Essentialism*, Cambridge; Cambridge University Press.

FUNK, C. (1912), 'The Etiology of the Deficiency Diseases"; *The Journal of State Medicine*, vol. 20, pp. 341-368.

GILLETT, C. (2006), 'The Metaphysics of Mechanisms and the Challenge of the New Reductionism'; in Schouten, M. & de Jong, H. L., (eds.), *The Matter of the Mind*, Oxford, Blackwell, pp. 76-100.

GILLIES, D. (2011), 'The Russo-Williamson Thesis and the Question of Whether Smoking Causes Heart Disease'; in Illari, P., Russo, F., and Williamson, J., (eds.), *Causality in the Sciences*, Oxford, Oxford University Press, pp. 110–125.

— (2017), 'Mechanisms in Medicine'; *Axiomathes* vol. 27, pp. 621-34.

GLENNAN S. (2017), *The New Mechanical Philosophy*; Oxford, Oxford University Press.

HILL, B. (1965), 'The environment of Disease: Association or Causation?'; *Proceedings of the Royal Society of Medicine*, vol. 58, pp. 295–300.

HUGHES, R. E. (1983), 'From Ignose to Hexuronic acid to Vitamin C'; *Trends in Biochemical Sciences*, vol. 8, pp. 146-7.

ILLARI, P. M. and WILLIAMSON, J. (2013), 'In Defense of Activities'; *Journal for General Philosophy of Science*, vol. 44, pp. 69–83.

ILLARI, P. K. and WILLIAMSON, J. (2011), 'Mechanisms Are Real and Local'; in Illari, P., Russo, F., and Williamson, J., (eds.), *Causality in the Sciences*; Oxford, Oxford University Press, pp. 818 - 844.

IOANNIDIS, S. and PSILLOS, S. (2017), 'In Defense of Methodological Mechanism: The Case of Apoptosis'; *Axiomathes*, vol. 27, pp. 601-619.

—. (2018), 'Mechanisms in Practice: A Methodological Approach'; *Journal of Evaluation of Clinical Practice*, vol. 24, pp. 1177–1183.

— (2018), 'Mechanisms, Counterfactuals and Laws'; in Glennan, S. and Illari, P. M. (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, New York, Routledge, pp. 144–156.

LAKHANI S., DILLY, S. and FINLAYSON, C. (2009), *Basic Pathology: An Introduction to the Mechanisms of Disease*, 4th edn.; London, Hodder Arnold.

LIND J. (1753), *A Treatise of the Scurvy, in Three Parts, Containing an Inquiry into the Nature, Causes, and Cure of That Disease. Together with a Critical and Chronological View of What Has Been Published on the Subject*; Edinburgh, Sands, Murray and Cochran.

LINSTER, C. L. and VAN SCHAFTINGEN, E. (2007), "Vitamin C Biosynthesis, Recycling and Degradation in Mammals'; *FEBS Journal*, vol. 274, pp.1–22.

MACHAMER, P., DARDEN, L. and CRAVER, C. (2000), 'Thinking About Mechanisms'; *Philosophy of Science*, vol. 67, pp. 1–25.

MAGIORKINIS, E. BELOUKAS, and A. DIAMANTIS (2011), 'Scurvy: Past, Present and Future'; *European Journal of Internal Medicine*, vol. 22, pp. 147–152.

MENZIES, P. (2012), 'The Causal Structure of Mechanisms'; *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, pp. 796–805.

NELSON, D. L., LEHNINGER, A. L., and COX, M. M. (2008), *Lehninger Principles of Biochemistry*, 5th edn.; New York, W. H. Freeman.

PEARL, J. and MACKENZIE, D. (2018), *The Book of Why the New Science of Cause and Effect*; New York, Basic Books.

PSILLOS, S. (2004), 'A Glimpse of the Secret Connexion: Harmonising Mechanisms with Counterfactuals'; *Perspectives on Science* 12, pp. 288-319.

— (2018), 'Review of Glennan's *The New Mechanical Philosophy*'; *Australasian Journal of Philosophy*. DOI: 10.1080/00048402.2018.1526197.

RUSSO, F. & WILLIAMSON, J. (2007), 'Interpreting Causality in the Health Sciences'; *International Studies in the Philosophy of Science*, vol. 21, pp. 157–170.

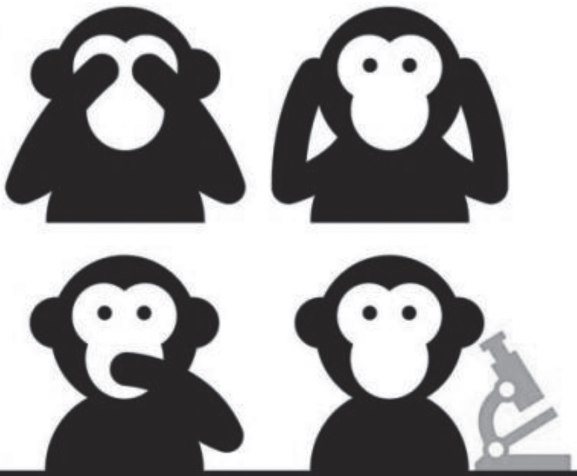
SCHULTZ J. (2002), 'American Chemical Society National Historic Chemical Landmarks. The Discovery of Vitamin C by Albert Szent-Gyögyi'; *http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/szentgyorgyi.html* (accessed February 11, 2019).

WILLIAMSON, J. (2011), 'Mechanistic Theories of Causality Part II'; *Philosophy Compass*, vol. 6, pp. 433–444.

WILLIAMSON, J. and WILDE, M. (2016), 'Evidence and Epistemic Causality'; in Wiedermann W., & von Eye, A. (eds.), *Statistics and Causality: Methods for Applied Empirical Research*, Wiley, pp 31-41.

WOODWARD, J. (2002), 'What is a Mechanism? A Counterfactual Account', *Philosophy of Science*, vol. 69, pp. S366–S377.

Lee McIntyre Author of Post-Truth



The Scientific Attitude

Defending Science from Denial, Fraud, and Pseudoscience

teorema
Vol. XXXVIII/3, 2019, pp. 77-94
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 77-94]

The Search for Generality in the Notion of Mechanism

Saúl Pérez-González

RESUMEN

En este artículo, introduzco y analizo un principio general compartido por los nuevos mecanicistas: *la búsqueda de generalidad*. Los nuevos mecanicistas consideran que una noción de mecanismo aceptable ha de ser adecuada para la mayoría de las áreas científicas en que los mecanismos son relevantes. El desarrollo de nociones de mecanismo generales se lleva a cabo mediante dos estrategias diferentes y alternativas, a las cuales denomino *la estrategia de extrapolación* y *la estrategia a-través-de-las-ciencias*. Después de analizar ejemplos paradigmáticos de éstas, planteo que ambas estrategias tienen problemas significativos y que las posibilidades de superarlos son escasas. Se concluye que sería recomendable abandonar la búsqueda de generalidad.

PALABRAS CLAVE*: mecanismo, explicación científica, generalidad, mecanismo de transmisión monetaria, selección natural.*

ABSTRACT

In this paper, I introduce and discuss a general principle shared by new mechanists: *the search for generality*. New mechanists agree that an appropriate notion of mechanism has to be suitable for most of the fields of science where mechanisms are relevant. The development of general notions of mechanism is pursued with two different and alternative strategies, which I call *the extrapolation strategy* and *the across-the-sciences strategy*. After analysing paradigmatic examples of them, I argue that both strategies face outstanding difficulties and that the prospects for overcoming them are dim. It is concluded that it would be advisable to abandon the search for generality.

KEYWORDS: *Mechanism, Scientific Explanation, Generality, Monetary Transmission Mechanism, Natural Selection.*

I. INTRODUCTION

The new mechanism emerged in the mid-90s.[1] *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (1993) by William Bechtel and Robert C. Richardson marked the beginning of this approach. Although it became much more influent some years later with

the publication of "Mechanisms and the Nature of Causation" (1996) by Stuart Glennan and "Thinking about Mechanisms" (2000) by Peter Machamer, Lindley Darden, and Carl F. Craver (henceforth MDC). The new mechanism is both a philosophy of science (i.e. philosophical inquiry into science) and a philosophy of nature (i.e. philosophical inquiry into the constituents of real things). Not only is it concerned about the role of mechanisms in science, but also about the nature of mechanisms, which are part of the real world.

The aim of this paper is to discuss a general principle of the new mechanism that I call *the search for generality*. Its structure is as follows. Section II introduces the main features of the new mechanism. Section III characterizes the search for generality and the two strategies that are adopted in order to achieve that purpose. Section IV argues that both strategies for achieving generality face outstanding difficulties. Section V shows that the problems of the search for generality undermine some arguments in support of the mechanistic approach (e.g. the mechanistic account of explanation). Finally, section VI concludes.

II. The New Mechanism

Within the framework of the new mechanism, several proposals have been raised. The most relevant ones are those of Glennan (1996), (2002), (2017), MDC (2000), Bechtel and Abrahamsen (2005), (2010), and Illari and Williamson (2012). In spite of the disagreements among those proposals, some general ideas are shared by all of them. Recent books such as *The New Mechanical Philosophy* (2017) by Glennan and *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (2018) edited by Glennan and Illari have underlined the great deal of consensus existing within the new mechanism.

New mechanists consider that a mechanism is an organized compound that is part of the real world. They discriminate between a mechanism, which is a real entity, and a model of it, which is often a piece of scientific reasoning. In this sense, Glennan says that "mechanisms and their constituents are things in the world that exist independently of the models we made of them" [Glennan (2017), p. 10]. They also agree that mechanisms are nested and form a hierarchy [Machamer, Darden, and Craver (2000), p. 13]. A component part of a mechanism is often a mechanism itself. For instance, a heart is both a mechanism and a component part of a mechanism (e.g. a circulatory system). Nevertheless, this idea does not lead them to reductionism regarding mechanisms. They reject that it is possible

to reduce higher level mechanisms to lower level mechanisms [Andersen (2014), p. 281]. Another shared idea is that a mechanism is always a mechanism for some phenomenon [Bechtel and Abrahamsen (2005), p. 42; Craver (2007), p. 123; Glennan (1996), p. 52]. The identification and delimitation of a mechanism (i.e. the fixation of a mechanism's boundaries) depend on the phenomenon for which it is responsible [Kaiser (2018)]. In the new mechanism, the notion of mechanism is not equivalent to the notion of machine. Although human-built machines (e.g. a vending machine) can often be considered mechanisms, most mechanisms are not machines.

There are also some general agreements regarding the principles that guide new mechanists' research. All their proposals emerge form a focus on scientific practice [Glennan (2017), p. 12]. Scientists' considerations about mechanisms are the main reference for the development of new mechanists' notions of mechanism. Another shared trait is the interest in how the discovery and decomposition (i.e. the identification of components and their organization) of mechanisms works [Bechtel and Richardson (1993); Darden (2018)]. They are not only interested in the role of mechanisms in science, but also in how scientists discover and decompose them. Due to the fact that a mechanism is always a mechanism for some phenomenon, the discovery of a mechanism begins with the identification of a puzzling phenomenon.

According to the new mechanism, the role of mechanisms in science is usually associated with the scientific objective of explaining. New mechanists have developed a mechanistic account of scientific explanation. They consider that a phenomenon is explained by means of specifying the mechanism that is responsible for it.[2] In this sense, MDC say: "To give a description of a mechanism for a phenomenon is to explain that phenomenon" [Machamer, Darden, and Craver (2000), p. 3]. A well-known example of mechanistic explanation is the standard explanation of the phenomenon of chemical transmission at synapsis [Machamer, Darden, and Craver (2000)]. This phenomenon is explained by the interactions (e.g. transporting, inserting, diffusing…) among cell membrane, vesicles, microtubules, molecules, and ions that are responsible for it. The mechanistic approach to scientific explanation has been developed as an alternative to the covering-law model [Bechtel and Abrahamsen (2005); Craver (2014)].[3] The covering-law model, which was developed by Carl G. Hempel (1965), is based on the idea that to explain a phenomenon is to subsume it under a law. This proposal gave rise to a consensus regarding the notion of scientific explanation that lasted from the late 1940s to the mid-1960s [Salmon (1989), p. 3]. However, since

the early 1960s, several critiques and counterexamples have noted that subsuming a phenomenon under a law is neither necessary nor sufficient condition for explaining it. The mechanistic approach, as other current approaches (e.g. unificationist account of explanation, pragmatic theories of explanation…), try to account for scientific explanation avoiding covering-law model's problems.

New mechanists consider that mechanisms are relevant in science and that their relevance is manly associated with explaining. Thus, they support a mechanistic account of scientific explanation. According to it, a phenomenon is explained by means of specifying the mechanism that is responsible for it. In what follows, I will discuss one problematic aspect of the new mechanism related to the notion of mechanism, i.e., the search for generality.

III. The Search for Generality

Through the previous section, several well-known general principles of the new mechanism have been noted. Nevertheless, there is another one that has not been previously identified and deserves attention. It is *the search for generality*. New mechanists consider that mechanisms are relevant in most scientific fields. And they agree that an appropriate notion of mechanism has to be suitable for most of the fields where mechanisms are relevant. In this sense, in his foundational "Mechanisms and the Nature of Causation", Glennan [(1996), p. 50] claims that mechanisms are relevant in all scientific fields except fundamental physics. His proposal aims to suit those fields.

Despite the fact that the search for generality is a trait shared by all mechanistic proposals, it is not always pursued with the same strategy. This difference with respect to the strategies has made difficult to identify this common feature. Within the new mechanism, there are two strategies for proposing general notions of mechanism. I call them *the extrapolation strategy* and *the across-the-sciences strategy*.

The extrapolation strategy consists of developing a notion of mechanism taking one or a few fields of science as reference, and then applying that notion to many other fields. This strategy goes from certain kind of mechanisms to a general notion of mechanism, so that, the notion is applied to kinds of mechanisms that were not taken into account for its development. MDC (2000), for instance, follow the extrapolation strategy. Their notion of mechanism was developed taking neurobiological and molecular mechanisms as reference. The mechanisms of chemi-

cal transmission at synapsis and protein synthesis are their paradigmatic examples of mechanisms. But they consider that their notion of mechanism could be applied to other fields of science. In this sense, they say: "We suspect that this analysis is applicable to many other sciences, and maybe even to cognitive or social mechanisms" [Machamer, Darden, and Craver (2000), p. 2].[4] The extrapolation strategy was also adopted by Glennan (1996), (2002). He developed his notion of mechanism taking physical mechanisms (e.g. a float valve, a voltage switch…) as reference. Nevertheless, he considered that it suited many other kinds of mechanisms: "my analysis is in no way limited to mechanisms that are physical in nature. It is meant to equally apply to chemical, biological, psychological and other higher-level mechanisms" [Glennan (1996), p. 61].

James Woodward (2002) also follows the extrapolation strategy in his study of mechanisms.[5] He focuses on mechanics (e.g. a block sliding down an inclined plane), although he takes into account molecular biology too. However, he claims that "a notion of mechanism very similar to that characterized by **MECH** is employed in many other fields of science — for example, in psychology" [Woodward (2002), p. S376].

The across-the-sciences strategy consists of thinking about mechanisms across all the sciences and developing a notion of mechanism that includes their common features. This strategy goes from all mechanisms to a general notion of mechanism. All kinds of mechanisms to which the notion is applied are taken into account for its development. The across-the-sciences strategy was introduced by Illari and Williamson (2012), who underlined its difference from the previous developments of notions of mechanism (i.e. the extrapolation strategy). Their aim is to consider mechanisms in general and to propose "a characterization that gives an understanding of what is common to mechanisms in *all* fields" [Illari and Williamson (2012), p. 120]. They underline the need of a consensus account of mechanisms in order to address several philosophical issues (e.g. causal explanation, inference, and modelling). The across-the-sciences strategy has recently been adopted by Glennan (2017). He has abandoned his previous notion of mechanism [see Glennan (1996), (2002)] and proposes a minimal characterization of it, which tries to include what all mechanisms share in common. His aim is to develop a notion of mechanism "broad enough to capture most of wide range of things scientists have called mechanisms" [Glennan (2017), p. 18]. In the next section, I will show that both the extrapolation strategy and the across-the-sciences strategy face outstanding difficulties.

IV. The Difficulties of the Search for Generality

Generality is a valuable purpose here. There are several kinds of mechanisms in science (e.g. molecular mechanisms, social mechanisms, computing mechanisms…). A notion that could account for all of them would be useful for both scientific research and philosophical understanding. It would facilitate collaboration among fields of science where mechanisms are relevant. Besides, the will of generality is also present in many philosophical issues related with mechanisms. For instance, philosophers of science try to develop a notion of causal explanation that suits all explanations where the *explanans* makes reference to the causes of the *explanandum* phenomenon. A general notion of mechanism would help to address these issues. However, I will argue below that both suggested strategies for pursuing generality face outstanding difficulties. In order to identify and analyse those difficulties, I will focus on MDC's and Illari and Williamson's proposals, which are paradigmatic examples of the extrapolation strategy and the across-the-sciences strategy respectively.

IV.1. *The Difficulties of the Extrapolation Strategy*

MDC's proposal, one of the most relevant ones in the current debate, follows the extrapolation strategy for developing a general notion of mechanism. However, their notion of mechanism, which is developed taking certain kind of mechanisms as reference, does not suit many other kinds of mechanisms. MDC define mechanism as follows:

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions [Machamer, Darden, and Craver (2000), p. 3].

A mechanism is an organized collection of entities (with their properties) and activities. Entities are things that engage in activities. They are usually spatiotemporal located, structured, and oriented. Examples of entities are neurotransmitters, neurons, DNA bases… Activities are productive happenings. They have temporal order, rate, and duration. Examples of activities are transporting, neuromodulating, recycling… Mechanisms' components are organized. Their organization has temporal, spatial, and active aspects [Craver and Darden (2013), p. 20]. MDC hold that mechanisms are regular and "work always or for the most part in the same way under the same conditions" [Machamer, Darden, and Craver (2000), p. 3].

Although MDC consider that their notion of mechanism could be applied to most fields of science (see section III), it is unlikely the case. For

instance, the application of MDC's notion of mechanism to evolutionary biology would be very problematic. Evolutionary biologists often refer to evolutionary causes (e.g. natural selection, mutation, migration…) as mechanisms that bring about changes in populations. In this sense, Graham Bell says: "Selection is an effective mechanism for producing adaptation" [Bell (2008), p. 499]. Other evolutionary biologists who refer to several evolutionary causes as mechanisms are Jon C. Herron and Scott Freeman (2014). However, MDC's notion of mechanism is not able to account for evolutionary causes as mechanisms. Robert A. Skipper and Roberta L. Millstein (2005) have argued that natural selection does not meet MDC's characterization of mechanisms.[6] For instance, relevant productive relationships among component entities of natural selection cannot always be understood as activities. Natural selection often depends on passive selection processes (e.g. being poisonous, having certain colour…) that can hardly be considered activities. Skipper and Millstein also say that natural selection does not satisfy the requirement of regularity.[7]

MDC's notion of mechanism is also unable to account for economic mechanisms. Economists often refer to economic mechanisms (e.g. markets, price mechanisms…). Well-known examples of economic mechanisms are monetary transmission mechanisms. A monetary transmission mechanism is a mechanism responsible for the influence of a central bank in output, employment, prices, and inflation of a country or a political and economic union (e.g. European Union) [Samuelson and Nordhaus (2010), p. 484]. It is an organized collection of entities (e.g. banks, central banks, securities broker-dealers…) and activities (e.g. buying and selling government securities, trading reserve balances at a central bank, changing the legal reserve-ratio requirements…). It could seem that MDC's notion of mechanism suits monetary transmission mechanisms, but it is unlikely the case.

MDC consider that mechanisms are regular. Lane DesAutels (2016) has recently showed that in order to meet MDC's requirement of regularity a mechanism has to be process regular and not be affected by internal sources of irregularity. However, monetary transmission mechanisms are not process regular and are affected by internal sources of irregularity. Process regularity consists in that "the constituent entities and activities of a mechanism behave in roughly the same way each time the mechanism operates" [DesAutels (2016), p. 16]. But, component entities of monetary transmission mechanisms do not always behave in the same way. For instance, given an undesirably low level of inflation, a central

bank (e.g. European Central Bank, U.S. Federal Reserve System…) does not always behave in the same way in order to influence in it. Central banks often buy government securities for increasing the level of inflation. But sometimes they also modify the reverse-ratio requirements or borrow money with a discount rate. Monetary transmission mechanisms are affected by internal sources of irregularity too. For example, a change in the behaviour of the U.S. Federal Reserve System can be the result of a change in which presidents of regional Federal Reserve Banks are voting members of the Federal Open Market Committee.

Other aspect of MDC's proposal that does not suit monetary transmission mechanisms is the fixation of mechanisms' boundaries. It is considered that "mechanisms are always mechanisms *of* a given phenomenon" [Craver (2007), p. 123]. Regarding boundaries, Craver says: "The boundaries of mechanisms —what is in the mechanism and what is not— are fixed by reference to the phenomenon that the mechanism explains" [Craver (2007), p. 123]. A mechanism *of* certain phenomenon is composed of those entities, activities, and organizational features that are part of the system whose behaviour is the phenomenon of interest and are relevant for that phenomenon. A part is relevant for a phenomenon if it meets the requirement of mutual manipulability [Craver (2007)]. Therefore, a part X is a component of the mechanism *of* phenomenon Y if some interventions on X bring about changes in Y, and vice versa. Craver appeals to the notion of intervention developed by Woodward (2003). Woodward claims that "an intervention on some variable X with respect to some second variable Y is a causal process that changes the value of X in an appropriately exogenous way, so that if a change in the value of Y occurs, it occurs only in virtue of the change in the value of X" [Woodward (2003), p. 94]. Nevertheless, this proposal does not suit monetary transmission mechanisms. A monetary transmission mechanism is a mechanism of a phenomenon (i.e. the influence of a central bank in output, employment, prices, and inflation). But it is not composed of all entities, activities, and organizational features that are part of the system whose behaviour is the phenomenon of interest and are relevant for that phenomenon. Consider the South Korean monetary transmission mechanism. That mechanism is responsible for the influence of the Bank of Korea in the South Korean output, employment, prices, and inflation. Samsung Electronics, which is the largest South Korean firm, is part of the system whose behaviour is the phenomenon of interest (South Korea). It also meets the requirement of mutual manipulability and is relevant for the phenomenon. Some interventions on Samsung Electronics produce changes in the influence of the Bank of Korea in the South Kore-

an economy, and vice versa. However, it is not part of the South Korean monetary transmission mechanism. In sum, MDC's proposal does not properly fix monetary transmission mechanisms' boundaries.[8]

MDC follow the extrapolation strategy for developing a general notion of mechanism. However, as it has been argued, their notion does not suit many kinds of mechanisms. It introduces certain requirements that are not met by those kinds of mechanisms. This also seems to be the case for the other proposals that follow the extrapolation strategy. For instance, Glennan's (2002) and Woodward's (2002) proposals do not suit neither economic mechanisms nor evolutionary mechanisms. They consider that properties of mechanisms' parts must remain stable in absence of interventions. However, properties of economic mechanism's parts (e.g. firms) and evolutionary mechanisms' parts (e.g. populations) may change even if no intervention has been done [Skipper and Millstein (2005)]. Consider the following hypothetical example of a firm. During a lunch at the office, there is a strong discussion between the CEO of firm X and the director of its department of publicity. As a consequence of this event, the CEO decreases the budget of the department of publicity. Due to the reduction of the budget, the department of publicity has to introduce changes in the advertising strategy of the firm (e.g. the number of ads on TV is decreased, while the number of ads on radio is increased). In this example, different properties of firm X (e.g. budget of its departments, number of ads on TV…) changed without any exogenous intervention.

IV.2. *The Difficulties of the Across-the-Sciences Strategy*

Illari and Williamson, who introduced the across-the-sciences strategy, follow it for developing a general notion of mechanism. Nevertheless, on my view they propose a vacuous and overly broad notion of mechanism. Illari and Williamson offer the following definition of mechanism:

> A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon. [Illari and Williamson (2012), p. 120]

A mechanism is an organized collection of entities and activities. No restrictions on regularity, internal structure, size, boundaries or robustness are imposed on them. Examples of entities are electrons, stars, black holes, x-rays… While examples of activities are colliding, relaxing, collapsing, radiating… Mechanisms' organization is merely defined as "whatever relations between the entities and activities discovered pro-

duce the phenomenon of interest" [Illari and Williamson (2012), p. 128]. There are several possible forms of organization (e.g. spatial, temporal, equilibrium, self-organization, feedback…). What forms of organization are relevant in a particular mechanism is an empirical question.

My main objection here is that Illari and Williamson's notion of mechanism is vacuous.[9] Their definition of mechanism relies on the concepts of entity, activity, organization, and being responsible for. Nevertheless, they do not properly characterize those concepts. For instance, consider the concept of entity. They offer neither a definition of entity nor a set of necessary conditions to be an entity nor a set of sufficient conditions to be an entity. Besides, they refuse to introduce restrictions on entities. They certainly present some examples of component entities of astrophysical and molecular mechanisms. Nevertheless, those examples are not numerous and diverse enough to properly characterize entities across the sciences. It could be argued that the concept of entity is too general, and a proper characterization of entities is not possible. However, other authors have already raised more concrete characterizations of component entities [see Machamer, Darden, and Craver (2000)]. Illari and Williamson's notion of mechanism is not precise enough. It is not clear what features would characterize mechanisms. They increase the scope of their notion of mechanism (i.e. the number of mechanisms that it subsumes) at the cost of decreasing its precision.

Another problem for Illari and Williamson's notion of mechanism is that it is overly broad. Although they consider that their characterization "is not so broad that it captures non-mechanisms" [Illari and Williamson (2012), p. 129], it subsumes things that could hardly be accepted as mechanisms. For instance, their notion would admit a group of cows grazing in a field as a mechanism. It is an organized collection of entities (e.g. cows) and activities (e.g. grazing) that are responsible for a phenomenon (i.e. the removal of the grass of the field). However, it could hardly be considered a proper mechanism. It is actually a mere aggregate, whose components are not actively organized [Craver and Darden (2013), p. 20]. Cows do not interact and make a difference to each other in order to remove the grass. Other examples of things that Illari and Williamson's notion would wrongly admit as mechanisms are a traffic jam and a group of babies napping.

The aim of Illari and Williamson is to develop a wide notion of mechanism that could encompass mechanisms of all fields. As it has been noted, they decrease the precision of their characterization in order to increase its scope. Nevertheless, Illari and Williamson's notion of mechanism does not suit many kinds of mechanisms. For instance, it

would not fit evolutionary mechanisms. Illari and Williamson, as MDC, consider that mechanisms are collections of entities and activities. Hence, their notion of mechanism cannot account for those cases of evolution in which natural selection depends on passive selection processes that can hardly be considered activities. Their notion of mechanism would also be unable to account for economic mechanisms. Like MDC, they consider that a mechanism for a phenomenon is composed of those parts (e.g. entities, activities…) that are relevant for it. In this sense, they claim: "mechanisms are functionally individuated by their phenomena" [Illari and Williamson (2012), pp. 123-124]. But, as it has been argued, this idea does not suit economic mechanisms' (e.g. monetary transmission mechanisms) boundaries. In conclusion, despite of the fact that the across-the-sciences strategy is developed as an alternative to the extrapolation strategy, Illari and Williamson do not avoid the problem of the latter (see subsection IV.1).

Illari and Williamson follow the across-the-sciences strategy for developing a general notion of mechanism. Nevertheless, they do not satisfactorily achieve that purpose. The main problems of their notion of mechanism are that it is vacuous and overly broad. The other proposals that follow the across-the-sciences strategy seem to face the same kind of problems. For example, Glennan's (2017) recent proposal is overly broad too. It subsumes things that could hardly be accepted as mechanisms (e.g. a group of cows grazing, a traffic jam…). In addition, Glennan's notion is also unable to account for many kinds of mechanisms. As Illari and Williamson, he considers that mechanisms are collections of entities and activities, and that a mechanism for a phenomenon is composed of those parts that are relevant for it. Therefore, it does not suit evolutionary mechanisms and economic mechanisms either.

V. The Search for Generality and the Arguments in Support of the Mechanistic Approach

Mechanisms are relevant in several fields of science. In most of those fields (e.g. neuroscience, cognitive science, molecular biology…), a mechanistic stance has been adopted. From this fact, new mechanists have raised an argument in support of the mechanistic approach. They argue that the adoption of the mechanistic approach in a field of science would improve its relationship with the numerous fields in which this approach has already been adopted. Mechanisms would be the subject of

study of all of them. The fields would only differ in which kind of mechanisms they study. Different fields of science would just study different parts of the hierarchy of mechanisms. In this sense, Craver and Alexandrova say that one reason why neuroeconomics should be a mechanistic science is that "the rest of neuroscience, cognitive science, and biology have adopted a largely mechanistic stance […] The search for mechanisms provides a common goal toward which researchers in different fields can contribute" [Craver and Alexandrova (2008), p. 398].

A related argument can be raised in favour of the mechanistic account of scientific explanation [Hedström and Ylikoski (2011)]. As it has been said (see section II), supporting a mechanistic account of scientific explanation is a trait of the new mechanism. New mechanists consider that a phenomenon is explained by means of specifying the mechanism that is responsible for it. The explanation of a phenomenon is often presented by means of a mechanistic model. A mechanistic model has two components: phenomenal description and mechanistic description [Glennan (2017), p. 66]. The phenomenal description is a model of the phenomenon and the mechanistic description is a model of the mechanism responsible for it. In a mechanistic model, the phenomenal description is (or represents) the *explanandum* and the mechanistic description is (or represents) the *explanans* [Glennan (2005), p. 448]. Due to the fact that mechanisms are relevant in several fields, new mechanists argue that a mechanistic account of scientific explanation could be adopted in most fields of science. Hence, it would be an all-encompassing account of scientific explanation. This can be presented as an argument in favour of it.

As it was pointed at the beginning of this paper, the mechanistic account of scientific explanation has been developed as an alternative to the covering-law model. A well-known problem of the covering-law model is that there are fields of science where only a few laws are known, such as evolutionary biology [see Beatty (1995); Scriven (1959)]. A nomological account of explanation can hardly be defended for those fields. Nevertheless, several mechanisms are often known in those fields where laws are not available. A mechanistic account of explanation could be adopted for them. The broad applicability of the mechanistic account of scientific explanation would be an argument to prefer it rather than other options.

Both previously presented arguments rely on the same assumption. They assume that the same notion of mechanism exists in all the fields where mechanisms are relevant. It is considered that the adoption of the mechanistic approach in various fields would improve the relationships among them because their subjects of study would be very similar. But

they would be similar only if mechanisms are understood in a similar way in all fields. If those fields understood mechanisms in a very different way, the adoption of the mechanistic approach would not imply similar subjects of study. Likewise, it is considered that the mechanistic account of scientific explanation could offer a unified account of explanation because in several fields of science phenomena could be explained by means of referring to mechanisms. But it would be a unified standpoint only if mechanisms are similarly understood in all fields. If those fields understood mechanisms in a very different way, the mechanistic account of scientific explanation would not be unifier. Although several fields could refer to mechanisms, they would not refer to the same kind of things.

The assumption on which both arguments rely is challenged by the difficulties faced by the search for generality. It is considered that the same notion of mechanism is shared by all fields of science where mechanisms are relevant. The search for generality constitutes the real attempt of identifying that shared notion. New mechanists try to propose a notion of mechanism that suits most of the fields where mechanisms are relevant. Nevertheless, as it has been argued, both strategies for pursuing generality (i.e. the extrapolation strategy and the across-the-sciences strategy) face outstanding difficulties. This means that the assumption that the same notion of mechanism is shared by all fields is not justified and requires additional support. Not only is that shared notion unknown, but also several attempts of achieving it have failed. Therefore, both previously presented arguments in support of the mechanistic approach are not acceptable in their current form. They should not be used for endorsing the mechanistic approach.

VI. Conclusion

The search for generality is a general principle of the new mechanism. New mechanists consider that an appropriate notion of mechanism must be suitable for most of the fields of science where mechanisms are relevant. In order to propose a general notion of mechanism, two strategies have been adopted: the extrapolation strategy and the across-the-sciences strategy. As it has been argued, both of them face outstanding difficulties. The problems of the search for generality undermine some arguments in support of the mechanistic approach, which rely on the assumption that the same notion of mechanism exists in all fields where mechanisms are relevant.

It seems that current notions of mechanism are unable to properly account for all kinds of mechanisms. Moreover, it is doubtful that a general notion of mechanism could be developed. As Petri Ylikoski argues, "[t]he entities and processes studied by different sciences are quite heterogeneous, and it is probably impossible to propose a mechanism definition that would be both informative and cover all the prominent examples of mechanisms" [Ylikoski (2012), p. 22]. Giving this scenario, it would be advisable to abandon the search for generality. In each field of science, a notion of mechanism must be developed taking the activity of the scientists of that field as the main reference. How mechanisms are understood in other fields should not heavily influence in that development. It does not mean that philosophers of science must not think about similarities among mechanisms across the sciences. It could be very useful, after the development of the field-specific notions of mechanism, to compare the different notions of mechanism and identify their common features. Nevertheless, the output of that comparison would not be a proper notion of mechanism in general. In the same way that a list of the common features of English laboratories would not be a general definition of English laboratory. It would just be a list of traits that are shared by mechanisms across the sciences. It is an open question if, in spite of not being a definition, some of those shared traits may be necessary or sufficient conditions to be a mechanism. But that question exceeds the scope of this paper.

Departament de Filosofia
Universitat de València
Av. Blasco Ibáñez 30, 46010 Valencia, Spain
E-mail: saul.perez@uv.es

ACKNOWLEDGEMENTS

I would like to thank Marc Artiga, Raffaella Campaner, and Valeriano Iranzo for their valuable comments. This work was supported by the Spanish Ministry of Science, Innovation and Universities under grants FPU16/03274 and FFI2017-89639-P.

NOTES

[1] Several classifications of the proposals raised within the new mechanical (or mechanistic) philosophy have been proposed [see Kuorikoski (2009); Reiss (2008)]. A particularly useful classification has been recently offered by Stuart

Glennan and Phyllis Illari (2018). They have distinguished two trends within the new mechanical philosophy: the new mechanism and the social scientific mechanism. In this paper, I will focus on what they call the new mechanism. It involves the senses of the term 'mechanism' that Holly Andersen (2014) has named as mechanism$_1$ and mechanism$_2$. The main ideas of the social scientific mechanism can be found in the works of Jon Elster (1989), (1999) and Peter Hedström (2005).

[2] There is a disagreement among new mechanists about whether mechanistic explanations are ontic (i.e. they explain because they fit the *explanandum* phenomenon into the causal structure of the world) or epistemic (i.e. they explain because they successfully increase our understanding of the world) [see Illari (2013)]. Nevertheless, all of them agree that mechanistic explanations refer to the mechanism responsible for the *explanandum* phenomenon.

[3] Bert Leuridan (2010) has claimed that mechanistic accounts are not genuine alternatives to nomologic accounts. Taking the pragmatic account of laws by Sandra Mitchell (1997) as his starting point, he argues that mechanistic models epistemologically depend on laws and cannot replace them as a model of explanation in science. However, Andersen (2011) shows that mechanistic models are not dependent on laws, but on regularities, which are not synonymous with laws.

[4] Peter Hedström [(2005), p. 25] has developed the application of MDC's notion of mechanism to sociology.

[5] Although Woodward is not properly a new mechanist, his work has strongly influenced many new mechanists [see Craver (2007); Glennan (2002); Woodward (2002), (2011)].

[6] Since the publication of Skipper and Millstein's paper, there has been a debate about how natural selection could be understood as a mechanism [see Barros (2008); DesAutels (2016); Illari and Williamson (2010); Pérez-González and Luque (2019)].

[7] Lane DesAutels (2016) has recently argued that natural selection is only irregular in aspects that are not relevant in order to meet MDC's requirement (e.g. product regularity, regularity regarding external sources of irregularity…).

[8] For other critiques against the mutual manipulability account of constitutive relevance see Leuridan (2012).

[9] Rosenberg (2018) has underlined the strategic vagueness of some mechanistic proposals. He claims that mechanists are often cagey in order to avoid counterexamples against their proposals.

REFERENCES

ANDERSEN, H. K. (2011), 'Mechanisms, Laws, and Regularities', *Philosophy of Science*, vol. 78, pp. 325-331.

— (2014), 'A Field Guide to Mechanisms: Part I', *Philosophy Compass*, vol. 9, pp. 274-283.

BARROS, D. B. (2008), 'Natural Selection as a Mechanism', *Philosophy of Science*, vol. 75, pp. 306-322.

BEATTY, J. H. (1995), 'The Evolutionary Contingency Thesis', in Woltersand, G. and Lennox, J. G. (eds.), *Concepts, Theories, and Rationality in the Biological Sciences. The Second Pittsburgh-Konstanz Colloquium in the Philosophy of Science*, Cambridge, MIT Press, pp. 45-81.

BECHTEL, W. and ABRAHAMSEN, A. (2005), 'Explanation: A Mechanist Alternative', *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 36, pp. 421-441.

— (2010), 'Dynamic Mechanistic Explanation: Computational Modelling of Circadian Rhythms as an Exemplar for Cognitive Science', *Studies in History and Philosophy of Science*, vol. 41, pp. 321-333.

BECHTEL, W. and RICHARDSON, R. C. (1993), *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, Princeton, Princeton University Press.

BELL, G. (2008), *Selection. The Mechanism of Evolution*, Oxford, Oxford University Press.

CRAVER, C. F. (2007), *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, Oxford, Clarendon Press.

— (2014), 'The Ontic Account of Scientific Explanation', in Kaiser, M. I., Scholz, O. R., Plenge, D. and Hüttemann, A. (eds.), *Explanation in the Special Sciences: The Case of Biology and History*, Dordrecht, Springer, pp. 27-52.

CRAVER, C. F. and ALEXANDROVA, A. (2008), 'No Revolution Ncessary: Neural Mechanisms for Economics', *Economics and Philosophy*, vol. 24, pp. 381-406.

CRAVER, C. F. and DARDEN, L. (2013), *In Search of Mechanisms: Discoveries across the Life Sciences*, Chicago, University of Chicago Press.

DARDEN, L. (2018), 'Strategies for Discovering mechanisms', in Glennan, S. and Illari, P. M. (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, Abingdon, Routledge, pp. 255-266.

DESAUTELS, L. (2016), 'Natural Selection and Mechanistic Regularity', *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 57, pp. 13-23.

ELSTER, J. (1989), *Nuts and Bolts for Social Sciences*, Cambridge, Cambridge University Press.

— (1999), *Alchemies of the Mind. Rationality and the Emotions*, Cambridge, Cambridge University Press.

GLENNAN, S. (2002), 'Rethinking Mechanistic Explanation', *Philosophy of Science*, vol. 69, pp. S342-S353.

— (2005), 'Modeling Mechanisms', *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 36, pp. 443-464.

— (2017), *The New Mechanical Philosophy*, Oxford, Oxford University Press.

GLENNAN, S. and ILLARI, P. M. (2018), 'Introduction: Mechanisms and Mechanical Philosophies', in Glennan, S. and Illari, P. M. (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, Abingdon, Routledge, pp. 1-9.

GLENNAN, S. S. (1996), 'Mechanisms and the Nature of Causation', *Erkenntnis*, vol. 44, pp. 49-71.

HEDSTRÖM, P. (2005), *Dissecting the Social. On the Principles of Analytical Sociology*, New York, Cambridge University Press.

HEDSTRÖM, P. and YLIKOSKI P. (2011), 'Analytical Sociology', in Jarvie, I. C. and Zamora-Bonilla, J. (eds.), *The SAGE Handbook of The Philosophy of Social Sciences*, London, SAGE Publications, pp. 386-398.

HEMPEL, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York, The Free press.

HERRON, J. C. and FREEMAN, S. (2014), *Evolutionary Analysis*, Upper Saddle River, Pearson Education.

ILLARI, P. M. (2013), 'Mechanistic Explanation: Integrating the Ontic and Epistemic', *Erkenntnis*, vol. 78, pp. 237-255.

ILLARI, P. M. and WILLIAMSON, J. (2010), 'Function and Organization: Comparing the Mechanisms of Protein Synthesis and Natural Selection', *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 41, pp. 279-291.

— (2012), 'What is a Mechanism? Thinking about Mechanisms *across* the Sciences', *European Journal of Philosophy of Science*, vol. 2, pp. 119-135.

KAISER, M. I. (2018), 'The Components and Boundaries of Mechanisms', in Glennan, S. and Illari, P. M. (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, Abingdon, Routledge, pp. 116-130.

KUORIKOSKI, J. (2009), 'Two Concepts of Mechanism: Componential Causal System and Abstract Form of Interaction', *International Studies in the Philosophy of Science*, vol. 23, pp. 143-160.

LEURIDAN, B. (2010), 'Can Mechanisms Really Replace Laws of Nature?', *Philosophy of Science*, vol. 77, pp. 317-340.

— (2012), 'Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms', *The British Journal for the Philosophy of Science*, vol. 63, pp. 399-427.

MACHAMER, P., DARDEN L. and CRAVER, C. F. (2000), 'Thinking about Mechanisms', *Philosophy of Science*, vol. 67, pp. 1-25.

MITCHELL, S. D. (1997), 'Pragmatic Laws', *Philosophy of Science*, vol. 64, pp. S468-S479.

PÉREZ-GONZÁLEZ, S. and LUQUE, V. J. (2019), 'Evolutionary Causes as Mechanisms: A Critical Analysis', *History and Philosophy of the Life Sciences*, vol. 41, pp. 1-23.

REISS, J. (2008), *Error in Economics. Towards a More Evidence-Based Methodology*, Abingdon, Routledge.

ROSENBERG, A. (2018), 'Making Mechanism Interesting', *Synthese*, vol. 195, pp. 11-33.

SALMON, W. C. (1989), *Four Decades of Scientific Explanation*, Minneapolis, University of Minnesota Press.

SAMUELSON, P. A. and NORDHAUS, W. D. (2010), *Economics,* Boston, McGraw-Hill/Irwin.

SCRIVEN, M. (1959), 'Explanation and Prediction in Evolutionary Theory', *Science*, vol. 130, pp. 477-482.

SKIPPER, R. A. and MILLSTEIN, R. L. (2005), 'Thinking about Evolutionary Mechanisms: Natural Selection', *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 36, pp. 327-347.

WOODWARD, J. (2002), 'What is a Mechanism? A Counterfactual Account', *Philosophy of Science*, vol. 69, pp. S366-S377.

— (2003), *Making Things Happen,* New York, Oxford University Press.

— (2011), 'Mechanisms Revisited', *Synthese*, vol. 183, pp. 409-427.

YLIKOSKI, P. (2012), 'Micro, Macro, and Mechanisms', in Kincaid, H. (ed.), *The Oxford Handbook of Philosophy of Social Science*, New York, Oxford University Press, pp. 21-45.

teorema
Vol. XXXVIII/3, 2019, pp. 95-120
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 95-120]

Equilibrium Explanation as Structural Non-Mechanistic Explanations: The Case of Long-Term Bacterial Persistence in Human Hosts

Javier Suárez and Roger Deulofeu

RESUMEN

Philippe Huneman ha cuestionado recientemente los límites en la aplicación de los modelos mecanicistas de la explicación científica en base a la existencia de lo que denomina "explicaciones estructurales", en las que el fenómeno se explica en virtud de las propiedades matemáticas del sistema en que el fenómeno ocurre. Las explicaciones estructurales pueden darse en formas muy diversas: en virtud de la forma de *pajarita* (*bowtie*) de la estructura, de las propiedades topológicas del sistema, de los equilibrios alcanzados, etc. El papel que juegan las matemáticas en las explicaciones que apelan a la estructura de pajarita o a las propiedades topológicas del sistema ha sido recientemente examinado en varios trabajos. Sin embargo, el papel exacto que juegan las matemáticas en el caso de las explicaciones en términos de equilibrio aún no ha sido totalmente clarificado, y diferentes autores defienden interpretaciones contradictorias, algunas de las cuales las asemejarían más al modelo defendido por algunos filósofos mecanicistas que al modelo estructural de Huneman. En este trabajo, tratamos de cubrir ese déficit estudiando el papel que juegan las matemáticas en el modelo de equilibrio anidado (*nested equilibrium*) elaborado por Blaser y Kirchner para explicar la estabilidad de las asociaciones ontogenética y filogenéticamente persistentes entre humanos y microorganismos. De nuestro análisis se desprende que su modelo es explicativo porque i) se identifica una estructura matemática del sistema que viene dada por un conjunto de ecuaciones diferenciales que satisfacen una estrategia evolutivamente estable; ii) la estructura anidada del modelo hace que la estrategia evolutivamente estable sea robusta ante posibles perturbaciones; iii) esto es así porque las propiedades del sistema empírico son isomorfas a, pero no causalmente responsables de, las propiedades de la estrategia evolutivamente estable. La combinación de estas tres tesis hace que las explicaciones en términos de equilibrios se asemejen más al modelo estructural de explicación que al modelo mecanístico.

PALABRAS CLAVE: *explicación científica; mecanismos; explicación en términos de equilibrio; explicaciones estructurales; explicaciones no causales; estrategia evolutivamente estable*.

ABSTRACT

Philippe Huneman has recently questioned the widespread application of mechanistic models of scientific explanation based on the existence of structural explanations, i.e. explanations that account for the phenomenon to be explained in virtue of the mathematical properties of the system where the phenomenon obtains, rather than in terms of the mechanisms that causally produce the phenomenon. Structural explanations are very di-

verse, including cases like explanations in terms of bowtie structures, in terms of the topological properties of the system, or in terms of equilibrium. The role of mathematics in bowtie structured systems and in topologically constrained systems has recently been examined in different papers. However, the specific role that mathematical properties play in equilibrium explanations requires further examination, as different authors defend different interpretations, some of them closer to the new-mechanistic approach than to the structural model advocated by Huneman. In this paper, we cover this gap by investigating the explanatory role that mathematics play in Blaser and Kirschner's nested equilibrium model of the stability of persistent long-term human-microbe associations. We argue that their model is explanatory because: i) it provides a mathematical structure in the form of a set of differential equations that together satisfy an ESS; ii) that the nested nature of the ESSs makes the explanation of host-microbe persistent associations robust to any perturbation; iii) that this is so because the properties of the ESS directly mirror the properties of the biological system in a non-causal way. The combination of these three theses make equilibrium explanations look more similar to structural explanations than to causal-mechanistic explanation.

KEYWORDS: *Scientific Explanation; Mechanisms; Equilibrium Explanations; Structural Explanations; Non-Causal Explanations; Evolutionarily Stable strategy*.

In the last few years, a new trend in the debates about scientific explanation has flourished in philosophy of science. This new trend, "new-mechanism," emphasizes the role of mechanisms in scientific discourse in general, and in scientific explanation in particular [Machamer et al. (2000); Glennan & Illari (2017)]. Inspired by the developments in molecular biology, new-mechanists redefine causalism and argue that to explain a phenomenon consists in providing the mechanism that produces it. In the new-mechanist tradition, mechanisms are taken to be a set of *entities* (parts) and *activities* (operations) with a particular *organization* such that their causal interactions bring the phenomenon to be explained about [Glennan (2002); Bechtel & Abrahamsen (2005); Craver & Darden (2013); Craver (2007); Nicholson (2012); Issad & Malaterre (2015); Deulofeu & Suárez (2018)]. Thus, for a scientific explanation to be mechanistic, it must fulfill two necessary and sufficient conditions. First, it must identify a *model of mechanism* in which the mechanism is individuated by its parts, operations and organization. Second, it must provide a story of how the components of the mechanism are causally connected in such a way that they produce the *explanandum*.

New-mechanists share a basic commitment to a causal view of the world combined with: 1) the rejection of the Hempelian idea that explanations take the form of logical arguments, either inductive or deductive, and 2) the notion that mechanisms provide the causal "ingredient" that scientific explanations require to be genuinely explanatory[1]. Furthermore, they often assume a hierarchical view of mechanisms, acknowledging the existence of a diversity of scientific explanations in every science, thus

neither renouncing to the explanatory role of the special sciences, nor to the possible existence of mechanistic inter-level (hierarchical) explanations among different sciences [Krickel (2018)].

The wide scope of the New Mechanism account of scientific explanation in biology has been questioned due to the existence of explanations that seem to lack the causal ingredient that new-mechanists demand. One of the traditional explanatory types where this happens is in equilibrium explanations, where the mathematical properties of the empirical system (i.e. the fact that it reaches an equilibrium point) are taken as explanatory, irrespectively of the causal-mechanistic details of the system. Starting with Sober (1983), equilibrium explanations have been hypothesized to constitute an alternative to purely causal-mechanistic explanations [Batterman & Rice (2014); Rice (2015); Huneman (2018b), (2018c)]. However, it has also been argued that some equilibrium explanations admit a causal interpretation, if "causality" is understood in Woodward's interventionists terms [Woodward (2003); Kuorikoski (2007); Potochnik (2015)]. If the later were the case, as some new-mechanists are committed to an interventionist Woodwardian view of causation [Craver (2007); Kaplan & Craver (2011)], it could be argued: first, that the mathematical components that are present in equilibrium explanations describe the causal relationships among the entities of the system; second, that equilibrium explanations do not then constitute a real exception to the new-mechanist trend. The existence of these contradictory interpretations of the nature of equilibrium explanations (causal vs. non-causal) creates an important gap to understand how they gain their explanatory force, as well as about the specific role of causality in scientific explanation: is causality — at some level — a necessary ingredient in every scientific explanation, or are non-causal explanations also legitimate in certain cases?

In this paper, we aim to clarify this issue by studying Blaser & Kirschner's (2007) nested equilibrium model (NEM, hereafter) of the persistence of bacteria in human hosts. Our choice of this case is motivated by two reasons: on the one hand, Blaser & Kirschner's NEM explains the phenomenon in terms of the existence of an evolutionarily stable strategy (ESS, hereafter) among the different interacting organisms, a feature that makes it sufficiently analogous to most cases of equilibrium explanations reviewed in the philosophical literature so that our conclusion can shed light on the nature of scientific explanation; on the other hand, the explanatory force of their model is also conditional on the existence of a nestedness among different biological scales, i.e. on the

existence of a hierarchy of interrelated ESSs. As the acknowledgment of the existence of a hierarchy of mechanisms is a hallmark of the new-mechanist account of scientific explanation, and, to our knowledge, cases of nested equilibria have never been studied before in the philosophical literature, we believe that our case study could bring new light to the study of the old phenomenon of equilibrium explanations. Our aim is thus to analyse the explanatory role that the appeal to the existence of equilibria at different levels plays in the NEM. In that vein, we intend to provide a better understanding of the nature of equilibrium explanation, and to the role of causality in scientific explanation[2]. To do so, we frame the paper in the context of the debate between Huneman's structural account of scientific explanation and the causal-mechanistic account.

In section I, we introduce the general account of structural explanations presented by Huneman (2018a) and motivate the necessity of discussing the precise nature of equilibrium explanation to understand whether, and if so, to what extent, equilibrium explanations fit Huneman's account, or are rather a special case of causal-mechanistic explanations. In section II, we present our case study. In section III, we present our philosophical analysis. We first argue that the explanatory force of Blaser & Kirschner's NEM is mainly provided by the concept of ESS, plus the mathematical modelling that defines each strategy at each of the levels of the hierarchy, rather than by the causal-mechanistic details of the system. Additionally, the nested nature of the different ESSs plays a role in making the system robust to every possible intervention at different levels. Thirdly, and connected to this last point, we argue that no role is left for any causal element in their model, thus suggesting that their explanation constitutes a case of structural explanation as Huneman has defined it. Finally, in section IV, we present our conclusions.

I. Explaining With and Without Causes: The Role of Mathematics in Equilibrium Explanations

In recent years, the universal application of the "new-mechanist" account of scientific explanation in biology has been questioned on the basis of the existence of a family of explanations that do not rely on any causal features of the system whose properties they explain, but rather on its mathematical properties [Huneman (2010), (2018a), (2018b); Woodward (2013); Rice (2015); Kostic (2018), (2019); Deulofeu et al. (2019)]. Huneman has called these explanations "structural", and defines them as follows:

> Family of explanations for which the mathematical tools used in the description of an explanandum system belong to a mathematical structure whose properties are directly explanatory of some aspects of the system (such as equilibria, behaviour, limit regime, asymptotic behaviour, etc.) (…) They explain by accounting for the explananda through pinpointing structural relations that are mathematical relations of some sort. Mathematics here are not representing a dependence between structures in the world, but they are constituting the structural dependence itself, (…) and in virtue of that they are explanatory [Huneman (2018a), p. 695].

In contrast with mechanistic explanations, structural explanations do not include any mechanism, nor any causal story in their *explanans*. Furthermore, the inclusion of any of these elements would usually be taken as counterproductive to account for the *explanandum*. Structural explanations are abundant in systems biology, where an extensive amount of data has to be interpreted by using mathematical and computational tools [Green (2016), (2017); Green & Jones (2017); Brigandt et al. (2017)]. Huneman explicitly argues that some of the properties of the biological systems studied under the label of "systems biology" can only be explained by appealing to the formal (mathematical) properties that characterize those systems. A well-known example of this, studied by Jones (2014), is the vulnerability of the immunological system to attacks to the CD4+ T-cells. Drawing upon Kitano & Oda's (2006) case study, Jones argues that what explains the vulnerability of the human immune system to attacks on this particular component is its bowtie structure: because the human's immune system has a bowtie structure such that CD4+ T-cells are non-redundant elements in the core of the bowtie, the system is vulnerable to attacks on this type of cells (Figure 1). What is more important is that the vulnerability to attacks on CD4+ T-cells is not a consequence of the causal-mechanistic processes that produce the vulnerability: it is a consequence of the topological properties of the architecture (organization) of the immunological system. These topological properties determine its vulnerability to attacks on its core, as it is the only non-redundant element of the network, which is furthermore a necessary step for every other immunological process. Huneman summarized this kind of explanation as follows: "what is epistemically proper to this network modelling is that the topological properties found in the networks are such that they explain some of the properties one is interested in [vulnerability to attacks on CD4+ T-cells], (…) the instantiation of these properties is explained by the fact that the network is of such topological nature" [Huneman (2018b) p 127].

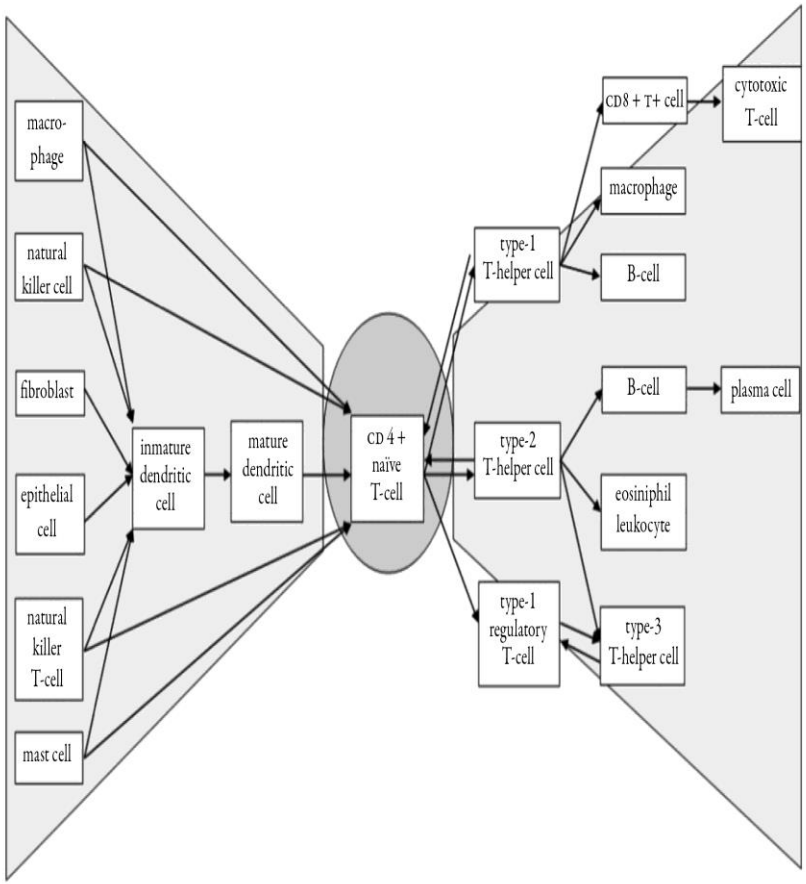


FIGURE 1. Bowtie structure of the immune system, with the CD4+ T-cells in the core of the bowtie. From Jones (2014), p. 1138, Fig. 1.

A second point that is epistemically proper to this kind of explanation is that the mechanisms that "sustain" the realization of such topological properties are irrelevant for explaining those properties (namely, the vulnerability of the network) [Huneman (2018c) pp. 6-8; Deulofeu et al. (2019); Moreno & Suárez, (submitted)][3].

Structural explanations are not restricted to cases of topological explanation, though. In his (2018c), p. 6, Huneman outlines the case of explanations in microeconomics, particularly the "ice cream vendors" problem — a direct application of the theory of Nash equilibrium to

human behaviour. In this situation, we imagine that there are two vendors standing on a beach and need to decide where to situate their stall in order to maximize their sales. Microeconomics, relying on game theory, says that the vendors will situate their stall in the middle of the beach, next to each other, to attract customers both in the area around them and in their extremes. By placing themselves in the middle of the beach, the vendors generate a Nash equilibrium, a situation where none of the players (the vendors) can change their strategy without decreasing their benefits (potential customers). Let us suppose we have to explain a scenario where there are two vendors placed in the middle of the beach. What explains the fact that both of them place their stalls in the middle? Huneman replies: "the fact that it simultaneously maximizes the share of each of them, or in other words, that it instantiates a Nash equilibrium." And adds: "[t]he mechanisms through which vendors move, decide, sell or buy, etc. are not explanatory relevant to this precise question" [Huneman (2018c), p. 6].

Nonetheless, Huneman just sketches the elements that make the Nash equilibrium explanatory in the case of the "ice cream vendors" but does not specify in detail what explaining with equilibria exactly entails, nor what is his reason to believe that mechanisms do not play any explanatory role in equilibrium explanations. Previous analyses of the role of equilibria in scientific explanations had been presented in Sober (1983) and Kuorikoski (2007). However, both authors reach opposing conclusions about where equilibrium explanations gain their explanatory force from: while the former argues that "equilibrium explanations show how the cause of an event can be (statistically) *ir*relevant to its explanation", and that their explanatory force comes exclusively from their mathematical structure [Sober (1983), p. 201], the latter believes that "explanations of singular events are indeed causal, even those supplied by equilibrium models" [Kuorikoski (2007), p. 149]. These opposing conclusions are interesting because they leave open whether equilibrium explanations must be considered a subtype of structural explanation (Sober), or a subtype of causal-mechanistic explanation (Kuorikoski), thus creating an important gap in how to understand the role of mathematics in this type of explanation. In addition to that, they leave open a question about the role of causality in scientific explanation in general for, if as Kuorikoski argues, even equilibrium explanations are in the end causal, then it could be argued that causality is a necessary ingredient in every genuine case of scientific explanation.

In the next section, we introduce Blaser & Kirschner's NEM of the persistence of bacteria in human hosts as a case study that we will use to motivate our response to these two questions.

II. A Nested Equilibrium Explanation of the Persistence of Bacteria in Human Hosts

Humans harbour an abundant number of microbes in their guts that constitute the human microbiome [Huttenhower et al. (2012); Lozupone et al. (2012)][4]. Among those microorganisms, some persist in our guts throughout our entire whole life cycle, whereas others are mainly transient, or appear in specific moments of our development, disappearing afterwards [Chiu & Gilbert (2015)]. Furthermore, some of those are hypothesized to have established long-term associations with humans over millions of years, with some people speculating that they might constitute co-evolved systems or hologenomes [Rosenberg & Zilber-Rosenberg (2014), (2016); Díaz (2015); Suárez (2018); Suárez & Triviño (2019); cf. Moran & Sloan (2015); Douglas & Werren (2016)]. Irrespectively of the evolutionary nature of those associations, the fact that organisms from different species engage in persistent long-term associations with each other is paradoxical from the perspective of the neo-Darwinian model of life and evolution. According to this model, when two individuals of different species associate, i.e. when they share the same habitat or niche, each one will pursue its own fitness interests. In this scenario, it might happen that the two organisms coexist peacefully for a period of time but, normally, peaceful coexistence will tend to break down: on the one hand, in the moment in which an opportunity for one of the organisms to benefit in detriment of the other appears, it will tend to grow to maximize its fitness until the other organism is destroyed (appearance of cheaters); on the other hand, it is also not infrequent that in a stable biological population where one out of two different survival strategies has been adopted among the members, the population becomes invaded by individuals that adopt an alternative strategy, until the point where the population collapses (external invasion). For these reasons, peaceful associations among organisms of different species are rare and will normally be short-term. Then, how is it possible that humans and some of their microbes establish persistent infections that are not disrupted by cheaters[5]? And which are the mechanisms that allow long-term associations that survive the challenges of sharing a habitat and are not perturbed by external invaders?

Blaser and Kirschner have recently developed a model "*to explain* the common features of microbial persistence in their human hosts" [(2007), p. 847, emphasis added)], i.e. to explain why humans and some specific microorganisms have overcome the difficulties of co-habitation[6]. They speculate that those situations represent a successful phenotype that must be maintained according to certain eco-evolutionary rules. In their view:

> persistence represents the evolved selection for balancing host and microbial interests, resulting in an equilibrium that, by definition, is long-term but not necessarily forever stable. We hypothesize that maintenance of this equilibrium requires a series of evolved, nested equilibria to achieve the overall homeostasis [Blaser & Kirschner (2007), p. 843].

They argue that such nested equilibria will be observed at different timescales: microscopic, at the level of the interactions between the immunological system of the host and cell-receptors of the microbes; mesoscopic, at the level of tissue function; tissue in which the microbe population inhabits; macroscopic, where evolutionary changes in the host and the microbe will occur to guarantee microbe transmission[7]. Blaser and Kirschner believe that any of these levels conforms to Nash equilibria in the form of an ESS that allows the persistence of the relationship. This is so because both the host and the microorganism will have developed a very specific hierarchy of cross-signalling mechanisms that generate a set of positive and negative feedback loops with each other that guarantee that the overall equilibrium is not disrupted.

Blaser and Kirschner's model begins by defining five populations at the microlevel whose changes with respect to certain variables are followed over time [see also Blaser & Kirschner (1999); Blaser & Atherton (2004); Blaser (2006)]. In the case of *Helicobacter pylori*, the variables include: *M*, which represents the population of mucus-living *H. pylori* (rate of change); *A*, which represents the *H. pylori* population that adhere to epithelial cells; *N*, which represents the concentration of nutrients available to bacteria derived from inflammation; *E*, which represents the concentration of effector molecules (molecules that the microbes generate to achieve some aims, such as suppressing immune response by the host); and *I*, that stands for the host response. Blaser and Kirschner's NEM includes five differential equations that track the changes in the variables of their model, as well as how they interact with each other[8].

For instance, to study how the concentration of mucus-living *H. pylori* varies over time due to the interaction with the other populations, they introduce the following differential equation:

$$\frac{dM}{dt} = g_m \alpha N(t) - \mu_m M(t) - \alpha M(t)(K - A(t)) + \delta A(t) \qquad (1)$$

where, g_m, α, μ_m and δ are parameters, whose value will depend on the situation; N, M, A (mentioned above) and K (the epithelial carrying capacity) are variables that together will determine the rate of change of the mucus-living population M. In (1), $g_m \alpha N(t) M(t)$ represents the potential growth of the population in virtue of the nutrient availability; $\mu_m M(t)$, represents the loss of *H. pylori* due to the process of mucus shedding; and $\alpha M(t)(K - A(t)) + \delta A(t)$ represents the potential loss/gain of *H. pylori* due to migration between the epithelial and the mucus-living populations. Obviously, migration from M to A can only happen when $A < K$, namely, when there is still room for more adherence to epithelial cells, and the opposite is the case for migration from A to M. Adherent sites are always limited or otherwise *H. pylori* would grow too much, risking the stability of the symbiotic association.

The inflammation induced by the bacteria on the host is captured by measuring the change of nutrient concentration over time:

$$\frac{dN}{dt} = \frac{b}{(b + I(t))} \beta E(t) - g_m N(t) M(t) - g_\alpha N(t) A(t) \qquad (2)$$

In (2), b, β, g_m and g_α are parameters. $N(t)$ is characterized by a gain term that is a function of the concentration of effector molecules, E, and the host response I. The equation shows the direct proportionality that exists between E and N, and the inverse proportionality between I and N. In other words, it shows the limiting effect that the host response has over the nutrient concentration, as well as the inducing effect of the bacteria on the nutrient concentration. (2) also specifies the rate of assimilation of nutrients of the mucus-living bacterial population and of the adherent epithelial populations.

Furthermore, for a microbe-host association to be *evolutionarily* persistent, the microbe needs to develop strategies for transmission. R_o captures this concept, quantifying "the transmission potential of a microparasite as the average number of secondary infections occurring when a single infectious host is introduced into a universally susceptible host population" [Blaser & Kirschner (2007) p. 844].

$$R_o = \frac{BN}{(x + b + v)} \tag{3}$$

In (3), *BN* measures the transmission rate as a function of the population size, *x* measures the rate of host mortality due to the microbe (measure of virulence), *b* is the rate of mortality of the host population independently of the microbe (measure of lifespan), and *v* is the rate at which the host recovers from the microbe infection (measure of immunity). Usually, for $R_o > 1$ microbial transmission is sustained whereas for $R_o < 1$ microbial transmission goes extinct.

Blaser and Kirschner show that in a persistent microbe-host association those five differential equations remain constant, and any deviation in one of the equations gets immediately counter-balanced by the adjustment of the other equations, keeping the equilibrium stable. Thus, Blaser and Kirschner claim this can only be possible if the system behaves according to a Nash equilibrium, and if the strategies followed by microbe and host conform to an ESS. Let us now see how an ESS can account *explanatorily* for observed constancy.

II.1. *The Role of the Evolutionarily Stable Strategy in Blaser and Kirschner's Model*

Nash equilibrium is a very common situation in game theory. It obtains when two players in a non-cooperative game adopt a strategy such that no individual change will render greater benefits to any of them, i.e. such that every change in the strategy that one of the players adopts independently will result in lower individual profit for that player. Nash equilibria are not necessarily, however, optimal strategies. It is sometimes possible to obtain a better net result if both players change their strategy simultaneously and a new equilibrium is reached. Nonetheless, this will only occur if *both partners* modify their strategy co-ordinately, but not if they do so independently. Therefore, no player has any incentive to modify his strategy individually. The prisoner's dilemma constitutes a typical example of a game whose solution is provided by a Nash equilibrium (Table 1). In this situation, two individuals — A and B — are accused independently of a crime, and each of them is interrogated separately and offered a deal: 1) if A betrays B and accuses her of having committed the crime, while B stays silent, A will have 4-years reduction of sentence and B will have no reduction (and the same, but inverted, occurs if B betrays A while A remains silent); 2) if both stay silent, each

of them will have a 3-years reduction of sentence; 3) if both betray each other, each will have a 1 years reduction of sentence. In this scenario, the Nash equilibrium is reached in situation 3), when both players betray each other. Of course, the result that they obtain is not optimal (each of them will only get 1 year reduction of sentence), but is such that none of them has any incentive to change her strategy individually, unless the other also does so, as otherwise she will have a bigger individual cost, i.e. she will have less years of reduced sentence [Nash (1950a), (1950b); Gintis (2000)].

A \ B	Betrays	Remains silent
Betrays	**1, 1**	4, 0
Remains silent	0, 4	**3, 3**

TABLE 1. Payoff matrix for the prisoner's dilemma. The numbers represent the amount of years that each subject would have as reduction of sentence. The optimal strategy is that where both remain silent (italics). Only the strategy where both betray constitutes Nash equilibria (bold).

An ESS is a biological strategy that, when it is adopted in a population, natural selection alone will keep the population safe from "intruder populations", in so far as the organisms that adopt an alternative strategy will be selected against. All ESSs are cases of Nash equilibria, but the opposite is not the case. If a solution to a non-cooperative game represents Nash equilibrium that is not an ESS, the solution could be disrupted by an alternative strategy that drives the population towards an alternative Nash equilibrium that constitutes an ESS [Smith & Price (1973); Smith (1974); Easley & Kleinberg (2010), pp. 209-227]. For instance, take the case of the stag hunt game (Table 2). This is a two players' game, where each player has two possible exclusive strategies: hunt-hares or hunt-stags. In this situation, there are three possible scenarios: 1) that both individuals are hare-hunters (case where both obtain a fitness benefit of 2); 2) that both individuals are stag-hunters (both obtain a fitness benefit of 3); 3) that one of the individuals is a hare-hunter whereas the other is a stag-hunter (in which case the hare-hunter obtain a fitness benefit of 3, whereas the stag-hunter obtains a fitness benefit of

0). In this situation, strategies 1) and 2) constitute a Nash equilibrium, for none of the players could get a better payoff by changing strategy. However, only 1) constitutes an ESS: while a hare-hunter and a stag-hunter do equally well when they are paired with a stag-hunter (fitness benefit of 3), hare-hunters score better than stag-hunters when they are paired with hare-hunters (hare-hunters score 2, while stag-hunters score 0). That means the stag-hunting strategy is not an ESS because if a hare-hunter is introduced in a population of stag-hunters, the population will evolve towards a population of hare-hunters. On the other hand, a population where all the individuals are hare-hunters represents an ESS, because if a stag-hunter is introduced in the population, it will be eventually extinct, for its fitness benefit will be lower than the fitness benefit of hare-hunters.

	Stag-hunter	Hare-hunter
Stag-hunter	**3, 3**	0, 3
Hare-hunter	3, 0	***2, 2***

TABLE 2. Payoff matrix for the stag hunt game. The numbers represent the net benefit for the individuals in the population that engage in the game. Cases where all the individuals in the population hunt exclusively stags or exclusively hares represent Nash equilibria (bold). However, only the case where both individuals hunt hares represent an ESS (italics).

Blaser and Kirschner apply this type of reasoning to persistent long-term host-microbe associations to argue that the situation must be the one that is obtained in Nash equilibrium, particularly in ESSs, where both positive and negative feedback between the host and the microbe occur, so that the equilibrium persists over time. The core idea of their model is that the equilibrium obtained at the microscopic level immediately affects the equilibrium at superior levels (mesoscopic and macroscopic). At the same time, the equilibrium at the higher levels affects in a specific way the possibility of new microbe-host persistent associations. The equilibria are nested and the association does not get in principle disrupted. The interaction among levels, partially captured by the equations (1)-(3), is as follows:

first, on the microscopic level one would find the microbial population, localized on an organ or tissue of the host, and the population of immune host cells responsible of recognizing the microbe population. The structure of both populations will depend on the nature of the original founder strain, the possibility for generating genetic variants, the selective pressures from other microbial cells in the same tissue and, more importantly, from the selection that the persistent microbe and the immune cells exert on each other [e.g. (Pradeu et. al 2013); Pradeu & Vivier (2016); Eberl (2016)]. The nature of the interactions between the organisms in the microscale will shape tissue function (or malfunction), and thus will partially determine the viability of the host, as well as the opportunity for microbial transmission (mesoscale). Finally, the effects of the microbe on the viability of the host will determine the host population structure (macroscale) that in return will affect microbial transmission (mesoscale) (Figure 2).

Even if the model illustrated in Figure 2 looks like a multilevel mechanism, for it appeals to a model of mechanism, it lacks the adequate type of causal stories that new-mechanists demand to have a proper explanation. First, because multilevel causation is mysterious, as Craver and Bechtel illustrate (2007), since causal relations happen exclusively intra-level. Second, because the type of inter-level readjustments of the system are symmetrical, occurring both top-down (e.g. from the macroscale to the mesoscale, or from the latter to the microscale), and bottom-up (e.g. from the microscale to the mesoscale, or from the latter to the macroscale), while relations between cause and effect are always asymmetrical. Third, because even if there could be a way to capture inter- and intra-level causal relations, this would be at odds with the information that NEM conveys and appeals to. NEM does not specify the causal way in which the entities at one level affect the entities at another level. It only specifies that the disruption of the equilibrium at one level will either prompt the collapse of the system (i.e. its death), or it will prompt the re adjustment of the equilibrium at that level due to the equilibria that exist in the other scales. In other words, NEM is not specific about how the equilibrium will be readjusted, it only predicts that it will be readjusted, provided that the other levels keep their equilibrium states. The causal elements (if any) that will bring this readjustment are irrelevant for the explanation of this behaviour in terms of NEM. What matters is exclusively the nested structure of the host-symbiont system (see section 4 for the full details).

In that vein, the nested structure of the model and the level of complex interactions between the different elements at the three scales (Nash equilibria, ESS) grant the persistence of the association. As it was

said before, one of the reasons why host-microbe associations do not normally last long is due to the presence of cheaters, organisms that enjoy the profits of the associations without paying the cost. Nash equilibria avoid the appearance of cheaters: cheaters are players that change their strategy unilaterally; in Nash equilibria, every player that does so is condemned to failure, and thus will be removed from the population. Furthermore, as the Nash equilibria that are reached in the population adopt the form of an ESS, it is not possible that an external invader adopting an alternative strategy disrupts the persistence of the association.

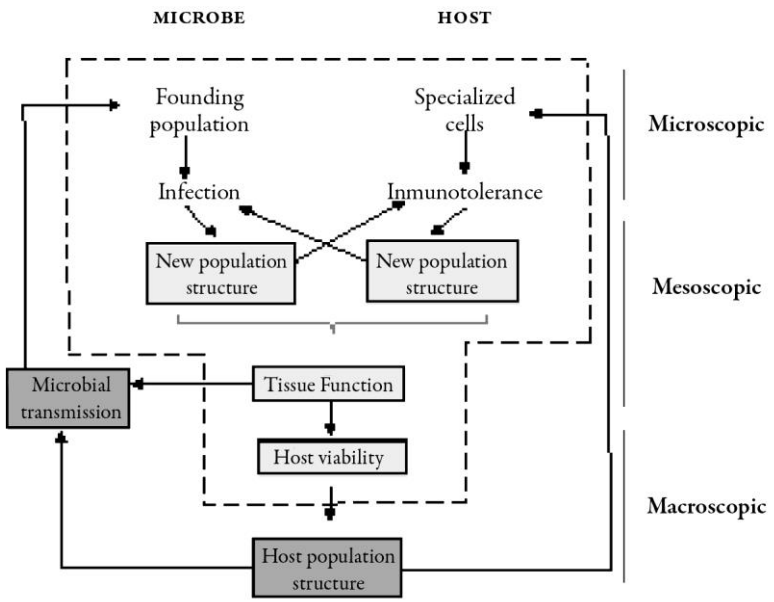


FIGURE 2. Nested equilibrium model. The dashed box represents those events that occur within the host. Adapted from Blaser & Kirschner (2007), p. 845, Fig. 2).

III. EQUILIBRIUM EXPLANATIONS AS STRUCTURAL AND NON-MECHANISTIC EXPLANATIONS

Blaser and Kirchner's NEM was developed to account for the persistence and the long-term character of certain human-microbe associations. Concretely, the authors seek to explain two paradoxes: first, why the association is not disrupted by the appearance of cheaters, i.e. entities that ben-

efit from the association without paying the costs; second, why the bacterial population is not entirely substituted by an intruder/external invader that deploys a different strategy. Only if those two phenomena are avoided, persistent host-bacterial associations can be successful. We will now argue that Blaser and Kirchner's NEM explains how those phenomena are avoided by appealing to mathematical, *but not causal*, properties, of host-microbial associations. In other words, we will argue that the alleged explanatory force of the NEM lies in the fact that: (i) it provides a mathematical structure in the form of a set of differential equations that together satisfy an ESS; (ii) that the nested nature of the ESSs makes the explanation of host-microbe persistence robust to any perturbation; (iii) that this is so because the properties of the ESS directly mirror the properties of the biological system in a non-causal way.

First of all, as shown in section II, Blaser and Kirschner's NEM consists in a series of differential equations that describe how the concentration of bacteria in different host tissues, their effector cells, their nutrient availability, the immunological response and their rate of transmission will change over time. These equations, as we explained, do not contain *a priori* any information about the persistence of the host microbe relationship. However, they provide information about how the different variables must be related to each other so that persistence obtains. Particularly, the equations measure the impact of host immunological response on bacterial colonization and, in doing so, allow determining the level at which host's response will abruptly disrupt colonization, as well as the levels at which bacterial inflammation will trigger a decrease in nutrient availability that in the end will disrupt colonization. And, in addition, they provide information about the way in which the solutions to these equations that guarantee the persistence of the symbiotic relation relate to: a) the rate of transmission of the symbiont (R_o), b) the viability of the host (tissue function and evolutionary advantages).

The set of equations can be resolved for a concrete host-symbiont system, and the evolution of the variables under study, as well as their interrelation, can be analysed. This will provide information about how they relate and how they are maintained constant, allowing predictions about empirical system[9]. However, notice that they would still provide no information about our *explanandum*, i.e. about what makes the host-microbe relationship persistent. To do so, the set of equations must be embedded in the framework of ESSs, i.e. it must model the biological situation as a non-cooperative game of two players, such that if any of the players (host, microbe) follows a unilateral strategy, the consequences

will be detrimental for the player that does so. That this is so can be seen by studying how changes in the equations that relate the concentration in nutrient availability, immune response, microbial concentration, etc. will relate to each other to make the system collapse if the change is unilateral. However, as we argued, the explanatory character of the equations comes exclusively from the possibility of embedding them in the framework of ESS. In other words, they are explanatory sound because it is possible to realize that no unilateral change that disrupts the system is possible without generating a chain reaction that either reverses the change or destroys the system. The ESS thus explains stability by ruling out two alternative scenarios: one where cheaters spread in the population, and another when an invader population entirely substitutes the actual one.

Second, the explanatory force of the ESS is reinforced in Blaser and Kirschner's NEM due to its nested nature. The nested nature of the equilibria works as a check and balances system which prevents that a disruption of the ESS at one of the levels (microscopic, mesoscopic and macroscopic) spreads across the other levels and destroys the host-microbe association. Let us explain this with an example: take the case of a disruption at the mesoscale that substitutes the microbe population for an invader. As we are at the mesoscale, the invader will disrupt tissue function in its own benefit, e.g. growing more than what the original microbial population would have grown, while at the same time escaping from the barriers of the immunological system. This type of change, totally beneficial for the bacteria at the mesoscale, would trigger two responses: First, a response at the macroscale that would be immediately detrimental for the bacteria. At this level, host viability, which is affected by the tissue function, will be reduced and, as a consequence, bacterial transmission will substantially decrease in relation to the transmission of those bacteria that cause no damage in tissue function. Secondly, at the microscale, where the invader population will not have generated immunotolerance, the invader population will be systematically blocked by the specialized immunological cells, especially the cells of the adaptive immune system. Furthermore, it is expected that the host will reduce nutrient availability, so that it affects in the long-run the intruders' population structure. Remember, as we said in section II, that the key of the ESS is that no player that changes its strategy unilaterally will be better. In this situation, even if the "player" might be better in one particular scale (mesoscale), the same will not be true for the other scales, and thus no possibility for invasion exists[10].

Third, and more concretely about the nature of ESS, we believe that Blaser and Kirschner's NEM, as any explanation that appeals to the existence of an ESS, explains the stability of host-microbe persistent associations in a non-causal way. Let us argue why we believe this to be so.

1) Blaser and Kirschner's NEM appeals to general properties of ESSs, and they make their model explanatory in virtue of the equivalence between the theoretical ESSs framework and the general properties of persistence host-symbiont associations. The strategy is the general strategy of Huneman's structural explanations: first, build a system *S'* whose properties match the properties of the real system *S* whose behaviour you aim to track. Second, study the behaviour of *S'* and attribute its properties to *S*. In Blaser and Kirschner's NEM, the strategy is applied as follows: first, build the ESS model for host-microbe persistent associations, as a case of a non-cooperative game of two players; second, study the behaviour of the ESS model, i.e. why the existence of an ESS, as the optimal solution for both players (Nash equilibrium), excludes the possibility of cheaters and invasive populations; third, attribute the properties of the ESS model to the empirical phenomenon, i.e. to empirical cases of host-microbe persistent associations. Notice that in this schema the explanatory force comes because the mathematical system that is built, in this case an equilibrium model, behaves in a certain way that (allegedly) is the way in which the empirical system will behave. But, importantly, it is irrelevant how the empirical phenomenon causally realizes the properties that it is attributed. And this is so in a double sense: on the one hand, because the NEM neither mention, nor needs to mention the specific species that interact to generate the ESS; on the other, because the causal connections between the entities (*if any*) are epistemologically irrelevant for the explanation of the phenomenon.

2) Despite the highly problematic way of identifying interlevel causal relations in a multilevel mechanism, as Craver and Bechtel (2007) explain, one could still try to appeal to Woodward's interventionist strategy to identify the supposed causes explaining the persistence of host-microbe associations. However, we believe NEM rules out the possibility of generating or even heuristically imagining any intervention *à la* Woodward, thus contradicting Kuorikoski and Potochnick's interpretation of equilibrium explanations. Let us explore this via an example. Recall that the *explanandum* is the phenomenon of persistence host-microbe associations. How would an intervention look like in Blaser and Kirchner's NEM? The only possibility would be to generate a situation such that the ESS disappears. However, no possible intervention is imaginable without destroying the system. Or,

in other words, any imaginable intervention that would make host-microbe associations non-persistent would directly change the system we are trying to explain, and thus the information it will provide will turn out to be irrelevant to account for the phenomenon. Recall the structure of ESS (Table 2). The only possibility of imagining a significant intervention would be via a change in the expected payoffs for the actions of each player. However, this intervention would not give any relevant information about why the association is stable in certain circumstance, because it would directly shift the focus of attention towards a new system, namely, one where there is not an ESS. Or, in other words, a causal explanation would consist in saying that the ESS is explanatory because if there were not an ESS the host-microbe association would not be stable. But this kind of reasoning is uninformative and, in our view, unexplanatory. The structural interpretation *à la Huneman*, on the contrary, offers a plausible account of how Blaser and Kirschner's NEM gains its explanatory force.

More importantly, the nested nature of the model, far from moving its explanatory force in a causal-mechanistic direction, generates the opposite effect. It just makes any possible intervention less imaginable. Because even if one causal intervention could be imagined for one specific level, how would it possibly work, if its effects would be cancelled out due to the existence of ESSs in the other levels? Or, in other words, how is it possible to imagine an intervention that causally escapes the inter-level connection? This connection is just a property of any host-microbe persistent association, and the explanatory power of the nestedness resides, precisely, in its possibility to cancel out the effect of every possible intervention. Therefore, we argue, a causal interpretation of the explanatory power of Blaser and Kirschner's NEM is not possible, since it would simply make the explanatory force of the model completely mysterious.

Of course, one might agree with what we just said, and still believe that our argument does not rule out the fact that the most appropriate interpretation of the explanatory force of Blaser and Kirchner's NEM is indeed causal. For instance, Blaser and Kirschner explicitly argue that specific host-microbe associations (human-*H. pylori*, human-*Salmonella typhi*, etc.) are "not necessarily forever stable" [(2007), p. 843], as obviously context (environment) matters, and in a changing context (environment) it is possible that concrete associations go selected against, simply because the environment selects against that coevolved system [see Díaz (2015); Suárez & Triviño (2019)]. In this context, it is possible to investigate the causes that made the system collapse, and if this is so, then the same must

be true for the cases in which the association is persistent. Nonetheless, we disagree, because that will entail changing the *explanandum* in two senses: first, making it specific to particular species; second, explaining the disruption of the persistence, instead of the persistence itself. And remember that our original *explanandum* was why some host-microbe associations are persistent, and the cases to rule out are the cases of cheaters and invasive populations. In our view, their model should be interpreted counterfactually: if a host-microbe association is persistent throughout the host's life cycle and evolutionarily long-term, then it will satisfy the conditions of the NEM reached through an ESS. And this situation will be so irrespectively of the species that interact, and thus irrespective of the causal-mechanisms that host and microbe could have developed to reach that equilibrium. As in the case of the ice vendors (section I), where the psychological mechanisms that have driven the vendors to put their stalls in the middle of the beach are explanatorily irrelevant to understand why their stalls are there, in the case of persistent associations causal-mechanistic details are simply superfluous. One can perfectly omit all those details and the explanation would still be epistemically sound.

Alternatively, an enumeration of the causes (if any) that would determine whether a concrete host-microbe association is stable will be irrelevant to explain its persistence if it is not conceived as a consequence of an ESS. This is because it would still be possible to imagine the existence of cheaters or invasive populations that deploy the same causal-mechanistic "machinery" to escape e.g. immunitary controls, without paying the cost of the symbiotic association. However, as we explained, because the host-microbe association constitutes a nested ESS, both the cheater and the invader population will end up disappearing from the population, just because the host-microbe persistent system has the structure that appears in the mathematical formulation of ESSs. Importantly, we are not here saying that Blaser and Kirchner's NEM rules out the possibility of telling a causal story of why concrete host-microbe associations are, sometimes, persistent, although some story about how to speak about interlevel causation should be provided.[11] Furthermore, we believe that such causal stories *could* be told to explain specific host-microbe associations, even when these must be complemented with the appeal to ESSs. Our point is rather *epistemological*: causal stories that seek to explain the existence of persistent host-microbe associations are neither required, nor explanatory in themselves. The element that provides the explanatory strength in equilibrium explanations is purely structural (in Huneman's terms), and it is connected with the possibility of accounting for the existence of an equilibrium (in Blaser and Kirschner's NEM, a nested ESS).

IV. Conclusion

In this paper, we have examined the explanatory force of equilibrium explanations, and have studied whether the explanatory force of equilibrium explanations can be better justified by applying the causal-mechanistic model of scientific explanation, or Huneman's structural model. Concretely, we have examined the role that mathematical vs. causal properties play in the explanation of the stability of persistent long-term host-microbe associations. Explaining the stability of this type of associations is paradoxical, as it requires explaining two facts: first, the absence of cheaters; second, the impossibility of the population being substituted by an intruder population. We have used Blaser and Kirschner's NEM to illustrate that the explanation of host-microbe persistent associations does not seem to be causal, but structural, relying solely on the non-causal mathematical properties of the association to explain its long-term persistence [Huneman (2018a), (2018b)]. We have argued that Blaser and Kirschner's NEM is explanatory of the long-term persistence of host-microbe associations because (i) it provides a mathematical structure in the form of a set of differential equations that together satisfy an ESS; (ii) that the nested nature of the ESSs makes the explanation of host-microbe persistence robust to any perturbation; (iii) that this is so because the properties of the ESS directly mirror the properties of the biological system in a non-causal way. In this vein, our case study shows how equilibrium explanations, even if nested, gain their explanatory force from the mathematical structure that describes the system, instead of from the causal interactions among its components. Our analysis supports two theses: first, that equilibrium explanations, even if nested (in a hierarchical setting), are structural rather than causal-mechanistic; second, that causality, even if necessary in some explanations, is not a universally necessary requirement of every scientific explanation.

Logos – Research Group in Analytic Philosophy / Barcelona Institute of Analytic Philosophy
Department of Philosophy
University of Barcelona
C/Montalegre nº6, 08001. Barcelona, Spain
E-mail: javier.suarez@ub.edu
E-mail: roger.deulofeu@gmail.com

ACKNOWLEDGMENTS

A previous version of this paper was presented in the meeting 'Process epistemology: A workshop with Bill Bechtel', University of Exeter, May 2017. The authors want to thank all the participants for their feedback, and especially to Bill Bechtel for a wonderful discussion. José Díez, John Dupré, Philippe Huneman, Thomas Pradeu, Johannes Findl and two anonymous reviewers are acknowledged for their comments on earlier versions of the paper. The following institutions are formally acknowledged: Javier Suárez and Roger Deulofeu, Spanish Ministry of Economy and Competitiveness (FFI2016-76799-P); Roger Deulofeu, Spanish Ministry of Economy and Competitiveness (BES-2013-063239); Javier Suárez, Spanish Ministry of Education (FFU16/02570).

NOTES

[1] The commitment to a causal view of the world does not entail either a physical reductionism [as in Salmon (1984)] or an "ontic" interpretation of scientific explanation [as in Craver (2014)]. Cf. Glennan (2002), Bechtel & Abrahamsen (2005), for a model-based interpretation of mechanisms.

[2] There are other cases where equilibrium models have been used to explain the stability of biological associations [Baalen & Jansen (2001); Selosse *et al.* (2006)]. We have chosen to analyse Blaser & Kischner's NEM for its generality, and because it is a case of equilibrium explanation generally accepted among biologists. Nonetheless, our conclusions also apply to these cases. Thanks to Philippe Huneman for pointing this fact to us.

[3] Following Brigandt (2013), we consider that an element of an *explanans* is explanatory relevant if and only if removing it from the explanation entails that the *explanandum* does not follow, and it's explanatory irrelevant otherwise [(2013), p. 480].

[4] "Microbiota" refers to "[t]he assemblage of microorganisms present in a defined environment", and "microbiome" is used to denote "the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions" in a given environment [Marchesi & Ravel (2015), p. 1]. For the purposes of this paper, we will not distinguish the two concepts, and they will be used to refer only to the community of microorganisms present in a given environment.

[5] In biology, persistent infection refers to lifelong associations between a host and some species of microbes that do not necessarily harm the host, although they might do it in the long-term. The term should not be confused with its medical use, where "infection" is usually employed in reference to pathogens, or disease-causative agents.

[6] Their model is in principle developed exclusively for pair associations, between one host and one microorganism.

[7] Those different levels have both a temporal and a scale correlation: the macroscale refers to the evolutionary time, the mesoscale refers to organismal development and the microscale refers to the interactions among different cell types.

[8] Since our purpose is only to illustrate the main features of the model and their relation to Blaser and Kirschner's explanation, for a matter of simplicity we only introduce two of the equations.

[9] Information about the values that the variables must take for a concrete (empirically real) host-microbe association, if the association is known to be stable.

[10] It exists, but if and only if the intruder changes the situation *in the three scales*. That is precisely the nature of the nested model.

[11] See Craver & Bechtel (2007) for a proposal.

REFERENCES

BATTERMAN, R.W., and C. C. RICE (2014), "Minimal Model Explanations"; *Philosophy of Science* 81. 3, pp. 349-376.

BECHTEL, W., and A. ABRAHAMSEN (2005), "Explanation: A Mechanist Alternative"; *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2), pp. 421–441.

BLASER, M. J. (2006), "Who Are We? Indigenous Microbes and the Ecology of Human Diseases"; *EMBO Rep* 7, pp. 956–960.

BLASER, M. J. and J. ATHERTON (2004), "*Helicobacter pylori* Persistence: Biology and Disease"; *J. Clin. Invest*. 113, pp. 321–333.

BLASER M. J., and D. KIRSCHNER (1999), "Dynamics of *Helicobacter pylori* Colonization in Relation to the Host Response"; *Proc Nat Acad Sci* 96, pp- 8359-8364.

— (2007), "The Equilibria that Allow Bacterial Persistence in Human Hosts"; *Nature* 449, pp. 843–849.

BRIGANDT, J. (2013), "Systems Biology and the Integration of Mechanistic Explanation and Mathematical Explanation"; *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44, pp. 477–492.

BRIGANDT, J., S. GREEN, and M. A. O'MALLEY (2017)," Systems Biology and Mechanistic Explanation"; in S. Glennan and P. Illari (eds.) *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, London: Routledge (chapter 27).

CHIU, L., and S. F. GILBERT (2015), "The Birth of the Holobiont: Multi-species Birthing Through Mutual Scaffolding and Niche Construction"; *Biosemiotics* 8 (2), pp. 191–210.

CRAVER, C. F. (2007), *Explaining the Brain*; New York: Clarendon Press.

— (2014), "The Ontic Account of Scientific Explanation"; in Kaiser, M. I., Scholz, Plenge, R. D. Hüttemann, A. (eds.), *Explanation in the Special Sciences: The Case of Biology and History*, Springer Verlag. pp. 27–52.

CRAVER, C. F., and W. BECHTEL (2007.), "Top-Down Causation Without Top-Down Causes"; *Biology & Philosophy* 22(4), pp. 547-563.

CRAVER, C. F., and L. DARDEN (2013), *In search for Mechanisms: Discovery Across the Life sciences*; Chicago: University of Chicago Press.

DEULOFEU, R. and J. SUÁREZ (2018), "When Mechanisms Are Not Enough: The Origin of Eukaryotes and Scientific Explanation"; in Christian A., Hommen D., Retzlaff N., Schurz G. (eds) *Philosophy of Science*. European Studies in Philosophy of Science, vol 9. Springer, Cham.

DEULOFEU, R, J. SUÁREZ and A. PÉREZ-CERVERA (2019), "Explaining the Behaviour of Random Ecological Networks: The Stability of the Microbiome as a Case of Integrative Pluralism"; *Synthese. https://doi.org/10.1007/s11229-019-02187-9.*

DÍAZ, J. S. (2015), "El Mecanismo Evolutivo de Margulis y los Niveles de Selección"; *Contrastes* XX (1), pp. 7-26.

DOUGLAS A. E. and J. H. WERREN (2016), "Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts"; *mBio* 7 (2), e02099-15.

EASLEY, D. and KLEINBERG, J. (2010), *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*; Cambridge University Press.

EBERL, G. (2016), "Immunity by Equilibrium"; *Nat. Rev. Immunol.* 16, pp. 524-532.

GINTIS, H. (2000), *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior*.; Princeton University Press.

GLENNAN, S. (2002), Rethinking Mechanistic Explanation; *Philosophy of Science* 69 (S3), pp. S342–S353.

GLENNAN, S. and ILLARI, P. (Eds.) (2017), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, Taylor and Francis.

GREEN, S. (2016), *Philosophy of System Biology*; Dordrecht: Springer.

— (2017), "Philosophy of Systems and Synthetic Biology"; *The Stanford Encyclopedia of Philosophy* (Edition Spring 2019), Edward N. Zalta (ed.), URL = <*https://plato.stanford.edu/archives/spr2019/entries/systems-synthetic-biology/*>.

GREEN, S. and JONES, N. (2016), "Constraint-Based Reasoning for Search and Explanation Strategies for Understanding Variation and Patterns in biology"; *Dialectica* (70)3, pp. 343-374.

HUNEMAN, P. (2010), "Topological Explanations and Robustness in Biological Sciences"; *Synthese* 177, pp. 213–245.

— (2018a), "Outlines of a Theory of Structural Explanation"; *Philosophical Studies* 175 (3), pp. 665–702.

— (2018b), "Diversifying the Picture of Explanations in Biological Sciences: Ways of Combining Topology with Mechanisms"; *Synthese* 195, pp. 115–146.

— (2018c); "Realizability and the Varieties of Explanation"; *Studies in History and Philosophy of Science. <https://doi.org/10.1016/j.shpsa.2018.01.004>*

HUTTENHOWER C, GEVERS D, KNIGHT R, CREAS HH, *et al.* (2012) "Structure, Function and Diversity of the Healthy Human Microbiome"; *Nature* 486, pp. 207–214.

ISSAD, T., and C. MALATERRE (2015.), "Are Dynamic Mechanistic Explanations Still Mechanistic?"; in P. A. Braillard and C. Malaterre (eds.) *Explanation in*

Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences. Dordrecht: Springer, pp. 265–292.

JONES, N. (2014), "Bowtie Structures, Pathway Diagrams, and Topological Explanations"; *Erkenntnis* 79 (5), pp. 1135–1155.

KAPLAN, D. M. and C. F. CRAVER (2011), "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience. A Mechanistic Perspective"; *Philosophy of Science* 78.4, pp. 601-627.

KITANO, H., and K. ODA (2006), "Robustness Trade-Offs and Host-Microbial Symbiosis in the Immune System"; *Molecular Systems Biology* 2, pp. 1–10.

KOSTIĆ, D. (2018), "The Topological Realization"; *Synthese*, 195(1), pp. 79-98.

— (2019), "Minimal Structure Explanations, Scientific Understanding and Explanatory Depth", *Perspectives on Science*, 27 (1), pp. 48-67.

KRICKEL, B. (2018), *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*; (Vol. 13). Springer.

KUORIKOSKI, J. (2007), "Explaining with Equilibria"; in Persson, J., and Ylikoski, P. (Eds.), *Rethinking explanation*; Springer, Dordrecht, pp. 149-162.

LOZUPONE, C. J. I. STOMBAUGH, J. I. GORDON, J. K. JANSSON, and R. KNIGHT (2012), "Diversity, Stability and Resilience of The Human Gut Microbiota"; *Nature* 489 (7415), pp. 220–230.

MACHAMER, P., DARDEN, L., and C.F. CRAVER (2000), "Thinking About Mechanisms"; *Philosophy of science*, 67(1), pp. 1-25.

MARCHESI, J. R., and J. RAVEL. (2015), "The Vocabulary of the Microbiome Research: A Proposal"; *Microbiome* 3, p. 31.

MORAN, N., and D. B. SLOAN (2015), "The Hologenome Concept: Helpful or Hollow?"; *PLoSBiol* 13 (12), e1002311.

MORENO A., and J. SUÁREZ (submitted), "Plurality of Explanatory Strategies in Biology: Mechanisms and Networks".

NASH, J. F. (1950a), "The Bargaining Problem"; *Econometrica: Journal of the Econometric Society* 18 (2), pp. 155-162.

— (1950b), "Equilibrium Points in N-Person Games"; *Proceedings of the National Academy of Sciences* 36 (1), pp. 48-49.

NICHOLSON, D.J. (2012), "The Concept of Mechanism in Biology"; *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1), pp. 152–163.

POTOCHNIK, A. (2015) "Causal Patterns and Adequate Explanations"; *Philosophical Studies*, 172(5), pp. 1163-1182.

PRADEU, T., JAEGER, S., and E. VIVIER (2013), "The Speed of Change: Towards a Discontinuity Theory of Immunity?"; *Nature Reviews Immunology*, 13(10), p. 764.

PRADEU, T., and E. VIVIER (2016), "The Discontinuity Theory of Immunity"; *Sci. Immunol.* 1 (1): aag0479.

RICE, C. (2015), "Moving Beyond Causes: Optimality Models and Scientific Explanation"; *Noûs* 49.3 pp. 589-615.

ROSENBERG E and I ZILBER-ROSENBERG (2014), *The Hologenome Concept*. London, Springer.

— (2016) "Microbes Drive Evolution of Animals and Plants: The Hologenome Concept"; *mBio* 7 (2): e01395-15.

SALMON W. (1984), *Scientific Explanation and the Causal Structure of The World*; Princeton: Princeton University Press.

SELOSSE, M. A., RICHARD, F., HE, X., and S.W. SIMARD (2006), "Mycorrhizal Networks: des Liaisons Dangereuses?"; *Trends in Ecology and Evolution* 21 (11), pp. 621-628.

SMITH, J. M. (1974), "The Theory of Games and the Evolution of Animal Conflicts"; *Journal of Theoretical Biology*, 47(1), pp. 209-221.

SMITH, J. M., and G. R. PRICE (1973), "The Logic of Animal Conflict"; *Nature*, 246(5427), p. 15.

SOBER, E. (1983), "Equilibrium Explanation"; *Philosophical Studies* 43.2, pp. 201-210.

SUÁREZ, J. (2018). "The Importance of Symbiosis in Philosophy of Biology: An Analysis of the Current Debate on Biological Individuality and its Historical Roots"; *Symbiosis* 76(2) pp. 77-96.

SUÁREZ, J. and V. TRIVIÑO (2019), "A Metaphysical Approach to Holobiont individuality: Holobionts as Emergent Individuals"; *Quaderns de Filosofia* 6(1), pp. 59-76.

VAN BAALEN, M. and V.A. JANSEN (2001), "Dangerous Liaisons: The Ecology of Private Interest and Common Good": *Oikos*, 95(2), 211-224.

WOODWARD, J. (2003), *Making Things Happen: A Theory of Causal Explanation*; New York: Oxford University Press.

— (2013), "Mechanistic Explanation: Its Scope and Limits"; *Aristotelian Society Supplementary Volume* 87 (1), pp. 39–65.

teorema
Vol. XXXVIII/3, 2019, pp. 121-142
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 121-142]

Inference to the Best Explanation and the Screening-Off Challenge

William Roche and Elliott Sober

RESUMEN

Defendemos en Roche y Sober (2013) que la explicatividad es evidencialmente irrelevante, esto es, que Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*), donde *H* es una hipótesis, *O* una observación y *EXPL* es la proposición de que si *H* y *O* fueran verdaderas, entonces *H* explicaría *O*. Esta es una "tesis de neutralización" ["screening off" thesis, de ahí el nombre "SOT"]. En el presente artículo clarificamos esta tesis, replicamos a las críticas presentadas por Lange (2017), consideramos algunas formulaciones alternativas de la "Inferencia a la mejor explicación", defendemos dos versiones más fuertes de la tesis, que denominamos "SOT*" y "SOT**", y consideramos cómo estas inciden en la afirmación de que la virtud teórica de la unificación es evidencialmente relevante.

PALABRAS CLAVE: *bayesianismo; relevancia evidencial; explicatividad; inferencia a la mejor explicación; Lange; neutralización; unificación.*

ABSTRACT

We argue in Roche and Sober (2013) that explanatoriness is evidentially irrelevant in that Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*), where *H* is a hypothesis, *O* is an observation, and *EXPL* is the proposition that if *H* and *O* were true, then *H* would explain *O*. This is a "screening-off" thesis (hence the name "SOT"). Here we clarify SOT, reply to criticisms advanced by Lange (2017), consider alternative formulations of Inference to the Best Explanation, defend two strengthened screening-off theses called "SOT*" and "SOT**", and consider how they bear on the claim that unification is evidentially relevant.

KEYWORDS: *Bayesianism; Evidential Relevance; Explanatoriness; Inference to the Best Explanation; Lange; Screening-off; Unification.*

I. INTRODUCTION

We argue in Roche and Sober (2013) that explanatoriness is evidentially irrelevant in that Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*), where, here and throughout, *H* is a hypothesis, *O* is an observation, and *EXPL* is the proposition that if *H* and *O* were true, then *H* would explain *O*. This is a

"screening-off" thesis (hence the name "SOT") to the effect that *O* screens-off *EXPL* from *H* in that given *O*, *EXPL* has no impact on the probability of *H*. Suppose, for example, that you examine a large random sample of people older than age 50, and that on this basis, you arrive at the following population frequency estimate:

(1) Freq(heavy smoking before age 50 | lung cancer after age 50) = α

Suppose that from your perspective, Joe is a random member of the population (and was not in your sample). Let *H* be the hypothesis that Joe was a heavy smoker before age 50, *O* be the observation that Joe got lung cancer after age 50, and *EXPL* be the proposition that if *H* and *O* were true, then *H* would explain *O*. From your perspective, Joe is a random member of the population, so $Pr(H \mid O) = \alpha$. Given this, and given that *EXPL* should have no impact on your estimate of the frequency in (1), $Pr(H \mid O\&EXPL) = \alpha$. Hence $Pr(H \mid O\&EXPL) = Pr(H \mid O)$.

SOT and our defense of it have not been met with universal approval, to say the least. McCain and Poston (2014) were the first to chime in with objections. Climenhaga (2017) was next, and then Lange (2017) took his turn. We have responded to McCain and Poston [see Roche and Sober (2014)] and to Climenhaga [see Roche and Sober (2017b).[1] Here we respond to Lange, but that's not all. There are numerous non-equivalent versions of Inference to the Best Explanation (IBE) in logical space. We will argue that SOT refutes some of them, but not others. We then will defend two variants of SOT called "SOT*" and "SOT** and argue that they refute many of the remaining versions, specifically, versions of IBE that say that whether *H* is the best available potential explanation of *O* hinges on how *H* scores in terms of unification.

II. Response to Lange

Lange holds that there are many realistic cases where *EXPL* isn't screened-off from *H* by *O*, because $Pr(H \mid O\&EXPL) > Pr(H \mid O)$.[2] He gives two examples that he claims are of this sort. We discuss one of them in Section 2.1 and address the other in Section 2.2. We argue in each case that Lange fails to show, or even to make it plausible, that $Pr(H \mid O\&EXPL) > Pr(H \mid O)$. In Section 2.3, we clarify SOT, and provide a more general response to objections like Lange's.

II.1 *Lange's Robbery Example*

Lange's first alleged example of a realistic case where Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*) involves a robbery in which a jewel is stolen from a safe. After introducing the example, Lange argues that Pr(*H* | *O*) is greater than Pr(*H*) but less than maximal (i.e., less than unity):

> … suppose that *H* is that Jones is the person who stole the jewel from the safe, *O* is that the single strand of hair found inside the safe was blond, and the background information tells us that there was exactly one robber and one strand of hair found inside the safe, that Jones has blond hair, and that such a hair has a serious (though not overwhelming) likelihood to have been left by the robber during the robbery (though there are other ways in which the hair could have gotten into the safe). The background information also tells us that Jones is a serious suspect, unlike many other people with blond hair – although Jones is one among several serious suspects with blond hair and there is also a fair likelihood that the robber is not listed among our serious suspects. Background also tells us that if the hair were Jones's, then Jones would probably be the robber (since he would have left it during the robbery); Jones would have had no occasion to access the safe except to rob it.
>
> Accordingly, since the hair that was found is the same colour as Jones's hair, *O* lends some support to *H* – Pr(*H* | *O*) > Pr(*H*) – though this support is less than maximal, considering that the hair may not have come from the robber and that, even if it did come from the robber, the robber need not be Jones since many other people (including some other serious suspects) have blond hair [Lange (2017), p. 305].

Lange then adds *EXPL* to the mix,[3] and argues that Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*):

> But to *O* in the evidence let's now add *EXPL*: that if Jones were the robber and the single strand of hair found inside the safe were blond, then that Jones is the robber would explain why the strand of hair found in the safe is blond. The explanation would be that Jones left the hair in the course of the robbery. Without *EXPL*, the evidence's power to confirm *H* is rendered less than maximal by (among other things) serious doubt that the hair comes from the robber. But that doubt is removed by *EXPL* (at least in the event that Jones is the robber). Of course, the evidence's power to confirm *H* remains somewhat less than maximal because of other factors, such as doubt about whether the hair comes from Jones. But *EXPL* removes one consideration that mitigated the degree to which *O* pointed to Jones (namely, the possibility that even if Jones were the rob-

> ber, such a hair would not belong to Jones because it was not left by the robber). Consequently, *H* is better confirmed by *O*&*EXPL* than by *O* alone: Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*) [Lange (2017), pp. 305-306].

Lange's rationale, then, is two-fold. First, there's the claim that because of various doubts, Pr(*H* | *O*) is less than maximal. Second, there's the claim that *EXPL* eliminates some of the doubts in question, and as a result increases the probability of *H*.

What doubts make Pr(*H* | *O*) less than maximal? Lange initially mentions these:

(2) The strand of hair found in the safe wasn't left by the robber.

(3) The strand of hair found in the safe was left by the robber, but the robber wasn't Jones and instead was one of the other serious suspects with blonde hair.

He later points to this:

(4) Jones is the robber, but the strand of hair found in the safe doesn't belong to him because it wasn't left by the robber.

Lange claims that *EXPL* eliminates (4), and because of this, *EXPL* increases the probability of *H*. This is strange, since (4) is a possibility in which *H* is *true*. How is it that by eliminating a possibility in which *H* is true, *EXPL increases H*'s probability? Instead, why not think that by eliminating a possibility in which *H* is true, *EXPL decreases H*'s probability?

What Lange is describing can happen, but we doubt that it is true in his robbery example. To illustrate how an observation can raise the probability of a hypothesis and also eliminate a possibility in which the hypothesis is true, consider this example. A card is randomly drawn from a standard well-shuffled deck of cards. Let *A*, *D*, and *S* be understood as follows:

A: The card drawn is an Ace.

D: The card drawn is a Diamond.

S: The card drawn is a Spade.

Given that ~(*A*&*D*)&~*S* entails ~(*A*&*D*), it follows that ~(*A*&*D*)&~*S* eliminates *A*&*D* and thus eliminates a possibility in which *D* is true. Yet ~(*A*&*D*)&~*S* nonetheless increases *D*'s probability from 1/4 to 12/38.

The thing to notice here is this: although $\sim(A\&D)\&\sim S$ eliminates a possibility in which D is true, it also eliminates several possibilities in which D is *false* (for example, $A\&S$).

Is something similar true in Lange's robbery case? That is, is it the case that, though *EXPL* eliminates (4), it also eliminates various possibilities in which H is *false*? There's no doubt that *EXPL* eliminates *some* possibilities in which H is false. It eliminates, for example, $\sim H\&\sim EXPL$. But this, by itself, is insignificant, since, at the same time, and for the same reason, it also eliminates $H\&\sim EXPL$.

Consider, instead, these possibilities:

(5) Smith, not Jones, is the robber. The strand of hair found in the safe doesn't belong to Smith, and was instead planted by him as a distraction to the police.

(6) Smith, not Jones, is the robber. The strand of hair found in the safe was left by the owner of the jewel a few days before the robbery while she was taking something other than the jewel out of the safe.

Clearly, assuming, with Lange, that the background information on hand is realistic, *EXPL* doesn't eliminate (5) or (6).

It might be that Lange can flesh out this example of his so that it is clear that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$. We return to this possibility in Section II.3.

II.2. *Lange's Physics Example*

Lange's second attempt to provide a realistic example in which $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$ is taken from physics, where H is the hypothesis that the light quantum hypothesis is empirically adequate, and O is an equation specifying the black-body spectrum.[4] First, he argues that if H is true but the light quantum hypothesis is false, then it's a mere coincidence that various phenomena behave as if the light quantum hypothesis is true:

> If H holds but light is not quantized, then it is just a coincidence that various phenomena behave as if light is quantized: the fundamental natural laws of light (which do not include that light is quantized) entail the particular derivative laws (such as O) that govern various phenomena, and the light-quantum hypothesis also entails those laws, but there is no common reason why these two facts hold: this combination is 'nothing more than a

> curious property of light, without any physical significance' [Lange (2017), p. 309].

Second, he considers the possibility that *H* is true because the light quantum hypothesis is true too:

> On the other hand, if *H* holds because light is indeed quantized, then it is no coincidence that various phenomena behave as if light is quantized. Rather, there is a common reason why the fundamental laws of light and the light-quantum hypothesis are alike in sharing the property of entailing the derivative laws. The common reason is that the light-quantum hypothesis is one of those fundamental laws. If *H* holds, then if there are light-quanta, *H* can explain *O*; the explanation is roughly that since there are light-quanta, everything behaves as if there are (i.e., *H* holds), including the black-body spectrum, and *O* is what the black-body spectrum would be if *H*. However, if *H* holds, then *H* cannot explain *O* if there are no light-quanta, since then *H* is just a fluke [Lange (2017), p. 6].

Third, he argues that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O\&\sim EXPL)$:

> … compare $\Pr(H \mid O\&EXPL)$ to $\Pr(H \mid O\&\sim EXPL)$. Both $O\&EXPL$ and $O\&\sim EXPL$ confirm *H* to some degree. But $O\&\sim EXPL$ confirms *H* only by removing one way in which *H* could have gone wrong—that is, only by confirming one part of *H* (that the light-quantum hypothesis entails the correct equation for the black-body spectrum) and having no bearing on the rest of *H* (that the light-quantum hypothesis entails the correct equations for the photoelectric effect, for the Volta effect, …). In contrast, $O\&EXPL$ confirms *H* not just by confirming that the light-quantum hypothesis gets the black-body spectrum right, but also by confirming that the light-quantum hypothesis gets various other phenomena right. Therefore, $\Pr(H \mid O\&EXPL) > \Pr(H \mid O\&\sim EXPL)$ …. [Lange (2017), pp. 309-310)].

Fourth, and finally, he concludes that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$.[5] Lange (2017), p. 311, says that inequalities of the following form are central to his argument:

> (7) Pr(the … phenomenon behaves as if there were light-quanta | $O\&EXPL$) >> Pr(the … phenomenon behaves as if there were light-quanta | $O\&\sim EXPL$).

Consider, for example, the following, where *O** is the proposition that the photoelectric effect behaves as if there were light-quanta:

(8) $\Pr(O^* \mid O\&EXPL) >> \Pr(O^* \mid O\&\sim EXPL)$

It's not immediately obvious, however, how inequalities like (8) figure in his overall argument. The problem is that (8) makes no mention of *H*, whereas the claim that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$ does.

It might be that Lange is tacitly assuming Hempel's (1965) Converse Consequence Condition (here understood in terms of confirmation in the sense of increase in probability):

> CCC: For any propositions *X*, *Y*, *Z*, and *Z**, if (i) $\Pr(Z \mid Y\&X) > \Pr(Z \mid Y)$ and (ii) *Z** entails *Z*, then $\Pr(Z^* \mid Y\&X) > \Pr(Z^* \mid Y)$.

It follows from (8) that

(9) $\Pr(O^* \mid O\&EXPL) > \Pr(O^* \mid O\&\sim EXPL)$,

and this entails that:

(10) $\Pr(O^* \mid O\&EXPL) > \Pr(O^* \mid O)$.

Given (10), and given, suppose, that *H* entails *O**, it follows by CCC that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$. However, if this is how the argument is supposed to work, then the argument fails. As is well known, CCC has counterexamples.[6]

We now set this problem aside and suppose, for the sake of argument, that there's a legitimate way to get from inequalities such as (8) to the claim that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$.[7] Why should inequalities like (8) be accepted?

We don't understand Lange's answer here, but we have a conjecture based in part on what he says about a slight variant of an example described by a former time slice of one of the authors of the paper you're now reading [Sober (2015)]. Suppose that two of the students in a seminar you are teaching turn in word-for-word identical essays. You consider two hypotheses. *CC* (short for "Common Cause") says that they searched the Internet together, found a paper suited to the assignment, and agreed to plagiarize it. *SC* (short for "Separate Causes") says that they worked separately and independently. Sober (2015), pp. 103-104, formulates the following thesis, and says that it is a "first pass" in need of refinement:

(11) $\Pr(\text{the papers match} \mid CC) >> \Pr(\text{the papers match} \mid SC)$

Lange modifies the example slightly by letting "w" be a certain long sequence of words, and replacing (11) with the following:

(12) Pr(Smith's paper contains w & Jones's paper contains w | *CC*) >> Pr(Smith's paper contains w & Jones's paper contains w | *SC*)

Lange then claims that Sober would agree to (12) because he would also agree that:

(13) Pr(Smith's paper contains w | Jones's paper contains w & *CC*) >> Pr(Smith's paper contains w | Jones's paper contains w & *SC*)

Here is Lange's explanation of why Sober would endorse (12):

> For Smith's paper to contain w, given that Jones's paper contains w but *SC*, would be extremely unlikely. 'According to *SC*, the matching is a coincidence; according to *CC*, it is anything but' (Sober 2015: 103) [Lange (2017) p. 310].

The idea here (and elsewhere in Lange's discussion) seems to be that Sober would accept (12) because he would accept:

(14) Given *SC*, it would be a coincidence if both Smith's paper and Jones's paper were to contain w, and so given *SC* and that Jones's paper contains w, it is highly unlikely that Smith's paper also contains w, whereas given *CC*, it would not be a coincidence if both Smith's paper and Jones's paper were to contain w, and so, given *CC* and that Jones's paper contains w, it is not highly unlikely that Smith's paper also contains w.

Note the transitions from "would be a coincidence" to "is highly unlikely", and from "would not be a coincidence" to "is not highly unlikely".[8]

Lange seems to agree with Sober (as he reads him) on all of this, and further seems to think that analogous points hold in his physics case. This suggests that Lange accepts (8), for example, because he accepts:

(15) Given ~*EXPL*, it would be a coincidence if both *O* and *O** were true, and so given ~*EXPL* and *O*, it is highly unlikely that *O** is true, whereas given *EXPL*, it would not be a coincidence if both *O* and *O** were true, and so, given *EXPL* and *O*, it is not highly unlikely that *O** is true.

This, like (14), has transitions from "would be a coincidence" to "is highly unlikely", and from "would not be a coincidence" to "is not highly unlikely".

However, Sober (2015) *denies* that a common cause explanation of a "matching" between two events always has a higher likelihood than a separate cause explanation of that matching. Sometimes the matching of the two events favors a common cause explanation that says that the matching is not a coincidence, but sometimes it does not. Everything depends on the background assumptions that pertain. In the example of the student essays, it's easy to see how (12) could be false. Suppose that if the students work together and plagiarize, they will be loathed to include word sequence w, but if they work separately and independently, the chances of them including that sentence in their essays is much greater. Notice that our point here does not depend on this supposition's being realistic.[9]

If, as it seems, Lange's view is that (8) should be accepted because (15) should be accepted, then his argument is in trouble. For, (15), like (14), should be rejected.

We noted at the end of Section II.1 that it might be that Lange can flesh out his robbery example so that it's clear that $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$. The same is true with respect to his physics example.

II.3. *SOT's Scope*

It might seem that SOT *universally* quantifies over all logically possible cases:

> TOO STRONG: For any logically possible case in which $\Pr(H \mid O\&EXPL)$ and $\Pr(H \mid O)$ are well-defined, O screens-off $EXPL$ from H in that $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$.

Alternatively, it might seem that SOT is much more modest, in that it *existentially* quantifies over all logically possible cases:

> TOO WEAK: There are logically possible cases in which O screens-off $EXPL$ from H in that $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$.

In fact, neither reading is right. TOO STRONG is obviously false. Suppose, for instance, that $\Pr(H \mid O)$ is less than unity, and that the background information codified in Pr(-) includes the assumption that:

> (16) $(EXPL\&H) \vee (\sim EXPL\&\sim H)$

It follows that Pr(*H* | *O*&*EXPL*) = 1 > Pr(*H* | *O*). TOO WEAK, in turn, is obviously true but utterly boring. If, for example, Pr(*H* | *O*) equals unity, and Pr(*H* | *O*&*EXPL*) is well-defined, then, trivially, Pr(*H* | *O*&*EXPL*) = 1 = Pr(*H* | *O*).

How, then, should SOT be understood? Inspired by Goldilocks, we understand it to be saying this:

> JUST RIGHT: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *EXPL* from *H* in that Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*).

Our smoking example is a case of this sort, but there are many – very many – other examples of the same kind.

We acknowledge, though, that there are some potentially misleading passages in Roche and Sober (2013). Here is one:

> Our screening-off thesis is related to Van Fraassen's (1989) thesis that inference to the best explanation (IBE) is probabilistically incoherent, and therefore subject to a Dutch book. Van Fraassen thinks that IBE proposes a two-step rule for updating: if the evidence *O* increases *H*'s probability, then *H* receives a further boost in probability if *H* would provide a good explanation of *O*. Our argument aims to show that the explanatoriness of *H* *cannot* provide this additional boost; in addition, it side-steps the question of how the apparently prudential considerations introduced by Dutch book arguments are relevant to a non-prudential notion of rational degree of belief [Roche and Sober (2013), p. 665, emphasis added]

Our use of the word "cannot" may suggest that we meant TOO STRONG as opposed to JUST RIGHT, but that wasn't our intent. We meant *cannot* (*in cases like our smoking case where there's abundant frequency data on hand*). We regret not making this completely clear.

Given that SOT should be understood as JUST RIGHT, it follows that even if Lange fleshed out his robbery example or his physics example so that it's clear that Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*), SOT would remain unscathed. SOT allows for realistic cases in which Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*). It even allows for realistic cases in which the background information codified in Pr(-) includes frequency data such that Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*).[10]

Is SOT, understood as JUST RIGHT, trivial in the way that TOO WEAK is trivial? We think not, and now will explain why.

III. SOT AND IBE*

IBE can be formulated in different ways. Here is one:

> IBE*: If (i) *O*, (ii) *H* is a potential explanation of *O*, and (iii) *H* is better overall in terms of explanatory virtues v_1, v_2, …, and v_n than each of the available rival potential explanations of *O*, then it's rational to believe *H* and disbelieve each of the rivals.

This is a relatively standard formulation, but there are others. We discuss other formulations in Section V.

Where does probability come into play in IBE*? Let *BEST* be the proposition that *H* is the best overall available rival explanation of *O* in terms of virtues v_1, v_2, …, and v_n . We assume that IBE* entails that there are no cases in its scope where *O*, *EXPL*, and *BEST* are true but:

(17) $\Pr(H \mid O\&EXPL\&BEST) \leq 0.5$

For, presumably, if $\Pr(H \mid O\&EXPL\&BEST) \leq 0.5$, then it isn't rational to believe *H* and disbelieve each of the rivals.[11]

We further assume that IBE* entails that at least some cases in its scope where *O*, *EXPL*, and *BEST* are true are such that:

(18) $\Pr(H \mid O\&EXPL\&BEST) > t > \Pr(H \mid O)$

Here "*t*" is the threshold for high probability, and should be understood so that it is less than 1 and greater than or equal to 0.5.[12] If *EXPL*&*BEST* never increases *H*'s probability (given *O*) from low (not high) to high, then why build a theory of inference around *EXPL*&*BEST*?[13]

Now consider:

(19) $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$

Should IBE* be understood so that every case in its scope where *O* and *EXPL* are true is a case where (19) holds? SOT is relevant here. The claim that *H* is a potential explanation of *O* is in effect *EXPL*. This, at any rate, is how IBE-ists typically construe the notion of a potential explanation. Consider, for instance, the following:

> 'Explains' in the second premise [i.e., the premise, in our notation, that *H* explains *O*] cannot, without begging the question, mean 'actually explains'; rather, it is used in the sense of 'would explain if true' [Lycan (2002), p. 413].
>
> A *potential* explanation of the evidence is anything that *would* explain the evidence *if it were true* [Williamson (2016), p. 266, emphasis original].

Lycan doesn't use the expression "potentially explains", but that's the intended contrast with "actually explains".[14] If IBE* should be understood so that every case in its scope where *O* and *EXPL* are true is a case where (19) holds, and if IBE*'s scope includes all realistic cases, then SOT entails that IBE* is false.[15]

We aren't insisting, though, that IBE* should be understood in that manner. We are simply addressing one potential way of understanding it.

IV. SOT* AND IBE*

The present task is to consider the possibility that IBE* does *not* require that every case in its scope where *O* and *EXPL* are true is a case where (19) holds. The first point to note is that SOT has a cousin:

> SOT*: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *BEST* from *H* in that $\Pr(H \mid O\&BEST) = \Pr(H \mid O)$.

There's no explicit mention here of *EXPL*. But it's there implicitly, since *BEST* is logically equivalent to *EXPL*&*BEST*, which means that $\Pr(H \mid O\&BEST) = \Pr(H \mid O\&EXPL\&BEST)$. Is SOT* true, and if so, does it undermine IBE*?

It might seem that SOT* follows from SOT. For, *BEST* is logically stronger than *EXPL*, and it might seem that for any propositions *X, X*, Y*, and *Z*, if (i) $\Pr(Z \mid Y\&X) = \Pr(Z \mid Y)$ and (ii) *X** is logically stronger than *X*, then $\Pr(Z \mid Y\&X^*) = \Pr(Z \mid Y)$. But, as with CCC, there are exceptions, for example, where $1 > \Pr(Z \mid Y\&X) = \Pr(Z \mid Y)$, and *X** is the conjunction of *X* and *Z*.

There are many different explanatory virtues noted in the extant literature on IBE.[16] Some commonly cited examples are:

(a) empirical adequacy

(b) explanatory power

(c) fit with background data

(d) fertility
(e) internal consistency
(f) internal coherence
(g) mechanism
(h) parsimony
(i) precision
(j) scope
(k) unification

Different sets of explanatory virtues lead to different versions of IBE*.[17] We want to focus on versions of IBE* on which unification is included in v_1, v_2, …, and v_n. We assume for definiteness that a common cause explanation is always superior in unification to a separate cause explanation. This way of understanding unification has some *prima facie* appeal, and has been explicitly endorsed in the extant literature on unification.[18]

Suppose, adapting a case introduced in Lange (2004) and later modified in Blanchard (2018), that your background information includes the following frequency data:

(20) Freq(pleuritis & malar rash | lupus) = 0.891 > 0.0495 = Freq(pleuritis & malar rash)

(21) Freq(lupus) = 0.005

(22) Freq(lupus | pleuritis & malar rash) = 0.09

Let *L*, *M*, and *P* be understood as follows:

L: Jones has lupus.

M: Jones has a malar rash.

P: Jones has pleuritis.

Given (20), (21), and (22), and given that, suppose, Jones is a random member of the population from your perspective, you should have the following probabilities:

(23) Pr(*P*&*M* | *L*) = 0.891 > 0.0495 = Pr(*P*&*M*)

(24) Pr(*L*) = 0.005

(25) $\Pr(L \mid P\&M) = 0.09$

You then learn that $P\&M$, and as a result rightly increase your credence in L from 0.005 to 0.09.

Now let B and F be understood as follows:

B: Jones has Bloom's disease.

F: Jones has the flu.

Suppose that you subsequently come to learn that pleuritis is caused both by lupus and by the flu, and that malar rashes are caused both by lupus and by Bloom's disease. You then have two potential explanations of $P\&M$. One is L, which is a common cause explanation. The other is $F\&B$, which is a separate cause explanation. Which is better in terms of v_1, v_2, …, and v_n?

Given the assumption that a common cause explanation is always more unifying than a separate cause explanation, it follows that L is superior in unification to $F\&B$. Suppose that L is equal or superior to $F\&B$ in terms of each of the remaining explanatory virtues included in v_1, v_2, …, and v_n, and that you learn this and thus further learn that L is better overall in terms of v_1, v_2, …, and v_n than $F\&B$.[19] Should you increase your credence in L from 0.09 to some higher value? It seems clear that the answer is negative; your credence in L should remain at 0.09.

What if you learned not just that L surpasses $F\&B$ in terms of v_1, v_2, …, and v_n, but also that since there are no additional available potential explanations of $P\&M$, L surpasses each of its rival available potential explanations in terms of v_1, v_2, …, and v_n? What if, in other words, you learned *BEST*? The answer, it seems, is the same: *BEST* is screened-off in that your credence in L should remain at 0.09.

This verdict can be bolstered by adding some further details to the example. Suppose that your frequency data goes beyond (20), (21), and (22). Suppose in particular that it includes:

(26) Freq(pleuritis & malar rash | flu & Bloom's disease) = 0.891

(27) Freq(flu & Bloom's disease) = 0.005

(28) Freq(flu & Bloom's disease | pleuritis & malar rash) = 0.09

It follows that your frequency data is neutral between L and $F\&B$. For, you know that Jones has pleuritis and a malar rash (and know nothing else relevant about his symptoms), and you know that the frequency of

lupus among people who have pleuritis and a malar rash is equal to the frequency of the flu and Bloom's disease among such people. But then you, like your frequency data, should be neutral between *L* and *F*&*B*.

Things could have turned out differently. Your background information could have included things in addition to the frequency data given in (20), (21), (22), (26), (27), and (28), and this extra information could have had the result that your credence in *L* should be greater than your credence in *F*&*B*. You could have known, for instance, that Jones's partner has lupus, and that lupus is easily passed from person to person. We were supposing, though, that initially Jones was a random member of the population from your perspective, so you didn't have any such extra information.

There is nothing special about our lupus example. It's typical of many realistic cases in which the background information codified in Pr(-) includes frequency data such that, although *BEST* is true (because *H* is superior in unification to each of the available rival potential explanations of *O*), *BEST* is screened-off from *H* by *O* in that Pr(*H* | *O*&*BEST*) = Pr(*H* | *O*). Hence SOT* is true.

How does all this bear on IBE* (or, more specifically, on the versions of IBE* under consideration)? If we are right about our lupus case, and if IBE*'s scope includes all realistic cases, then at least some of the cases in virtue of which SOT* is true are cases where (17) is true and (18) is false because:

(29) $t > \Pr(H \mid O\&EXPL\&BEST) = 0.09 = \Pr(H \mid O)$

But then IBE* is false.

Denying that *BEST* is screened-off in our lupus case doesn't save IBE*. Even if Pr(*H* | *O*&*EXPL*&*BEST*) were greater than 0.09, surely it wouldn't be greater than 0.5, and thus surely it wouldn't be greater than *t*. Hence (17) would still be true.

We noted above that *L* is superior in unification to *F*&*B*, and then simply supposed, without argument, that *L* is equal or superior to *F*&*B* in terms of each of the remaining explanatory virtues included in v_1, v_2, …, and v_n. This seemed legitimate then, and still seems legitimate now. Nothing in the case as specified to this point requires that *L* be *inferior* to *F*&*B* in terms of any of empirical adequacy, explanatory power, fit with background data, or any of the other explanatory virtues noted above (or any additional ones for that matter).

V. SOT** and IBE**

IBE* is a relatively standard formulation of IBE, but there are alternatives. Here is one:

> IBE**: If (i) *O*, (ii) *H* is a potential explanation of *O*, (iii) *H* is better overall (in terms of explanatory virtues v_1, v_2, …, and v_n) than each of the available rival potential explanations of *O*, and (iv) *H*'s overall score in terms of v_1, v_2, …, and v_n is high, then it's rational to believe *H* and disbelieve each of the rivals.[20]

Let *HIGH* be the proposition that *H*'s overall score in terms of v_1, v_2, …, and v_n is high. If *HIGH* holds in cases like our lupus case, then:

> SOT**: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *BEST*&*HIGH* from *H* in that Pr(*H* | *O*&*BEST*&*HIGH*) = Pr(*H* | *O*).

It might be argued, however, that *HIGH* is false in cases like our lupus case. What then?

There would still be problems. The idea behind IBE** is that though *O*&*EXPL*&*BEST* (IBE*'s antecedent) leads to the *fully comparative* claim that *H*'s probability is *greater than* the probabilities of the other potential explanations in question, it doesn't lead to the *partially non-comparative* claim that *H*'s probability is *high*. *HIGH* is supposed to connect the two.[21] Our lupus example, though, shows that sometimes *O*&*EXPL*&*BEST* does *not* lead to the fully comparative claim that *H*'s probability is greater than the probabilities of the other potential explanations in question. This undermines the idea behind IBE**.

Let's set aside this worry and turn to the question of whether there's a legitimate way to motivate the claim that *HIGH* is false in our lupus example. Recall that given the frequency data on hand, Pr(*P*&*M* | *L*) = 0.891, and Pr(*L*) = 0.005. It might be argued that *L*'s explanatory power with respect to *P*&*M* is given by Pr(*P*&*M* | *L*), that *L*'s fit with background data is given by Pr(*L*), and that neither of these is high enough for *HIGH* to be true. What now?

It turns out that there are variants of our example in which each of the following holds:

(30) Freq(pleuritis & malar rash | lupus) = 1 = Freq(pleuritis & malar rash | flu & Bloom's disease)

(31) Freq(lupus) = 0.4 = Freq(flu & Bloom's disease)

(32) Freq(lupus | pleuritis & malar rash) = Freq(flu & Bloom's disease | pleuritis & malar rash) ≈ 0.493

Now Pr(*P*&*M* | *L*) = 1 (up from 0.891) and Pr(*L*) = 0.4 (up from 0.005), so that *L*'s ability to predict *P*&*M* (the observation to be explained), as given by Pr(*P*&*M* | *L*), is maximal, and though *L*'s fit with the background information on hand, as given by Pr(*L*), isn't maximal, it is nonetheless relatively high and much greater than 0.005. If this means that *HIGH* is true, then since your frequency data is still neutral between *L* and *F*&*B*, and since there are many additional examples like this variant of our lupus example, it follows that IBE** is false and SOT** is true.

Friends of IBE** could restrict IBE** to realistic cases where the background information on hand includes no relevant frequency data. Then our lupus examples would pose no threat, but friends of IBE** would be obliged to furnish an independent rationale for this restriction. If the sole reason for this restriction is that otherwise the theory would be open to counterexample, the restriction would be *ad hoc*. Furthermore, friends of IBE** should be worried about more than just frequency-data cases. For, arguably, things other than frequency data, for example, background theories, can enable *O* to screen-off *BEST*&*HIGH* from *H*.[22]

What if friends of IBE** simply excised unification from the set of virtues (or else simply denied the assumption that a common cause explanation is always superior in unification to a separate cause explanation)? Would they then be in the clear? Not necessarily. Our argument carries over to any version of IBE** on which the set of virtues includes a virtue *v* such that there can be realistic cases where *H* is superior in *v* to each of its rival available potential explanations, and yet substantial frequency data at hand is neutral between *H* and at least one of the rivals.[23]

VI. Conclusion

We have defended three screening-off theses, which we'll now repeat for convenience:

- SOT: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *EXPL* from *H* in that Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*).
- SOT*: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *BEST* from *H* in that Pr(*H* | *O*&*BEST*) = Pr(*H* | *O*).
- SOT**: There are many realistic cases in which the background information codified in Pr(-) includes frequency data such that *O* screens-off *BEST*&*HIGH* from *H* in that Pr(*H* | *O*&*BEST*&*HIGH*) = Pr(*H* | *O*).

These theses are all existential, but unlike TOO WEAK, they are far from trivial. First, if IBE* is understood so that every case in its scope where *O* and *EXPL* are true is a case where Pr(*H* | *O*&*EXPL*) > Pr(*H* | *O*), and if IBE*'s scope includes all realistic cases, then SOT refutes IBE*. Second, if IBE* and IBE** are understood so that unification is an explanatory virtue such that a common cause explanation is always more unifying than a separate cause explanation, and if IBE*'s and IBE**'s scopes include all realistic cases, then SOT* refutes IBE*, and SOT** refutes IBE**.

Department of Philosophy
Texas Christian University
Fort Worth, TX 76129, USA
E-mail: w.roche@tcu.edu

Department of Philosophy
University of Wisconsin, Madison
Madison, WI, 53706 USA
E-mail: ersober@wisc.edu

NOTES

[1] McCain and Poston (2017) have responded in kind to our response to their earlier objections. We don't have the space here to respond back.

[2] All references in this section to Lange are to Lange (2017).

[3] Strictly speaking, Lange uses "*E*" as opposed to "*EXPL*". We have changed his notation in the quoted passages below, so that it conforms to our notation.

[4] Lange construes the light quantum hypothesis as the hypothesis "that light comes in discrete quantities rather than continuous waves" [Lange (2017), p. 308]. But since he claims that the light quantum hypothesis entails *O*, which is an equation, we take it that he means for the light quantum hypothesis to be something more than just the hypothesis that light comes in discrete quantities ra-

ther than continuous waves. We shall assume for the sake of argument that the light quantum hypothesis when *fully specified* has the entailments claimed by Lange.

[5] It's a theorem of the probability calculus that for any propositions X, Y, and Z, $\Pr(Z \mid Y\&X) > \Pr(Z \mid Y)$ precisely when $\Pr(Z \mid Y\&X) > \Pr(Z \mid Y\&\sim X)$.

[6] See Roche (2017) for discussion of whether CCC and related theses can be repaired by modifying them in terms of explanation.

[7] We also want to set aside a further problem. In the second displayed passage in this subsection, Lange seems to assume that flukes are explanatorily impotent. However, that assumption is dubious. If Smith and Jones run into each other on State Street at noon on Tuesday by coincidence, then their running into each other is a fluke. But their running into each other can nonetheless explain why they each are smiling then.

[8] It is natural to think that two events that happen at the same time comprise a coincidence precisely when they are causally/explanatorily unconnected (neither causes/explains the other and there is no common cause/explanation). This does not mean that coincidences are inexplicable; that's what separate cause explanations provide [see Sober (2012), p. 362 for discussion].

[9] In addition, Sober (2015) does not commit to the thesis that if the common cause explanation has the higher likelihood, then its value is much bigger than 0.5 whereas the likelihood of the separate cause explanation is much smaller than 0.5.

[10] Lange argues not just that there are realistic case where $\Pr(H \mid O\&EXPL) > \Pr(H \mid O)$, but also that there are cases where $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$ because *EXPL* is a necessary truth. He writes:

> The problem with cases where *EXPL* is a logical necessity is that in such cases, although Roche and Sober are correct that $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$, this equality is trivial since $\Pr(EXPL) = 1$. The equality then fails to show that *H*'s explanatoriness counts for nothing in its confirmation [Lange (2017), p. 308].

We have three comments. First, even if Lange is right that there are cases where $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$ because *EXPL* is a necessary truth, this leaves it open, as per SOT, that there are also many realistic cases in which the background information codified in Pr(-) includes frequency data such that O screens-off *EXPL* from H in that $\Pr(H \mid O\&EXPL) = \Pr(H \mid O)$. Second, *EXPL* isn't a necessary truth in cases like our smoking case. Third, as we explain in Sections III IV, and V, SOT and its cousins SOT* and SOT** are far from trivial, given their implications with respect to various versions of IBE.

[11] Some theorists argue that belief and acceptance are distinct in that a subject can accept a given hypothesis without believing it. See, e.g., Elliott and Willmes (2013). We leave it open whether they are right; maybe there are cases in which $\Pr(H \mid O\&EXPL\&BEST) \leq 0.5$ where it's rational to accept H (though not to believe it).

[12] It might be that t varies from context to context. We take no stand on this.

[13] Cabrera (2017) has pointed out that some of the explanatory virtues are antithetical to high probability. For example, if theory T_1 entails theory T_2, then T_1 can't have a higher posterior probability than T_2, no matter what the evidence is. That said, T_1 may have wider scope than T_2.

[14] It might be that, strictly speaking, there can be cases where H is a potential explanation of O, and yet it's false that if H and O were true, then H would explain O, because something else would explain it [see Lipton (2004), Ch. 4 for discussion]. We're ignoring this possibility here (as is standard).

[15] This objection is distinct from van Fraassen's (1989) "best of a bad lot" objection. See Okasha (2000) for discussion of the latter.

[16] For a recent taxonomy of explanatory virtues, and for references, see Keas (2018). See also Beebe (2009); Douven (2017), sec. 2; Harman (1965); Lipton (2004), Chs. 7 and 8; Lycan (2002), sec. 3; McMullin (2008); and Psillos (2002).

[17] There are 2047 sets of one or more of (a)-(k). [18] See, for example, Blanchard (2018), Lange (2004), and Patrick (2018). There are ways of understanding unification on which our assumption is false. This is true, for example, of Friedman's (1974) account of unification, since it's restricted to propositions about laws of nature. For discussion of Friedman's ideas about unification, and, more generally, his unificationist theory of explanation, see Roche and Sober (2017a).

[19] We are assuming, as is natural, that for any rival available potential explanations H and H^* (of O), if (i) H is superior to H^* in at least one of v_1, v_2, …, and v_n and (ii) H isn't inferior to H^* in any of v_1, v_2, …, and v_n, then H is better overall in terms of v_1, v_2, …, and v_n than H^*.

[20] There's a variant of IBE** where the fourth condition in the antecedent is the condition that H's overall score in terms of v_1, v_2, …, and v_n is *significantly greater than* the overall scores of the other potential explanations in question [see Lycan (2002), p. 414 for discussion]. If this condition can be met even though *HIGH* is false, then there can be cases where IBE** and this variant of it come apart. Even so, what we say below about the former carries over to the latter.

[21] See Douven (2017), sec. 2, for further discussion and references.

[22] We suspect that the genetics example in Roche and Sober (2017b) can be adapted to show this.

[23] See Roche (2018) for related discussion on parsimony and background information.

References

BEEBE, J. (2009), "The Abductivist Reply to Skepticism"; *Philosophy and Phenomenological Research*, 79, pp. 605-636.

BLANCHARD, T. (2018), "Bayesianism and Explanatory Unification: A Compatibilist Account"; *Philosophy of Science*, 85, 682-703.

CABRERA, F. (2017), "Can there be a Bayesian Explanationism? On the Prospects of a Productive Partnership"; *Synthese*, 194, pp. 1245-1272.

CLIMENHAGA, N. (2017), "How Explanation Guides Confirmation"; *Philosophy of Science*, 84, pp. 359-368.
DOUVEN, I. (2017), "Abduction"; in E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (summer ed.). URL = <http://plato.stanford.edu/archives/sum2017/entries/abduction/>.
ELLIOTT, K., and D. WILLMES (2013), "Cognitive Attitudes and Values in Science"; *Philosophy of Science*, 80, pp. 807-817.
FRIEDMAN, M. (1974), "Explanation and Scientific Understanding"; *Journal of Philosophy*, 71, pp. 5-19.
HARMAN, G. (1965), "The Inference to the Best Explanation"; *Philosophical Review*, 74, pp. 88-95.
HEMPEL, C. (1965), "Studies in the Logic of Confirmation"; in C. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*; New York: Free Press, pp. 3-46.
KEAS, M. (2018), "Systematizing the Theoretical Virtues"; *Synthese*, 195, pp. 2761-2793.
LANGE, M. (2004), "Bayesianism and Unification: A reply to Wayne Myrvold"; *Philosophy of Science*, 71, pp. 205-215.
— (2017), "The Evidential Relevance of Explanatoriness: A Reply to Roche and Sober"; *Analysis*, 77, pp. 303-312.
LIPTON, P. (2004), *Inference to the Best Explanation* (2nd ed.); London: Routledge.
LYCAN, W. (2002), "Explanation and Epistemology"; in P. Moser (Ed.), *The Oxford Handbook of Epistemology*; Oxford: Oxford University Press, pp. 408-433.
MCCAIN, K., and T. POSTON (2014), "Why Explanatoriness is Evidentially Relevant"; *Thought*, 3, pp. 145-153.
— (2017), "The Evidential Impact of Explanatory Considerations"; in K. McCain and T. Poston (Eds.), *Best Explanations: New Essays on Inference to the Best Explanation*; Oxford: Oxford University Press, pp. 121-129.
MCMULLIN, E. (2008), "The Virtues of Good Theories"; in S. Psillos and M. Curd (Eds.), *The Routledge Companion to Philosophy of Science*; London: Routledge, pp. 498-508.
OKASHA, S. (2000), "Van Fraassen's Critique of Inference to the Best Explanation"; *Studies in the History and. Philosophy of Science*, 31, pp. 691-710.
PATRICK, K. (2018), "Unity as an Epistemic Virtue"; *Erkenntnis*, 83, pp. 983-1002.
PSILLOS, S. (2002), "Simply the Best: A Case for Abduction"; in A. Kakas and F. Sadri (Eds.), *Computational Logic: Logic Programming and Beyond*; Berlin: Springer-Verlag, pp. 605-625.
ROCHE, W. (2017), "Explanation, Confirmation, and Hempel's Paradox"; in K. McCain and T. Poston (Eds.), *Best Explanations: New Essays on Inference to the Best Explanation*; Oxford: Oxford University Press, pp. 219-241.
— (2018), "The Perils of Parsimony"; *Journal of Philosophy*, 115, pp. 485-505.

ROCHE, W., and E. SOBER (2013), "Explanatoriness is Evidentially Irrelevant, or Inference to the Best Explanation Meets Bayesian Confirmation Theory"; *Analysis*, 73, pp. 659-668.
— (2014), "Explanatoriness and Evidence: A Reply to McCain and Poston"; *Thought*, 3, pp. 193-199.
— (2017a), "Explanation = Unification? A New Criticism of Friedman's Theory and a Reply to an Old One"; *Philosophy of Science*, 84, pp. 391-413.
— (2017b), "Is Explanatoriness a Guide to Confirmation? A Reply to Climenhaga"; *Journal for General Philosophy of Science*, 48, pp. 581-590.
SOBER, E. (2012), "Coincidences and How to Reason About Them"; *European Philosophy of Science Association Proceedings*, 1, pp. 355-374.
— (2015), *Ockham's Razors – A User's Manual*; Cambridge: Cambridge University Press.
VAN FRAASSEN, B. (1989), *Laws and Symmetry*; Oxford: Oxford University Press.
WILLIAMSON, T. (2016), "Abductive Philosophy"; *Philosophical Forum*, 47, pp. 263-280.

teorema
Vol. XXXVIII/3, 2019, pp. 143-162
ISSN: 0210-1602
[BIBLID 0210-1602 (2019) 38:3; pp. 143-162]

Conjunctive Explanations and Inference to the Best Explanation

Jonah N. Schupbach

RESUMEN

Este artículo, de acuerdo con el tema de la sección monográfica, examina un modo en el que las discusiones sobre la naturaleza de la explicación científica pueden ser relevantes para formular adecuadamente la inferencia explicativa. La inferencia hacia la mejor explicación (IME) aconseja al agente razonador inferir una y solo una explicación. Esta recomendación parece convertirse en una limitación cuando abordamos "explicaciones conjuntivas", esto es, explicaciones distintas que, sin embargo, son explicativamente mejores cuando van juntas que cuando van separadas. Para hacer frente a ello, los "explicacionistas" matizan su formulación de la IME estipulando que esta forma de inferencia sólo se pronuncia entre hipótesis rivales. No obstante, una consideración más atenta de la naturaleza de tal competición revela problemas para esta tesis matizada. Según la explicación más común de la competición entre hipótesis, dicha tesis constriñe artificial y radicalmente el dominio de aplicación de la IME. Desde una acepción más sutil, y más reciente, de lo que cuenta como 'hipótesis rivales' se muestra que, para abordar las explicaciones conjuntivas, la tesis matizada es prescindible. A la luz de estos resultados, sugiero una estrategia diferente para acomodar las explicaciones conjuntivas. En vez de modificar la forma de la IME, planteo un modo nuevo de pensar la estructura del lote de hipótesis que cae bajo la consideración de IME.

PALABRAS CLAVE: *inferencia hacia la mejor explicación, explicación conjuntiva, pluralismo explicativo, competición entre hipótesis, razonamiento explicativo*.

ABSTRACT

Fitting with the theme of the special issue, this paper explores one way in which discussions of the nature of scientific explanation can inform the proper statement of explanatory inference. Inference to the Best Explanation (IBE) advises reasoners to infer exactly one explanation. This uniqueness claim apparently binds us when it comes to "conjunctive explanations," distinct explanations that are nonetheless explanatorily better together than apart. To confront this worry, explanationists qualify their statement of IBE, stipulating that this inference form only adjudicates between competing hypotheses. However, a closer look into the nature of competition reveals problems for this qualified account. Given the most common explication of competition, this qualification artificially and radically constrains IBE's domain of applicability. Using a more subtle, recent explication of competition, this qualification no longer provides a compelling treatment of conjunctive explanations. In light of these results, I suggest a different strategy for accommodating

conjunctive explanations. Instead of modifying the form of IBE, I suggest a new way of thinking about the structure of IBE's lot of considered hypotheses.

KEYWORDS: *Inference to the Best Explanation, Conjunctive Explanation, Explanatory Pluralism, Hypothesis Competition, Explanatory Reasoning*

I. THE CHALLENGE OF CONJUNCTIVE EXPLANATION

Sometimes two explanations are better than one. This may happen, for example, in cases of "explanatory pluralism" when theories each do qualitatively different explanatory work. An object's existence can be explained either by referring to its causes or its function—cf. Wright (1976). One hypothesis may explain an event by telling us a causal-mechanical story leading up to the event, while another may perhaps explain the same event by referring to a nomic regularity that the event instantiates — cf. Salmon's (1981), (2001), "friendly physicist" example. In such cases, accepting a plurality of explanations provides us with a richer understanding of the explanandum. More generally, several explanations are better than one just when the explanatory benefits of accepting them all outweigh the costs (in complexity and otherwise). In such cases, I will say that the distinct potential explanations in question are "conjunctive", and I will refer to the above observation as the phenomenon of "conjunctive explanation."

The observation that there exist conjunctive explanations might seem mundane. But conjunctive explanations apparently spell trouble for Inference to the Best Explanation (IBE), at least in its simplest formulations. In an exchange with Peter Lipton, Wesley Salmon criticizes IBE precisely for mishandling conjunctive explanations. The problem, as Salmon (2001), p. 67, presents it, is with IBE's "uniqueness claim": "The phrase, 'inference to *the best* explanation,' involves a uniqueness claim that is difficult to justify." IBE may bar us from inferring truths, gaining richer understanding, accepting otherwise appealing explanatory hypotheses, etc. in conjunctive explanation cases, since it ostensibly mandates that we only choose the single best explanation.

In response to this challenge of conjunctive explanation, Lipton offers what is now widely regarded as a necessary qualification on IBE: "[IBE] is meant to tell us something about how we choose between *competing* explanations: we are to choose the best of these. But among compatible explanations we need not choose" [Lipton (2001), p. 104; cf. Ibid. (2004), pp. 62-63]. Call this move "Lipton's hedge." The suggestion is that Salmon's criticism holds no sway against IBE once the latter is properly

qualified. This is because IBE so hedged does not even attempt to adjudicate between non-competitors, the working presumption being that conjunctive explanations cannot compete with one another.

A proper evaluation of Lipton's hedge must take into account what it means for potential explanations to *compete* with one another. This question results in a balancing act that Lipton and his followers must manage. To give IBE a formulation that does not force us to choose between conjunctive explanations, Lipton's hedge restricts IBE's domain of applicability to cases in which the alternative explanations compete. In aiming to rule out the problematic cases (of conjunctive explanation) and only these cases, this competition qualification is susceptible to two potential errors. On the one hand, the account of competition might be too strong, ruling out more than the problematic cases and overly restricting IBE's domain of applicability. The problem in this case would be that IBE does guide us in reasoning between explanations not considered to be competitors according to such a strong account. On the other hand, to the extent that the proposed account of competition is too weak (not ruling out all of the problematic cases), the challenge of conjunctive explanation remains. In this case, the problem would be that conjunctive explanations can compete in the salient, weaker sense. And we would not want IBE to force us to choose between such conjunctive explanations in such a case anymore than in cases where conjunctive explanations do not compete. The hope for Lipton and his followers then is that there is a plausible explication of competition that strikes the right balance in order to rule out all and only the problematic cases of conjunctive explanation.

In this paper, I will argue that no candidate account of competition manages to strike the desired balance. A stronger "all-out" reading of competition allows one to bypass the challenge of conjunctive explanation, but only at the expense of absurdly restricting IBE's domain of applicability. A weaker, more generally palatable explication of competition fails to meet the challenge of conjunctive explanation. This failure of contemporary accounts of competition to strike the desired balance motivates a reassessment of Lipton's hedge. I argue that Lipton's hedge was never needed in the first place by suggesting an alternative way of responding to the challenge of conjunctive explanation. The upshot is a defense of IBE, as traditionally formulated, with Lipton's hedge completely trimmed.

II All-Out Competition and the Ubiquity of IBE

Lipton's hedge requires that the explanations being compared in any instance of IBE compete with one another. But what exactly does it take for explanatory hypotheses to compete? In the above quote, Lipton goes along with a popular trend in philosophy of science and assumes (perhaps only for the sake of simplicity) that hypotheses compete only when they are incompatible. Potential explanations may be incompatible by virtue of being directly inconsistent. Mutually exclusive descriptions of flag pole height and position of the sun constitute incompatible, competing potential explanations of the length of the pole's shadow. But potential explanations may also be rendered incompatible by the evidence they aim to explain. The hypotheses that John committed the robbery and that Bill committed the robbery are compatible, but they may be rendered incompatible by evidence showing that there could only possibly have been one robber acting in the case. Either way, when explanatory hypotheses compete in the extreme sense of being incompatible, they cannot possibly be true together. Accepting either potential explanation accordingly provides us with a decisive case for rejecting the other. Call this extreme notion of competition "all-out competition"—to be contrasted with a less extreme sense of competition in the next section.

With respect to the challenge of conjunctive explanation, it is easy to see the appeal of explicating competition as all-out. Lipton's hedge provides a convincing response to this challenge, at least in part, *because* it invokes the extreme reading of competition. If IBE only adjudicates between competing hypotheses, *and competition amounts to incompatibility*, then IBE should manifestly require us to choose at most one hypothesis. Conjunctive explanations pose no challenge at all to IBE, because there is not a situation to which IBE applies in which we ever somehow miss out by only inferring one explanation. The underlying assumption here is that incompatible explanations cannot constitute conjunctive explanations, that it is never explanatorily better to accept an unsatisfiable conjunction of hypotheses. And that surely seems right.

So, Lipton's hedge, when teamed with the all-out notion of competition, rules out all of the problematic cases of conjunctive explanation. But the natural follow-up question to ask is whether it rules out *only* those cases. Are there cases in which IBE helpfully guides us to infer between potential explanations that are not all-out competitors? If so, then Lipton's hedge overly restricts IBE's domain of applicability. This question is all the more pressing once one recalls a familiar point commonly made by

Lipton and explanationists more generally: that IBE is ubiquitous, being an extremely useful inference form with an impressively expansive domain of applicability [Lipton (2004), pp. 1-2]. Does IBE lose its intuitively expansive reach in light of Lipton's hedge?

In fact, the answer is a troublingly emphatic and obvious yes. Many (indeed, plausibly *most*) canonical instances of IBE compare potential explanations that are compatible with one another. Indeed, this is true of nearly all of Lipton's own examples of IBE at work.

Lipton's foremost example is the Semmelweis case. Working in the maternity division of the General Hospital in Vienna in the 1840s, Ignaz Semmelweis struggled to explain why three times more women in the first maternity ward were dying of "childbed fever" than in the second ward of the same hospital. Ward one was staffed by medical students, whereas ward two was overseen entirely by midwives. The potential explanations considered and tested by Semmelweis included the following:

> H_1. The midwives in ward two encouraged women to give birth on their sides, whereas the medical students had women give birth on their backs. The latter birthing position somehow promotes childbed fever.
>
> H_2. A priest was more often seen in ward one on his way to administering last rites to dying patients. This has a pernicious psychological influence on birthing women, which subsequently promotes childbed fever.
>
> H_3. Unlike the midwives, medical students in ward two were routinely conducting autopsies. Childbed fever is promoted by an infection of "cadaveric matter" from the hands of such students.

Eventually, Semmelweis famously inferred H_3 as the best explanation of his accumulating evidence.

Lipton takes this to be a paradigmatic example of IBE at work, regularly drawing upon this example to develop his account of the nature and power of explanatory inference. What Lipton does not seem to recognize is that his use of this example clashes with his response to the challenge of conjunctive explanation. For if IBE is only meant to "tell us something about how we choose between *competing* explanations," and competition is cashed out as "incompatibility" then Lipton's favorite example of IBE at work is in fact not an example of IBE at all. After all, H_1, H_2, and H_3 are

manifestly compatible. Any combination of these hypotheses could have been true prior or posterior to considerations of Semmelweis's collected evidence. If Lipton is right that IBE guides reasoning in the Semmelweis case, then he is wrong that IBE only adjudicates between incompatible explanations.

Upon further reflection, it is plausible that such cases — in which IBE adjudicates between compatible alternatives — are more the norm than the exception. Explanationists commonly draw instances of IBE from such contexts as detective work, historical science, medical diagnosis, and diagnostic settings more generally (e.g., diagnosing the failure of a car engine from the observable "symptoms"). Unless explanatory hypotheses from such contexts are intentionally framed so as to exclude one another, it is straightforward to think of such cases as usually comparing compatible potential explanations.

Let us emphasize the point by glossing over a couple more examples. Lipton (2001), pp. 95-96, writes, "When a detective infers that it was Moriarty who committed the crime, he does so because this hypothesis would best explain the fingerprints, blood stains and other forensic evidence [...] Moriarty's guilt would provide a better explanation of the evidence than would anyone else's." Lipton uses this example to demonstrate the fallibilistic nature of IBE; some other potential explanation than the best may turn out to be the actual explanation. But this example can just as well be used to demonstrate the fact that it is possible, in very typical cases, for more than one of the potential explanations compared by IBE to be an actual explanation —i.e., that more than one of these turn out true, given their joint satisfiability. In this particular example, the crime may after all have been committed by Moriarty and someone else.

Scientists debate the explanation of the mass extinction at the Cretaceous- Paleogene (K-Pg) boundary — which included the mass extinction of the dinosaurs about 66 million years ago. Common potential explanations include bolide impact, massive outbreaks of volcanic activity, climate change, continental drift and sea level regression, and so on. Notably, while it is perhaps more common to look for a "smoking gun" amongst these alternatives [Cleland (2011), p. 554], many scientists opt instead for inferring some combination of these alternatives —e.g., Archibald et al. (2010). The scientists involved in this debate do not think of the various alternatives as incompatible; instead they argue either that one alone suffices as the best explanation of the evidence, or that more than one of the compatible alternatives should indeed be inferred —see Schupbach and Glass (2017) for further discussion of this example. Accordingly, this is

another case that the explanationist will be keen to describe as potentially involving IBE, despite the fact it involves reasoning between recognizably compatible alternatives.

In sum, if Lipton is right that IBE only provides a model of inference between incompatible explanations, then he is wrong to think of the above (and any number of other such examples) as instances of IBE. But he is not wrong about that; these are paradigm examples of IBE at work, instances of IBE if anything is. And so, Lipton must be wrong in thinking that IBE only properly provides a model of inference between incompatible explanations.

III. Lipton's Hedge Refined

Lipton's hedge, when combined with the all-out explication of competition provides a convincing response to the challenge of conjunctive explanation, but only by absurdly restricting IBE's domain of applicability. To pin the blame immediately on Lipton's hedge itself however would be too quick. It may be that Lipton is correct to think that IBE provides a model of inference only between competing explanations, and that he only goes astray when he explicates competition as all-out. That is, maybe the problem is not with the hedge per se, but with Lipton's identification of competition with incompatibility.

Recent work by Schupbach and Glass (S&G) proves helpful here. S&G (2017) argue that competition is not plausibly explicated as mutual exclusivity. As they claim and demonstrate through examples, it is simply too easy to think of actual cases from various contexts of human reasoning in which recognizably compatible hypotheses are thought of and inferentially treated as competitors. S&G thus offer a probabilistic explication of what it takes for hypotheses to compete with one another in light of a body of evidence (as well as a formal measure of the degree to which hypotheses compete with one another apropos some body of evidence). Most importantly for our present purposes, their explication allows for cases in which compatible hypotheses nonetheless compete.

S&G motivate their account of competition by suggesting that a "mutual exclusivity" account falls short of a satisfactory, general account of competition in no less than two ways. The first is that it implies that competition is all-or-nothing (this objection applies equally well to the more general incompatibility explication that Lipton invokes). But hypotheses may compete

by disconfirming each other to varying degrees without fully precluding one another. Consider the following simple variation on Lipton's detective case:

> *Moriarty and Smith, v1.*
>
> Moriarty and Smith are both house burglars working in the same area, but they are also sworn enemies who are extremely unlikely to ever burgle together. Bob reports to the police that his front window has been broken and that all of his valuable belongings are missing from the house.

A detective investigating the case may rightly view H_S*: Smith burgled the house* and H_M*: Moriarty burgled the house* as distinct (but compatible) explanations of the reported evidence. Moreover, given the background information about Moriarty and Smith's relationship, the detective might rightly view these potential explanations as competing *to the extent that* H_S and H_M disconfirm each other. To make sense of such cases, S&G require that hypothesis competition be accounted for gradationally, as a matter of degree.

Noting that hypotheses may compete to varying degrees helps shed light on the nature of competition in cases where hypotheses disconfirm one another, but to some less than maximal extent—as in Moriarty and Smith, v1. However, in many actual cases, competing hypotheses may not even disconfirm one another directly (i.e., prior to consideration of the explanandum). This is plausibly the case in at least two of the examples we have described above. In the Semmelweis example, it is not at all clear that birthing position's having an influence on childbed fever rates would somehow lower the probability of a priest's presence also having such an influence. This is unclear, but it is clear that these hypotheses compete in this case. The nature of their competition with one another must come down to something other than a disconfirmatory (probability-lowering) relation between them then. Similarly, in the K-Pg extinction case, far from disconfirming one another (does volcanic activity decrease the chance of bolide impact?!), some of these historical hypotheses may even *confirm* one another to some extent. Nonetheless, many scientists persist in viewing these as competitors.

These observations point to another shortcoming with the mutual exclusivity account. It neglects an important sense in which hypotheses can compete —indirectly, via the relevant body of evidence *E*. Consider another variation on Lipton's detective:

Moriarty and Smith, v2.

Moriarty and Smith often rob houses together in well-informed, carefully planned ways. They would never knowingly rob the police chief's house. Moriarty knows that 123 Main is Bob's house but doesn't know that Bob is the police chief; Smith knows Bob is police chief but doesn't know where he lives. E includes this information as well as Bob's recent discovery that his front window has been broken and that all of his valuable belongings are missing from the house.

While H_S and H_M may confirm one another in general, the detective may rightly consider them as competitors in this particular case; relative to *this* body of evidence E, it is unlikely they collaborated.

In cases of distinctively indirect competition, adopting either hypothesis undermines any support E provides for the other. This may be because the evidence itself places otherwise mutually confirming (or independent) hypotheses in a relation of mutual disconfirmation with one another. Or it may be because the evidence is fully accounted for by one of the hypotheses alone, in which case no support from E accrues any longer for the other hypothesis. In order to make sense of competition in cases like this, S&G require that an appropriate account of competition accommodate two distinct paths to hypothesis competition: a direct path and an indirect path via the evidence.

S&G develop a confirmation-theoretic measure of the "net" degree (taking into account both the direct and indirect paths) to which hypotheses H and H' compete with one another relative to a particular body of evidence E. The full official statement of this measure is somewhat complex, but S&G prove a theorem that simplifies matters. They show that net degree of competition is formally equivalent to average degree of disconfirmation conditional on E. Using the log-likelihood measure of confirmation, the degree to which a proposition φ disconfirms another ψ (conditional on a proposition χ) is measured as the degree to which φ confirms $\neg\psi$ (conditional on χ):

$$C_l(\varphi, \neg\psi \mid \chi) = \log \frac{P(\varphi \mid \neg\psi \wedge \chi)}{P(\varphi \mid \psi \wedge \chi)}$$

Thus, the net degree to which hypotheses H and H' compete with one another relative to a particular body of evidence E can be represented as follows:

$$Comp(H', H/E) = [C_l(H, \neg H' \mid E) + C_l(H', \neg H \mid E)] / 2$$

$$= \left[\log \frac{P(H \mid \neg H' \wedge E)}{P(H \mid H' \wedge E)} + \log \frac{P(H' \mid \neg H \wedge E)}{P(H' \mid H \wedge E)}\right] / 2.$$

Moreover, S&G explicate qualitative judgments of hypothesis competition as positive degree of net competition; i.e., the judgment that H and H' compete with one another relative to E is explicated using the inequality $Comp(H', H/E) > 0$. Because net competition can (as above) be represented as average degree of disconfirmation, we may equally well explicate the judgment that H and H' compete with one another relative to E as an assessed positive degree of disconfirmation, using the inequality $C_l(H, \neg H' \mid E) > 0$ (or using $C_l(H', \neg H \mid E) > 0$, since these imply one another). Or we may represent this judgment probabilistically as $P(H \wedge H' \mid E) < P(H \mid E) \times P(H' \mid E)$. Finally—and this will prove the most useful statement of all—Glass (2012), Theorem 1, proves that the following inequality is yet another equivalent condition for qualitative competition:

$$\log \left[\frac{P(E \mid H \wedge H')\, P(E \mid \neg H \wedge \neg H')}{P(E \mid H \wedge \neg H')\, P(E \mid \neg H \wedge H')}\right] + \log \left[\frac{P(H \mid H')\, P(\neg H \mid \neg H')}{P(H \mid \neg H')\, P(\neg H \mid H')}\right] < 0. \qquad (1)$$

If we accept Lipton's hedge, along with S&G's explication of competition, then IBE is no longer so absurdly restricted in scope. The resulting qualified inference form again guides us in Lipton's paradigmatic cases. Recall that in both the Semmelweis case and the K-Pg extinction case (and potentially also in the detective case, depending on the details), we take the potential explanations involved to be at once perfectly compatible apart from the explanandum but to compete relative to this evidence. The sense in which they compete is indirect. But additionally, note that they compete indirectly not because E introduces an incompatibility between them, but simply because any one of these explanations arguably suffices on its own to account for E. In light of E, reason compels us to choose between these potential explanations for the simple reason that accepting more than one would arguably be epistemically overblown.

S&G's account implies that this can be a genuine source of competition; this is easiest to see in terms of the last formal statement of competition.

Condition (1) involves the sum of two terms, which can roughly be seen as respectively explicating the notions of direct and indirect degrees of competition.[1] Examining these two summands carefully sheds formal light on some exact paths to competition. Most relevant to our current purposes, note that the second summand—corresponding to the notion of direct competition between H and H'—is effectively a wash when H and H' are approximately independent of each other. Focusing on the first summand in such cases then, H and H' will be deemed competitors with respect to E when the denominator is greater than the numerator, $P(E|H \wedge \neg H') \times P(E|\neg H \wedge H') > P(E|H \wedge H') \times P(E|\neg H \wedge \neg H')$.

Notably, this inequality can easily attain in cases where one, but really only one, of the hypotheses is needed to account for the evidence. When this is true, $P(E|H \wedge \neg H')$ and $P(E|\neg H \wedge H')$ will be quite high. In fact, even if both hypotheses account for the evidence somewhat better than either individually, it will still be the case that $P(E|H \wedge \neg H') \approx P(E|H \wedge H')$ and $P(E|\neg H \wedge H') \approx P(E|H \wedge H')$. By contrast, since one or the other hypothesis is needed to account for E in the envisioned scenario, $P(E|\neg H \wedge \neg H') \ll 1$. The upshot is that H and H' compete with respect to E in such a scenario, since $P(E|H \wedge \neg H') \times P(E|\neg H \wedge H') \gg P(E|H \wedge H') \times P(E|\neg H \wedge \neg H')$. This is a prime example of a scenario in which two hypotheses can strongly compete with one another (relative to some E) despite the fact that they are otherwise entirely compatible with—possibly even mutually supportive of—one another!

Qualified in this way, IBE requires us to choose between explanatory hypotheses so long as $Comp(H', H/E) > 0$. The new question is whether the proper balance is now struck. IBE is not absurdly restricted in its domain of applicability, but is it overly restrictive in what it allows us to infer? Unfortunately, it turns out that the challenge of conjunctive explanation comes back for revenge at this point; there are hosts of cases in which it is problematic for IBE to require us to choose between hypotheses that compete on S&G's generalized account.

IV. The Challenge of Conjunctive Explanation 2.0

Lipton's hedge, combined with the all-out reading of competition, restricted IBE to cases in which it indeed is always appropriate to infer at most one competing hypothesis. But IBE, so restricted, lost much of its

intuitively vast domain of applicability. S&G's construal of competition provides IBE with a wider, more intuitive domain of applicability. But it reopens the door to cases in which it is undesirable for IBE to keep us from inferring more than one of the alternative hypotheses.

To see this, first consider a final variation on Lipton's detective case:

> *Moriarty and Smith, v3*.
>
> Moriarty and Smith are far and away the busiest and most notorious house burglars working in a town. They are also sworn enemies who are extremely unlikely to ever burgle together. Moriarty always leaves an "M" pendant at the scenes of his crimes, while Smith's trademark is to leave the water running into a plugged sink, flooding the houses he robs. Bob reports to the police that his valuables are missing from his now-flooded house, where he has also discovered the familiar "M" pendant.

A detective examining this case may rightly be strongly compelled to accept both H_S: *Smith burgled the house* and H_M: *Moriarty burgled the house*. Given the evidence of the case E, this certainly seems to be a better option than accepting H_S or H_M alone. Two explanations are better than one here, making this a case of conjunctive explanation. Importantly however, these explanations are not better together because they mutually support one another but because they are both separately supported by the evidence. E strongly confirms both H_S and competitor H_M individually, but it does nothing to unify them. H_S and H_M disconfirm one another unconditionally and conditional on E.

Accordingly, S&G's account validates the judgment that these hypotheses compete. Applying condition (1) to this case, we have the following qualitative criterion for competition between H_S and H_M:

$$\log\left[\frac{P(E|H_s \wedge H_M)\, P(E|\neg H_s \wedge \neg H_M)}{P(E|H_s \wedge \neg H_M)\, P(E|\neg H_s \wedge H_M)}\right] + \log\left[\frac{P(H_s|H_M)P(\neg H_s|\neg H_M)}{P(H_s|\neg H_M)P(\neg H_s|H_M)}\right] < 0.$$

Regarding the first summand, this case is meant to inspire the following judgments. The evidence is by far best accounted for by the conjunction $H_S \wedge H_M$. However, it is partially accounted for (made somewhat probable) by either hypothesis taken alone (conjoined with the negation of the hypothesis). Since these two criminals are far and away the most active burglars working in the area (and given the presence in this case of their trademarks), the evidence remains very unlikely indeed if neither criminal was at work.

Probabilistically, we may explicate these judgments simply using the following inequalities: $P(E|H_S \wedge H_M) > P(E|H_S \wedge \neg H_M)$ $[P(E|\neg H_S \wedge H_M)] > P(E|\neg H_S \wedge \neg H_M)$. These inequalities fail to determine whether the first summand above is positive or negative, but they plausibly suggest that the term is negative or at least not strongly positive (Figure 1).

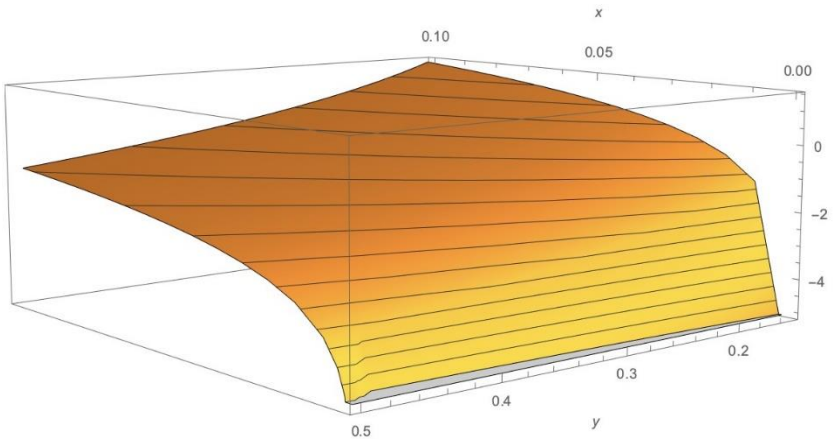


Figure 1: Ln $[(P(E|H_S \wedge H_M)P(E|\neg H_S \wedge \neg H_M))/(P(E|H_S \wedge \neg H_M)P(E|\neg H_S \wedge H_M))]$ plotted as a function of $y = P(E|\neg H_S \wedge H_M) = P(E|H_S \wedge \neg H_M)$ and $x = P(E|\neg H_S \wedge \neg H_M)$—with $P(E|H_S \wedge H_M)$ fixed at .95. Values in the displayed range tend to be negative (thus contributing to the fact that H_S and H_M compete), and they are never strongly positive.

Now consider the second summand. Given that they are sworn enemies, "extremely unlikely to ever burgle together," each criminal is much more likely to have burgled a house generally if the other suspect did not. The assumption that either criminal did in fact burgle a house thus makes it more likely that the other did not (than did). Alternatively, given the same considerations along with the suggestion that these burglars are collectively (though not typically conjointly) responsible for the vast majority of burglaries in the area, the assumption that either criminal did not burgle the house makes it more probable that the other criminal did (than did not). Probabilistically, $P(\neg H_S \,|H_M) \gg P(H_S \,|H_M)$ and $P(H_S \,|\neg H_M) > P(\neg H_S \,|\neg H_M)$. The upshot is that the second summand cannot but be negative, and the details of the case suggest that it is substantially so. It is

thus primarily because of this aspect of the case that we may safely conclude, on S&G's account, that H_S and H_M compete.

The fact that one would want to accept both H_S and H_M in light of E is evidently not because they do not compete. That is, this is not intended to be a counterexample to S&G's account of competition. It is right to think of these hypotheses as being (potentially strongly) in competition with one another with respect to this evidence for the simple reason that they (potentially strongly) disconfirm one another—i.e., lower each other 's respective probabilities—in the light of this evidence. Any justification accruing to $H_S \wedge H_M$ in this case is accounted for solely by way of E's providing strong support for H_S and H_M individually, and this despite the fact that each hypothesis goes some non-negligible way to rebutting the other hypothesis conditional on E (and unconditionally). The example thus serves to show that there are cases in which it would be explanatorily better to accept multiple, *competing* explanations of E; competing hypotheses can provide conjunctive explanations.

Guided by the above example, it is not difficult to characterize formally an entire family of such examples in which it can be explanatorily better to accept multiple competing explanations than to choose between them. The following jointly satisfiable[2], probabilistically explicable conditions are characteristic of such examples:

C1. H and H' compete with one another with respect to E: $Comp(H', H/E) > 0$, and so $P(H \wedge H' \,|E) < P(H|E)P(H' \,|E)$.

C2. E confirms each hypothesis individually, conditional on the other: $P(H|E \wedge H') > P(H|\,H')$ and $P(H' \,|E \wedge H) > P(H' \,|H)$.

C3. H and H' together account for the evidence better than either does individually: $P(E|H \wedge H') > P(E|H \wedge \neg H')$ and $P(E|H \wedge H') > P(E|\neg H \wedge H')$.

H and H' may plausibly provide conjunctive explanations in such cases, at least when E is *explanatorily* better accounted for by the conjunction $H \wedge H'$ than by either hypothesis taken alone. That is, allowing that the inequality in likelihoods used to explicate C3 may be achieved apart from explanatory considerations and contexts, still certain explanatory virtues plausibly have their logical effect via just such an inequality in likelihoods—e.g., power [Schupbach and Sprenger (2011); Schupbach, (2017)]. But then the problem of conjunctive explanation is once again a serious

problem, even for Lipton's hedged version of IBE. H and H' compete with one another *apropos* E, but it can be explanatorily better to accept both than to choose between them. IBE is problematic if it forces us to choose between them. We want to be able to infer the conjunction $H \wedge H'$ as "the" best explanation of E in cases like this, even if they compete.

The skeptical reader might wonder at this point whether it actually might be *epistemically* worse to do what is explanatorily better in these cases. That is, one might wonder whether the conjunction $H \wedge H'$ is overall worse off than $H \wedge \neg H'$ or $\neg H \wedge H'$, even if it happens to account for E better in the sense of making E more likely. If the answer is 'yes', then perhaps we really do want IBE to force us to choose between H and H' in such cases, even if this may mean not doing what is explanatorily best (though IBE would surely be in need of a new name in this case). Thus, it is very much worth highlighting the subset of cases demonstrating the consistency of C1, C2, and C3 with the following:

> C4. H and H' are overall more plausible together in light of the evidence than either is alone (i.e., conjoined with the negation of the other): $P(H \wedge H' \mid E) > P(H \wedge \neg H' \mid E)$ and $P(H \wedge H' \mid E) > P(\neg H \wedge H' \mid E)$.

Cases satisfying C1-C4 are ones in which $H \wedge H'$ may have not only explanatory considerations, but overall net epistemic considerations in its favor—at least assuming a Bayesian perspective from which net epistemic value is associated with posterior probability. An inference rule is surely problematic if it precludes us from even considering the possibility of inferring $H \wedge H'$ in such cases. But that is just what Lipton's hedge does, when combined with S&G's generalized explication of competition.

V. Trimming Lipton's Hedge

Let's take stock. It can be explanatorily best to accept multiple distinct explanatory hypotheses. Such conjunctive explanations seem to pose a serious challenge to IBE, since this inference form ostensibly guides us to infer the single best explanation of our explanandum. Responding to this challenge, Lipton's idea was to qualify IBE by restricting it to cases in which we are comparing competing explanations. However, given Lipton's own

interpretation of the notion of competition, this qualification greatly *over*-restricts IBE's domain of applicability, absurdly barring from the ranks many (perhaps most) of IBE's canonical instances. Thankfully, a subtler, generalized explication of competition does not seem to lead to this consequence; however, it results in a hedged version of IBE that again fails to meet the challenge of conjunctive explanation.

Explanations that compete, according to this generalized account, thereby each provide some (possibly strong) reason against accepting the other(s). However, we have highlighted the possibility that this reason may be outweighed by the explanandum's relation to each candidate explanation individually so that the conjunction (of competitors) indeed provides the overall best explanatory account of the explanandum. In other words, when explanations compete, there is some reason not to accept both. But what ultimately matters is whether the explanatory payoff outweighs the cost of accepting competitors. If there are net explanatory gains to accepting multiple, distinct explanations, then IBE should allow us to accept multiple explanations, regardless of whether they compete.

Attempts to meet the challenge of conjunctive explanation by hedging IBE such that it only adjudicates between competing explanations do not alas appear to be successful. Does this mean that the challenge of conjunctive explanation is devastating for IBE? That may be the moral of this story if a competition hedge were the only plausible way to respond to this challenge. But I want to suggest that Lipton's hedge was ultimately a distraction to IBE research. The final consideration offered in the previous paragraph suggests that the question of whether explanations compete is ultimately orthogonal to the central question of what is explanatorily best. This observation inspires a straightforward response to the challenge of conjunctive explanation that makes no mention of the notion of competition.

Indeed, I suggest the simplest (some would say naive) formulations of IBE can be understood as already providing a response to this challenge. The "challenge" of conjunctive explanation is thus only challenging for those who do not understand the statement of IBE in the way I propose. The central question is how to understand the phrase "the best explanation". It is common to point to one source of ambiguity in this phrase: the various, distinct dimensions along which humans tend to evaluate the explanatory goodness of hypotheses [Schupbach (2017) p. 41]. However, another ambiguity underlies the means by which IBE handles conjunctive explanations. On one reading, "the best explanation" might refer to the single most explanatory hypothesis. On another reading, "the best explanation" might refer to the inference that is explanatorily best, i.e. the most

explanatory conclusion. The phenomenon of conjunctive explanation teaches us that these two readings are indeed importantly different: the explanatorily best conclusion might involve inferring more than one explanatory hypothesis.

Conjunctive explanations pose no threat to IBE when "the best explanation" is understood in the second sense. If it is overall explanatorily best to infer multiple distinct explanations (i.e., explanatory hypotheses), then that is exactly what IBE will guide us to do when "the best explanation" refers to the conclusion or inferential move that is overall explanatorily best.

It is tempting to think that the word "explanation" is, or at least ought to be, always associated with a single explanatory hypothesis (however hypotheses may be individuated) rather than with an explanatory inference, move, or conclusion. It is certainly natural to use it in this way; this is in fact how I have largely been using the word throughout this paper. If one insists that this is the only proper usage, then the phrase "the best explanation" no longer is ambiguous in the required way. However, the same straightforward response to the challenge of conjunctive explanation plausibly remains available to IBE's defenders, with only minor adjustments to the statement of IBE. "Inference to the Best Explanation(s)" or the clunkier "Inference to the Most Explanatory Conclusion" suggest themselves as options.

That said, "explanation" and "the best explanation" do seem to be ambiguous in the required way, allowing for the suggested interpretation. Recall the K-Pg extinction case. As noted, scientists involved in this debate do not think of the various alternative hypotheses as incompatible; but they also do not think of them as necessarily being part of distinct explanations. While some of these scientists hold out hope that one hypothesis alone suffices as the best explanation of the evidence, many contemporary scientists accept more than one of the alternatives as jointly constitutive of *the* fullest explanation. No clash arises here between the inference of multiple explanatory hypotheses and IBE's uniqueness claim, if the uniquely best explanation (i.e., most explanatory conclusion) can take the form of the conjunction of hypotheses under consideration.

This last statement of the point raises one final important question: if "the best explanation" may refer to a conjunction of individual hypotheses, can it also refer to other Boolean combinations of these? More generally, if we understand IBE in the proposed way, then how should we think about the structure of IBE's lot of available, potential explanations? Plausibly, there is no principled ban on the sorts of compound hypotheses

we may consider for inference. Any form that may provide superior explanatory goodness is up for consideration. If conjunctions are allowed, why not disjunctions? Why not indeed? If a detective only knows that a house has been broken into in a neighbourhood where Smith and Moriarty typically work alone, but are collectively responsible for the vast majority of break-ins, the best explanation might in fact be the disjunction $H_S \vee H_M$. To commit to anything more specific than this would be to stretch the explanation beyond what the evidence and background information of the case allow. But if disjunctions are allowed, then why not material conditionals? And can't denials (in the form of negations) sometimes provide potential explanations?

If "the best explanation" is associated with the most explanatory individual hypothesis, then it is natural to think of the lot of considered potential explanations as simply being the set of individual hypotheses. By contrast, the present proposal amounts to thinking of the lot of potential explanations as the set containing these considered hypotheses along with their Boolean combinations. What matters is which combination of considered hypotheses best explains the explanandum, not what logical form the various options take. Regarding cases of conjunctive explanation, this move allows reasoners to infer more than one individual hypothesis when that is the explanatorily best option, all the while leaving IBE's domain of applicability entirely open.

Note that this proposal does not amount to offering a different qualification on IBE in the place of Lipton's hedge. The point of this response is that there is a way of understanding the simplest, unqualified formulations of IBE so that conjunctive explanations do not pose a challenge at all. The so-called challenge of conjunctive explanation does not point to a weakness in IBE properly construed; it rather betrays a misunderstanding of the most defensible statement of IBE. As such, there was never a need for Lipton's hedge in the first place. And that's a good thing, since this paper goes some way toward showing that this hedge ultimately proves unhelpful.

Department of Philosophy
University of Utah
417 CTIHB, 215 S. Central Campus Dr.
Salt Lake City UT 84112, USA
E-mail: jonah.n.schupbach@utah.edu.

ACKNOWLEDGMENTS

I owe special thanks to Vincenzo Crupi, David Glass, and attendees at the 11th MuST Conference in Philosophy of Science on "Models of Explanation" (University of Turin, Italy; June 11-13, 2018) for helpful conversations and feedback regarding this project. I am also very grateful for two sources of financial support for this project. While working on this paper, I was supported by the Charles H. Monson Mid-Career Award administered by the University of Utah's Philosophy Department. The research for this publication was also made possible through the support of a grant (#61115) from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

NOTES

[1] Though only roughly. S&G themselves explicate degrees of direct and indirect competition separately, and they prove that net degree of competition is a simple sum of these two component measures. Unfortunately, there is no neat correspondence between S&G's component measures and the two individual summands in condition (1).

[2] In all cases, satisfiability was established using Fitelson's (2008) decision procedure PrSAT as implemented in his corresponding *Mathematica* package, available at http://fitelson.org/PrSAT/.

REFERENCES

ARCHIBALD ET AL., J. D. (2010), "Cretaceous Extinctions: Multiple Causes; *Science*, 328, p. 973.

CLELAND, C. E. (2011), "Prediction and Explanation in Historical Natural Science"; *British Journal for the Philosophy of Science*, 62, pp. 551–582.

FITELSON, B. (2008), A decision Procedure for Probability Calculus with Applications"; *The Review of Symbolic Logic*, 1(1), pp. 111–125.

GLASS, D. H. (2012), "Can Evidence for Design be Explained Away?"; in Chandler, J. and Harrison, V. S., editors, *Probability in the Philosophy of Religion*, Oxford, Oxford University Press, pp. 79–102.

LIPTON, P. (2001), "Is Explanation a Guide to Inference? A reply to Wesley C. Salmon"; in Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, Dordrecht, Kluwer Academic, pp. 93–120.

— (2004), *Inference to the Best Explanation*; New York, NY, Routledge, 2nd edition.

SALMON, W. C. (1981), "Rational Prediction"; *British Journal for the Philosophy of-Science*, 32(2), pp.115–125.

— (2001), "Explanation and Confirmation: A Bayesian Critique of Inference to the Best Explanation."; in Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, Dordrecht, Kluwer Academic, pp.61-91.

SCHUPBACH, J. N. (2017), "Inference to the Best Explanation, Cleaned Up and Made Respectable"; in McCain, K. and Poston, T., editors, *Best Explanations: New Essays on Inference to the Best Explanation*, Oxford. Oxford University Press, pp. 39–61.

SCHUPBACH, J. N. and GLASS, D. H. (2017), "Hypothesis Competition Beyond Mutual Exclusivity"; *Philosophy of Science*, 84(5), pp. 810–824.

SCHUPBACH, J. N. and SPRENGER, J. (2011), "The Logic of Explanatory Power"; *Philosophy of Science*, 78(1), pp. 105–127.

WRIGHT, L. (1976), *Teleological Explanations*; Berkeley, University of California Press.