Brit. J. Phil. Sci. **60** (2009), 101–133

When Empirical Success Implies Theoretical Reference: A Structural Correspondence Theorem

Gerhard Schurz

ABSTRACT

Starting from a brief recapitulation of the contemporary debate on scientific realism, this paper argues for the following *thesis*: Assume a theory T has been empirically successful in a domain of application A, but was superseded later on by a superior theory T*, which was likewise successful in A but has an arbitrarily different theoretical superstructure. Then under natural conditions T contains certain theoretical expressions, which yielded T's empirical success, such that these T-expressions *correspond* (in A) to certain theoretical expressions of T*, and given T* is true, they *refer indirectly* to the entities denoted by these expressions of T*. The thesis is first motivated by a study of the phlogiston–oxygen example. Then the thesis is proved in the form of a *logical theorem*, and illustrated by further examples. The final sections explain how the correspondence theorem justifies scientific realism and work out the advantages of the suggested account.

doi:10.1093/bjps/axn049

Advance Access published on January 26, 2009

1 Introduction: Pessimistic Meta-induction vs. Structural Correspondence

Can we infer from the empirical success of scientific theories the approximate truth of their theoretical representations of reality? A celebrated argument that supports this inference is the *no-miracles argument*, or NMA for short (cf. Putnam [1975], p. 73). It says, roughly, that without the assumption of realism the empirical success of science would be a sheer miracle. More precisely, the best if not the only reasonable explanation of the continuous empirical success of scientific theories is the realist assumption that their theoretical 'superstructure' is approximately true and, hence, their central theoretical terms refer to real though unobservable constituents of the world.

However, it is doubtful whether the NMA in its unrestricted form is a reliable argument. Theoretical reasons against the reliability of the NMA will be given in Sections 7.1–2. The major empirical reason against the NMA is the *pessimistic meta-induction* argument (put forth by Laudan [1981]), which points to the fact that in the history of scientific theories one can recognize radical changes in the ontology, and hence in the theoretical superstructure of these theories, although there was continuous progress on the level of empirical success. On simple inductive grounds, it is unreasonable to expect that our contemporary theories are the only ones in the history of science that could escape this fate. One rather should expect that the ontology and also the theoretical superstructure of our presently accepted theories will be radically overthrown in the future, and hence can in no way be expected to be approximately true.

While the NMA is the point of support for a *full-fledged scientific realism*, the pessimistic meta-induction supports either an *anti-realist* attitude towards theories, or at least an *instrumentalist* position for which theories can only be regarded as more or less empirically adequate but not as true or false (van Fraassen [1980]).

In the face of Laudan's challenge, Boyd ([1984]) has pointed towards the existence of relations of *correspondence* between successive scientific theories, which reflect that even on the theoretical level *something is preserved* through historical theory change and, thus, has a justified realist interpretation. Laudan replies that there is no evidence for systematic retention of theoretical structures through successive theory-change with growing empirical success. Laudan ([1981], p. 121) gives a much debated list of examples of scientific theories that were strongly successful at their time but have assumed an ontology that is incompatible with or at least divergent from contemporary theories. Laudan concludes that scientists are well advised *not* to follow a retention strategy because this would impede scientific progress (ibid. p. 117, pp. 126f).

According to Worrall's *structural realism* ([1989], p. 122), what are preserved in successive theories are certain *structural relations* between the terms of the theories, but not the *content* of theoretical terms. In the view of other authors, however, the distinction between 'structure' and 'content' is problematic, because all we know about the 'content' of a theoretical entity is specified by the structure of the theory.[1] I agree with Psillos ([1995], p. 44) that we should replace the structure–content thesis by the thesis that not all but certain 'parts' of the theory's content are preserved through theory-change—parts that are explicable by structural relations. Can this thesis be defended? Worrall has supported his preservation-of-structure thesis by one example from Laudan's list, the correspondence relation between Fresnel's and Maxwell's equations describing the relative intensities of incident, reflected and refracted light beams. But as Worrall himself notices ([1989], p. 120), this is an especially pleasant case. Other examples such as the relation between the phlogiston and oxygen theory of combustion, the caloric and the kinetic theory of heat, or between classical and quantum mechanics, are much more difficult to handle. Even for simple cases as the Fresnel–Maxwell case, Laudan points out that one cannot say that the older theory is preserved in the later one, because the older theory assumes entities and mechanisms which according to the contemporary theory do not or even cannot possibly exist ([1981], pp. 127–31).

In my opinion, what is of primary importance is to find good answers to the following three questions:

Question 1: Do correspondence relations between seemingly disparate theories occur in the history of science for *systematic* reasons, and not as mere exceptions?
Question 2: Given that question 1 has a positive answer: are these systematic reasons *objective* ones? More precisely: are they the (necessary) result of the cumulatively increasing empirical *success* of the theories? Can we exclude that they are merely a consequence of the cognitive constitution of the human mind or brain, which prefers certain models over others?
Question 3: Given that question 2 has a positive answer: does this yield an improved way of justifying scientific realism, which can resist the objections against the NMA in its unrestricted version?

The debate in philosophy of science has produced sophisticated versions of realist positions, but it seems to me that a convincing answer to these questions

[1] Cf. (Psillos [1995], pp. 31ff; [1999], p. 155; Papineau [1996], p. 12; Votsis [2007]).

has so far not been achieved. In this paper, I will give a positive answer to these three questions. I will try to establish the following

Thesis: *Assume a theory T has been strongly (empirically) successful in a domain of application A, but was superseded later on by a superior theory T*, which was likewise successful in domain A but has an arbitrarily different theoretical superstructure (or 'ontology'). Then (under natural conditions) T contains certain theoretical terms or expressions that correspond (in A) to certain theoretical expressions of T* and, given T* is true, they refer indirectly to the entities denoted by these expressions of T*.*

I will first support my thesis with the historical example that has been most resistant against the discovery of correspondence relations, namely the phlogiston–oxygen case (Section 2). Then I will show that my thesis is a logical consequence of the assumption that the following five rather natural requirements are satisfied (Sections 3–4):

(Requirement 0:) The vocabulary of the *compared* theories can be divided into a non-theoretical or empirical vocabulary, which is *shared* by the theories, and a theoretical vocabulary that is *specific* for each theory. The non-theoretical vocabulary contains either perceptual concepts or concepts that depend on unproblematic pre-theories (e.g., theories of measurement) which are shared by the compared theories. For a large part of the modern history of science this requirement seems to be met—even if one assumes that what counts as non-theoretical is relative to given background beliefs or measurement technologies (cf. Laudan and Leplin [1991], p. 451), as long as these background parameters are shared by the compared theories. Logically speaking, the satisfaction of requirement (0) is a presupposition for the possibility of comparing the success of competing theories, and hence a presupposition of the entire problem.

While requirement (0) is common in the debate, the next four requirements (1–4) arise from specific insights into the logical situation, which underlies my correspondence theorem. Here I give only a brief exposition of them; they are motivated and precisely defined in Sections 3–4:

(Requirement 1:) The past theory T must have been capable of producing *novel* predictions, which at the time of the theory construction were neither known, nor expected by means of empirical induction alone. This requirement has also been defended by Worrall ([1989], p. 113), Psillos ([1999], p. 106) and Ladyman and Ross ([2007], Chapter 2.1.3). Specific to my account is my logical definition of *strong* (i.e. potentially novel) empirical–predictive success. The requirement of novel predictions alone is not sufficient to refute Laudan's pessimistic meta-induction, because several of his historical examples meet this condition.

(Requirement 2:) The strong success of the predecessor theory T must be *yielded* by one (or several) *theoretical* terms or expressions of T, which can be empirically *indicated* or *measured* in certain circumstances by way of *several* bilateral reduction sentences.

(Requirement 3:) The entailment of T's strong empirical success by the successor theory T* must *depend* on the theoretical part of T*; and finally,

(Requirement 4:) the compared theories T and T* are 'causally normal'.

In Section 4, I will prove that if requirements 0–4 are satisfied, a correspondence relation between that expression φ of T which *yielded* T's strong success, and a certain expression τ^* of T* follows from the union of T* with T_{restr}, where T_{restr} is a T*-consistent content-part of T, which entails T's strong success. The correspondence relation has the form of an *equivalence* relation restricted to the domain A of T's strong success. This φ–τ^* correspondence entails the possibility of a so-called φ–τ^* *reference shift*, by which I mean the assignment of τ^*'s intended interpretation $I(\tau^*)$ to φ such that this shifted interpretation makes true the content-part T_{restr} of T, provided that T* is true. I express this situation by saying that T_{restr} is *indirectly true* (i.e., true *under* the shifted interpretation), and that φ *refers indirectly* to $I(\tau^*)$ (i.e., a φ–τ^*-correspondence holds, which entails the possibility of a φ–τ^* reference shift). The content-part T_{restr} of T, which is preserved by the φ–τ^* correspondence is called φ's *outer structure* because it covers φ's (causal) relations to the observable phenomena (and maybe to other T-expressions), while φ's so-called *inner structure* may be completely lost in φ's shifted interpretation within T*.

In Section 5, I illustrate how my theorem applies to further historical cases, and in Section 6, I explain the philosophical substance of my theorem in the style of 'objections and replies'. In the final section, Section 7, I point out the epistemological advantages of my account in regard to the prospects of justifying scientific realism.

2 The Case of the Phlogiston Theory

The phlogiston theory of combustion was developed by J. J. Becher and his student G. E. Stahl in the late seventeenth and early eighteenth century. According to this theory, every material that is capable of being burned or calcinated contains *phlogiston*—a substance different from ordinary matter, which was thought to be the bearer of combustibility. When combustion or calcination takes place, the burned or calcinated substance delivers its phlogiston, usually in the form of a hot flame or an evaporating inflammable gas, and a dephlogisticated substance-specific residual remains. In the 1780s, Lavoisier introduced his alternative oxygen theory according to which combustion and calcination consists in the oxidation of the substance being burned or calcinated, that is, in

the formation of a chemical bond of its molecules with oxygen. The assumption of the existence of a special bearer of combustibility became superfluous in Lavoisier's theory. In modern chemistry, Lavoisier's theory is accepted in a *generalized* form and nobody believes in the existence of phlogiston any more.

Oxygen theory describes the calcination of a metal as a process of *oxidation*, i.e., a process in which oxygen is *added* to the metal:[2]

$$\text{Metal} + \text{oxygen} \rightarrow \text{calx} (= \text{metal–oxide}).$$

In terms of the phlogiston theory, this reaction is described as *de-phlogistication*, i.e., a process in which the metal delivers its phlogiston:

$$\text{Metal} \rightarrow \text{calx} + \text{phlogiston} \uparrow.$$

Metals, coal and oils were assumed to be rich in phlogiston. The extraction of metals from calxes (mineral ores) through heating in charcoal was described by phlogiston theory as *phlogistication*, i.e., the inversion of the process of dephlogistication, in which the charcoal gives phlogiston to the calx to form a metal:

$$\text{Calx} + \text{coal} \rightarrow \text{metal} + \text{ash} (= \text{dephlogisticated coal}) + \text{byproduct: phlogisticated air} \uparrow.$$

Oxidation theory describes this process as *reduction*, i.e., the inversion of the process of oxidation, in which the calx delivers its oxygen to the coal:

$$\text{Metal–oxide} + \text{carbonium (coal)} \rightarrow \text{metal} + \text{carbon dioxide} \uparrow + \text{byproduct: ash}.$$

'Proper' end products and 'by-products' exchanged their role. Phlogiston theory identified the evaporating fume (carbon dioxide) with phlogisticated air (not all of the coal's phlogiston combines with the calx) and the ash as dephlogisticated coal; for oxygen theory, carbon dioxide is a proper end product, and the ash is a residual of incompletely oxidized coal.

A further domain explained by the phlogiston theory was salt-formation through dissolution of metals in acids. In terms of modern chemistry, the qualitative reaction is the following (every acid has the chemical composition 'hydrogenium–X', where the hydrogenium atoms give off their electrons to their negatively charged partner X):

$$\text{Metal} + \text{acid} (= \text{hydrogenium–X}) \rightarrow \text{metal–X} (= \text{salt}) + \text{hydrogenium} \uparrow.$$

Since strong heating of the salt yielded the calx (oxide) of the metal, phlogiston theorists described this process as a dephlogistication of the metal, in which

[2] *Chemical notation:* the substances mentioned on the left of the arrow are input substances and those mentioned on the right of the arrow are output substances of the chemical reaction. '↑' at a substance means that the substance is an evaporating gas.

calx and acid combine into the salt (McCann [1978], p. 32). Cavendish believed that the evaporating 'inflammable air' (hydrogenium) was pure phlogiston. So, phlogiston theory modelled salt-formation as follows:

$$\text{Metal} + \text{acid} \rightarrow \text{calx–acid}\,(= \text{salt}) + \text{phlogiston (inflammable gas)} \uparrow.$$

Phlogiston theory entailed a variety of potentially novel predictions. For example, phlogiston theory predicted that the process of saltification should be possible for *every* kind of metal, provided that heat and acid are sufficiently strong—even for metals that have *never* been put into acid before, which were chemically so far unexplored, or for rare metals with atypical phenomenological properties. A particularly striking example of a novel prediction made by phlogiston theory has been described by Carrier ([2004]): after Cavendish had identified inflammable air with phlogiston, Priestley predicted in 1782 that it should be possible to invert the process of calcination by adding inflammable air to a metal calx. He heated several metal calxes in inflammable air and observed that the inflammable air was almost completely absorbed and that the calxes were slowly reconverted into the metals. Priestley had also recorded the emergence of water droplets in this reaction, but he assumed that the water was contained in the inflammable air from the beginning (Carrier [2004], p. 151). In modern terms, Priestley had performed the following reaction of *reduction*

$$\text{Metal–oxide} + \text{hydrogenium} \rightarrow \text{metal} + \text{water}$$

while in terms of the phlogiston theory he had performed the following reaction:

$$\text{Calx} + \text{phlogiston}\,(+ \text{water}) \rightarrow \text{metal}\,(+ \text{water}).$$

This was celebrated as a further success of the phlogiston theory.

There were several empirical reasons why Lavoisier later concluded that his oxygen theory was true and that the postulate of phlogiston was superfluous. For example, in some cases the process of combustion or calcination resulted in an increase of weight of the dephlogisticated substance, which was explained by different ad hoc assumptions (one of them was the attribution of 'negative weight' to phlogiston). Moreover, phlogiston theorists were unable to isolate phlogiston in a coherent way (Cavendish's identification of phlogiston with hydrogenium gas did not work in other domains). Nevertheless, if we restrict the phlogiston theory to a certain domain of application, such as the oxidation and saltification of metals and the retransformation of metal calxes into pure metals, then the phlogiston theory was strongly empirically successful with respect to these domains. Although Lavoisier's oxygen theory surpassed the success of the phlogiston theory, it also faced severe difficulties: for example, Lavoisier assumed that the saltification of metals through acid is always the effect of the oxygen contained in the acid. But oxygen is not contained in all

acids, e.g., it is not contained in hydrochloric acid. These difficulties had been overcome only by the *generalized* oxidation and reduction theory of modern chemistry.

Let us now turn to our thesis of Section 1: if this thesis were true, then we could conclude from the strong empirical success of phlogiston theory that there is a certain correspondence between some of its central theoretical concepts and modern chemistry, so that from the modern viewpoint there is at least something that phlogiston theory has got right. For Carrier ([2004], pp. 154f), what phlogiston theory has got right is the classification of dephlogistication (oxidation) and phlogistication (reduction) as inverse chemical reactions. This is certainly true, but it seems to be too weak, because the inversions of certain chemical processes had already been recognized before, and *independently* of the theoretical part of phlogiston theory. What we are after is something in the *theoretical superstructure* of the phlogiston theory that corresponds to something in modern chemistry. What could this be?

There is nothing that directly corresponds to 'phlogiston' from the viewpoint of modern chemistry. But this is no wonder, because as we have already explained, phlogiston theory did not provide a general criterion of how phlogiston can be empirically identified. So the theoretical term 'phlogiston' was empirically underdetermined in phlogiston theory. The theoretical expressions of phlogiston theory, which did all the empirically relevant work and were *not* empirically underdetermined, were the expressions of *phlogistication = assimilation of phlogiston*, and of *dephlogistication = release of phlogiston*. For these two expressions there indeed exists a correspondence with modern chemistry, which goes much further than the identification of phlogistication with oxidation in Lavoisier's sense. To explain this correspondence we need a bit more of modern chemistry.[3]

Every substance consists of molecules, and molecules consist of atomic elements bound together by chemical bonds. The *electropositivity* of an element measures its tendency to contribute electrons to its neighbouring atoms in electrically polarized or ionic bonds. Conversely, the *electronegativity* measures the tendency of an element to attract electrons from the neighbouring atom in polarized or ionic bonds. Metals and hydrogenium are electropositive; carbonium is in the middle of the spectrum, and nonmetals such as oxygen are electronegative, with the extremes being the halogens. Oxidation of an elementary substance X (a metal, coal, etc.) in the *generalized sense* consists in the formation of a polarized or ionic bond of X with a electronegative substance Y, in which the atomic elements of X are electropositive and thus donate electrons to the electronegative neighbour Y in the bond. Every process of combustion, calcination or saltification consists of such an oxidation process. The inversion

[3] For the following cf. (Oxtoby et al. [1999], Chapters 3, 6.3).

of the process of oxidation is called the process of reduction: here the polarized or ionic bond between an electropositive X-ion and its electronegative neighbour is broken, X regains its missing electrons and reappears in its pure elementary form. Therefore we have the following correspondence relations between phlogiston theory and modern chemistry:

Correspondence relations between phlogiston theory and modern chemistry:

> Dephlogistication of X corresponds to (and hence indirectly refers to) the donation of electrons of X-atoms to their bonding partner in the formation of a polarized or ionic chemical bond.
>
> Phlogistication of X corresponds to (and hence indirectly refers to) the acceptance of electrons by positively charged X-ions from their bonding partner in the breaking of a polarized or ionic chemical bond.[4]

What was wrong in phlogiston theory was that phlogiston was thought of as a special substance that is *emitted* during a dephlogistication process. The electrons do not *leave* the chemical substance but just move a little bit to the electronegative neighbours in the molecule. What really is emitted as the end product of an oxidation process (besides the oxidized material) depends on the oxidans, that is, the input substance, which causes the oxidation and which provides the electronegative partner. If the oxidans is an acid, then what is emitted is hydrogenium, whence in these cases phlogiston could be identified with hydrogenium. If the oxidans is pure oxygen and the oxidized material is coal, then carbon dioxide is emitted. In the combustion of phosphorus and sulphur nothing is emitted, and therefore the weight increases after dephlogistication—this was the most problematic case for the phlogiston theory. But apart from such cases, phlogiston theory was strongly successful, and this strong empirical success is *explained* by the above correspondence relation.

The proposed correspondence relation surely does *not* preserve *all* of the meaning of 'phlogistication'. In particular, the assumption that phlogiston is a special substance contained in other substances and leaving them during the process of combustion is not preserved under this correspondence. Therefore, the correspondence relation cannot be regarded as an *analytic* truth, as a semantic *translation*. There is nothing in modern electron theory that would *analytically* entail a connection with phlogiston theory; nor the other way round. Rather, the above correspondence principle has to be regarded as a *synthetic* statement that is true only in the domain of application in which phlogiston theory was empirically successful.

[4] As was pointed out to me by James Ladyman, a similar correspondence holds between the 'phlogiston-richness' and the 'electropositivity' (or 'phlogiston-poorness' and 'electronegativity') of an element.

3 Steps Towards a Systematic Correspondence Theorem

Now I start my attempt to establish the existence of correspondence relations of the above sort on systematic grounds. Where should these correspondences come from? I will try to show that, under natural conditions, correspondence relations can be obtained as consequences of the *union* of (parts of) the two theories in their *joined language*. Kuhnian philosophers of science will probably object: how can it be possible to unify the conceptual frameworks of two theories whose theoretical superstructures are *incommensurable*?

Assuming that theories are represented by sets of axioms and that requirement (0) of Section 1 is satisfied, two cases of incommensurability are to be considered: on the one hand, the two theories may be *theoretically incomparable* in the sense of a nonoverlap of theoretical terms, and on the other hand, they may be (non-theoretically or theoretically) *incompatible*. The *first case* is logically harmless. As Hintikka ([1988], p. 27) has emphasized, a situation of theoretical (or even total) incomparability does not prevent us from joining together the conceptual frameworks of two theories as follows: If L_1 is the language of T_1, and L_2 the language of the T_2, where L_1 and L_2 have a common logic and logical vocabulary, then the united language of both theories, L_{12}, is simply obtained as the set of all formulas expressible in the united vocabulary. In this united language we formulate the united theory $T_1 \cup T_2$ and see what follows from it.

The *second case* is (logical) *incompatibility* between T_1 and T_2. According to Carrier ([2001], pp. 67–9) and Hoyningen-Huene ([1993], Chapter 6.3(b)), this is a crucial aspect of incommensurability. We can distinguish two kinds of incompatibility. Non-theoretical (empirical) incompatibility arises when the two theories entail contradictory consequences in their joint non-theoretical vocabulary. More interesting are cases of *theoretical* incompatibility, because they are historically more stable. Theoretical incompatibility presupposes that the two theories have at least some theoretical concepts in common. Phlogiston and oxygen theories were theoretically incompatible, because they shared certain theoretical concepts, such as 'substance', 'mass', etc. To repeat their basic theoretical incompatibility: for phlogiston theory, combustion is the effect of a special substance (phlogiston), but there is no such substance for oxygen theory.

Also (non-theoretical or theoretical) incompatibility is *not* an obstacle for establishing correspondence relations, provided we *restrict* the outdated theory T. We cannot simply form a union of two incompatible theories $T \cup T^*$. Doing this would *trivialize* our claim that this union entails correspondence principles: since this union is contradictory, it implies everything whatsoever. Therefore we take pains to consider only a certain *part* of the content of the outdated theory T, which yielded T's strong empirical success but is nevertheless consistent with the contemporary theory T^*.

Our next task is to formulate natural conditions under which a correspondence between theoretical expressions of the two theories T and T^* can be established. I propose the following condition on the predecessor theory T, which comprises requirements 1 and 2 of Section 1: *T must contain one (or several) theoretical expressions φ, which yielded T's strong empirical success by way of empirical indication or measurement laws* connecting φ with several phenomena in several circumstances. The general format of an empirical indication or measurement law for a theoretical term φ is a (Carnapian) *bilateral reduction sentence* of the following form:

$$(BR_i)\ (\forall)\ A_i \rightarrow (\varphi(x) \leftrightarrow R_i), \quad \text{where } V_f(x) \subseteq V_f(R_i),\ V_f(x) \subseteq V_f(A_i).$$

In words: Under empirical circumstances A_i, the presence of φ is indicated or measured by an empirical phenomenon or process R_i.

Explanation of the notation: $\varphi(x)$ is a T-theoretical term or expression in the free variable(s) x, which describes the state of the individual system x in the T-theoretical language (x may also be a *vector* of variables describing a composed system consisting of several parts). The BRs for φ are *indexed* (BR_i), because it is a crucial condition that φ is characterized by *many* such BRs. The formulas A_i and R_i are possibly complex non-T-theoretical formulas, whose free variables obey the condition at the right side (where $V_f(\alpha)$ = the set of variables occurring free in expression α). Hence, the formulas A_i and R_i contain all variables (x) that refer to the individual system under consideration, but may in addition contain further variables—especially the *time* variable t, or variables referring to individuals in the external circumstances.[5] The bilateral reduction sentence is understood as *universally quantified*, indicated by the '(∀)' in front of it. I will frequently omit the variables and mention them only when necessary.

I understand bilateral reduction sentences, differently from Carnap, in a 'modernized' and nonreductionist sense. They are not analytically true (as assumed by the early Carnap), because as we shall see soon, several BR_is joined together have synthetic consequences (which was recognized by Carnap [1936], p. 451). Rather, they are synthetic statements, which express conditions for the empirical indication or measurement of the theoretical entity φ under specific conditions A_i. They are usually *not* part of T's axiomatization (so-called

[5] Certain subtleties are involved concerning the time variable t, which I mention only in the footnote. If φ is an intrinsic property of x that does not depend on time, then the bilateral reduction sentence has the explicit form (i) $\forall x,t: A_ixt \rightarrow (\varphi(x) \leftrightarrow R_ixt)$. This sentence is empirically creative insofar it logically implies (ii) $\forall x: \exists t(A_ixt \wedge R_ixt) \rightarrow \forall t(A_ixt \rightarrow R_ixt)$. This is *okay*: if a certain behaviour R_ixt under circumstances A_ixt (such as a certain chemical reaction) empirically indicates an intrinsic time-independent property of x, then it must be expected that x will exhibit this behaviour every time when it is put into circumstances A_i. More generally, if 't' stands for the vector of variables that are free in (A_i,R_i), but not in $\varphi(x)$, then BR_i has the explicit form (i) and logically implies (ii). This subtlety causes further subtleties discussed in footnotes 6 and 8. Note that our notation also admits φ being a temporal property—in this case, x contains the variable t; or more generally, $V_f(A_i,R_i) = V_f(x)$, and the explained subtlety does not arise.

'rules of correspondence', as assumed by the late Carnap), but are obtained as logical consequences of a suitably rich version of the theory. For example, if $\varphi(x)$ stands for 'x delivers phlogiston', and A_i stands for 'x is a metal that is put into hydrochloric acid (at a certain time t)', then R_i stands for 'x dissolves in the acid and inflammable air evaporates'. Bilateral reduction sentences subsume all important kinds of statements expressing empirical indication or measurement methods. In particular, this format subsumes *quantitative measurement laws* for theoretical terms, in which instead of the equivalence we have a numerical identity, because the two statements (1) and (2) are logically equivalent (where r is a variable ranging over real numbers, and f_i is a function term):

$$(\forall r :)\ A_i(x, u) \rightarrow (\varphi(x) = r \leftrightarrow f_i(x, u) = r). \quad (1)$$

$$A_i(x, u) \rightarrow (\varphi(x) = f_i(x, u)). \quad (2)$$

For example, if '$\varphi(x) = r$' stands for the expression 'mass-of-x = r grams' and A_i for the circumstance 'x is put on a balanced beam scales', then R_i stands for the condition 'the number of one gram units on the other side of the balanced beam is r', formalized as '$f_i(x,u) = r$'. The BR-sentence (1) is logically equivalent with (2), which is the standard format of a quantitative measurement law.

To avoid misunderstandings, that a theoretical term φ of T can be empirically indicated or measured does, of course, *not* mean that it becomes a non-theoretical term of T. The difference is, rather, that the measurement of φ presupposes the theory T, in other words, T is needed to derive φ's measurement conditions, while T's non-theoretical terms can be empirically measured independently of T, based on pure perception or non-theoretical knowledge (this insight goes back to Sneed [1971]). In the model-theoretic reading, a bilateral reduction statement for φ expresses a *class of measurement models* for φ in the sense of Balzer et al. ([1987], p. 64).

It is easy to see how the theoretical expression φ of T *yields* strong (i.e., potentially novel) empirical success: because the class of bilateral reduction sentences, which characterize φ, $BR(T) := \{BR_1(T), \ldots, BR_n(T)\}$, *entail* such a potentially novel success as follows:

$A_1 \rightarrow (\varphi(x) \leftrightarrow R_1)$	Potential novel predictions entailed by BR(T):
$A_2 \rightarrow (\varphi(x) \rightarrow R_2)$	(i) $(A_1 \wedge R_1) \rightarrow (A_2 \rightarrow R_2)$,
etc.	(ii) $(A_1 \wedge \neg R_1) \rightarrow (A_2 \rightarrow \neg R_2)$ etc.

With '$(A_1 \wedge \pm R_1) \rightarrow (A_2 \rightarrow \pm R_2)$' I mean a conditional of one of the two forms (i) or (ii) above. The class of conditionals $\{(A_i \wedge \pm R_i) \rightarrow (A_j \rightarrow \pm R_j)$: $1 \leq i \neq j \leq n\}$ is what I call the *strong (potential) empirical success* of T. These conditionals figure as potentially novel predictions because with their help one can infer something from what has happened in one domain of application about what will happen in another domain of application, without the other

domain of application having already been investigated.[6] For example, when φ represents a chemical model of oxidation, A_1 may describe the exposure of a metal to air and water, and R_1 the end products of the reaction of oxidation, A_2 the exposure of a metal to hydrochloric acid, and R_2 the end products of the reaction of salt-formation, etc. Even if the predictions asserted by scientific theories do not always have this explicit form, they can very often be reconstructed in this form.[7] The *causal* interpretation of the role of φ as described by BR(T) is the following: φ figures as a *common cause* of the observable behaviour R_i in circumstances A_i. The philosophical importance of this *common cause* condition is explained in Section 6.

Now let us see how for the theoretical term φ of T (satisfying requirements 1 and 2) a correspondence can be established to a theoretical expression of another theory T^*. This is possible whenever T^* contains a (possibly complex) expression τ^* for which T^* entails the *same* measurement conditions as T entails for φ. Then we have the following situation ('$\Vdash$' for 'logical consequence'):

$$\mathrm{BR(T)}: \ A_i \rightarrow (\varphi(x) \leftrightarrow R_i) \quad (\text{where } T \Vdash \mathrm{BR(T)}) \tag{3}$$

$$\mathrm{BR}(T^*): \ A_i \rightarrow (\tau^*(x) \leftrightarrow R_i) \quad (\text{where } T^* \Vdash \mathrm{BR}(T^*)). \tag{4}$$

The union of (3) and (4) entails that φ and τ^* are equivalent in the circumstances of the measurement:

$$T \cup T^* \Vdash \mathrm{BR(T)} \cup \mathrm{BR}(T^*) \Vdash A_i \rightarrow (\varphi(x) \leftrightarrow \tau^*(x)), \tag{5}$$

which is exactly a domain-restricted correspondence principle of the sort we are after. Since T and T^* have radically different theoretical superstructures, the existence of such a T^*-expression τ^* is neither obvious nor can it be expected to be directly evident from T^*'s axioms. What our theorem will show is that the existence of such a T^*-expression can indeed be *derived* from the assumption that our requirements (3) and (4) are satisfied.

[6] If $\varphi(x)$ is a time-independent property, then these potentially novel predictions have the exact form: $\forall x: \exists t(A_i xt \wedge R_i xt) \rightarrow \forall t(A_j xt \rightarrow R_j xt)$, saying that whenever x has behaved at *some* time in subdomain A_i in the way R_i, then every time x is put into subdomain A_j it will behave in the way R_j.

[7] For example, the novel prediction may have the form 'an x of kind α will exhibit reactions R_j under conditions A_j', where 'α' is a theoretical statement such as 'being rich of phlogiston'. But the *empirical indicators* for 'x is of kind α' will (after some transformations) again have the form 'under conditions $A_i x$, $R_i x$ has happened'. Even Cavendish's novel prediction following from the *inversion principle for chemical reactions* may be reconstructed in this form. The inversion principle asserts that for each BR_i describing a dephlogistication (oxidation) process, an inverted BR holds, abbreviated as $InvBR_i$, which describes a phlogistication (reduction) process ($\forall i \in \{1, \ldots, n\}$: $BR_i \leftrightarrow InvBR_i$). Let In(x) and Out(y) denote that x is an input and y an output substance of a chemical reaction (where y = fx), and '$\varphi(x,y)$' stand for 'phlogiston moves from x to y'. Then a BR for dephlogistication of x has the form $In(x,t) \wedge Out(fx,t) \wedge A(x) \rightarrow (\varphi(x,fx) \leftrightarrow B(fx))$, where A and B denote kinds of substances under certain conditions, and the inverted InvBR for phlogistication of x has the form $In(fx,t) \wedge Out(x,t) \wedge B(fx) \rightarrow (\varphi(fx,x) \leftrightarrow A(x))$. Taken together, BR and InvBR entail the novel prediction $\exists t(In(x,t) \wedge Out(fx,t) \wedge A(x) \wedge B(fx)) \rightarrow \forall t(In(fx,t) \wedge Out(x,t) \wedge B(fx) \rightarrow A(x))$.

Requirement (3) demands that the successor theory T^* entails T's strong empirical success in a way that *depends* on the theoretical part of T^*. The dependence on T^* is very natural, because from empirical descriptions of what goes on in a system x in domain A_i, nothing can be concluded by means of *empirical induction alone* about what goes on in system x in a qualitatively different domain A_j. For example, from observing reactions of metals in hydrochloric acid, nothing can inductively be concluded about the behaviour of metals in oxygen or water, and from observing projectiles on the earth, nothing can inductively be concluded about the planets moving around the sun. Connections of this sort can only be provided by a theory, and they are possible because in the theory one can infer from $A_ix \wedge R_ix$ a certain theoretical description $\tau(x)$ of the intrinsic properties of x, which in turn entails in the theory that $(A_jx \rightarrow R_jx)$ will hold. Therefore I assume that for every conditional of the form $(A_i \wedge \pm R_i) \rightarrow (A_j \rightarrow \pm R_j)$, which the new theory T^* entails, there exists a mediating theoretical description or *theoretical mediator* $\tau^*(x)$ such that T^* entails $(A_ix \wedge \pm R_ix \rightarrow \tau^*(x))$ as well as $(\tau^*(x) \rightarrow (A_jx \rightarrow \pm R_jx))$.[8] If a theory T has this property w.r.t. a partition $\{A_i{:}1 \leq i \leq n\}$ of a domain A into subdomains, then I say that the strong (potential) empirical success entailed by T is *T-dependent*.

The theoretical mediator $\tau^*(x)$ may depend both on i and j,

$$(A_i \wedge R_i) \rightarrow \tau^*_{ij}(x),\ \ \tau^*_{ij} \rightarrow (A_j \rightarrow R_j) \text{ for all } j \neq i,\ \ j \in \{1, \ldots, n\},$$

in which case the theory T^* utilizes for the prediction of the system's behaviour in each new domain of application A_j $(j \neq i)$ a different theoretical description. However, we can join these descriptions into a big conjunction

$$\tau^*_i(x) :\leftrightarrow \wedge\{\tau^*_{ij}(x) :\ \ 1 \leq j \leq n,\ \ j \neq i\},$$

which is T^*'s maximal theoretical description, which can be inferred from $(A_i \wedge R_i)$ and contains all theoretical details necessary to infer what goes on in all other domains. Thus we can assume that the theoretical mediator $\tau_i{}^*$ depends only on i. In Section 4, we shall see that we can even get rid of this dependence on index i.

With our requirement (3), we are *almost* where we want to be. As the proof of our theorem will show, the satisfaction of this requirement implies that T^* entails unilateral reduction sentences for the mediating T^*-theoretical expression τ^* in terms of the circumstances A_i and reactions R_i, which imply implications but not equivalences between φ of T and τ^* of T^*. To obtain equivalences and hence the desired correspondence principles, we will need one further natural assumption (requirement 4 of Section 1), namely that the considered theories

[8] If τ^* is time-independent (recall footnote 6), the T^*-entailed mediator statements have the form $\forall x{:}\ (\exists t(A_ixt \wedge R_ixt) \rightarrow \tau^*(x))$ and $\forall x{:}\ (\tau^*(x) \rightarrow \forall t(A_jxt \rightarrow R_jxt))$. These two statements are logically equivalent with the *universally* quantified statements $\forall x,t{:}\ (A_ixt \wedge R_ixt \rightarrow \tau^*(x))$ and $\forall x,t{:}\ (\tau^*(x) \rightarrow \forall t(A_jxt \rightarrow R_jxt))$, respectively. So it is possible to represent these statements without quantifiers.

are *causally normal* in the following sense. The non-theoretical predicates or parameters of the theory divide into a set of *causally independent* parameters, which describe the circumstances A_i of the special subdomains, and a set of *causally dependent* parameters. The behaviour of the system x w.r.t. its dependent parameters under given values of the independent parameters can be deduced from the theory. But it is impossible to derive from a purely theoretical description $\tau(x)$ of a system x any empirical assertion about the status of the independent parameters of x. This is again a very natural condition. For example, nothing can be concluded from the theoretical nature of a certain substance about what humans do with it, about whether they expose it to hydrochloric acid or to heat or whatever. Or, nothing can be concluded from a purely mechanical description of a physical body about its initial conditions, about whether the body was thrown into the air, or split into pieces, etc.

4 The Correspondence Theorem and Its Ontological Interpretation

The following definition summarizes the central notions that have been explained in the previous section (*note:* by definition, T entails a set of sentences Σ iff T entails all sentences in Σ).

Definition:

(1) A is a *partitioned domain* iff $A = A_1 \cup \ldots \cup A_n$, $n \geq 2$, where the A_i are mutually exclusive and qualitatively different subdomains described by non-theoretical means.

(2.1) A *strong potential empirical success* of a theory T w.r.t. partitioned domain $A = A_1 \cup \ldots \cup A_n$ is a set of conditionals of the form $(A_i \wedge \pm R_i) \rightarrow (A_j \rightarrow \pm R_j)$ for $i \neq j \in \{1, \ldots, n\}$, which are entailed by T, where the R_i describe the empirical (non-theoretical) behaviour of the system x under consideration in circumstances A_i.

(2.2) Such a strong potential empirical success is *yielded* by a theoretical expression φ of T iff T entails the bilateral reduction statements $BR(T) = \{A_i \rightarrow (\varphi(x) \leftrightarrow R_i) : 1 \leq i \leq n\}$.

(2.3) Such a strong potential empirical success of T is *T-dependent* iff for every conditional of the form $(A_i \wedge \pm R_i) \rightarrow (A_j \rightarrow \pm R_j)$ following from T, there exists a theoretical description $\tau_i(x)$ of the underlying system x such that $(A_i \wedge \pm R_i \rightarrow \tau_i(x))$ and $\tau_i(x) \rightarrow (A_j \rightarrow \pm R_j)$ follow from T (cf. footnote 8).

(3) A theory T is *causally normal* w.r.t. a partitioned domain $A = A_1 \cup \ldots \cup A_n$ iff (i) the non-theoretical vocabulary of T divides into a set of independent and a set of dependent parameters (predicates or function terms), (ii) the descriptions 'A_i' of the subdomains A_i are formulated solely by means of the independent parameters (plus logico-mathematical symbols) and (iii) no non-trivial claim about the state of the independent parameters of a system x can be derived in T from a purely T-theoretical and T-consistent description of x.

Correspondence theorem: Let T be a consistent theory that is causally normal w.r.t. a partitioned domain $A = A_1 \cup \ldots \cup A_n$ and contains a T-theoretical expression $\varphi(x)$, which *yields* a strong potential empirical success of T w.r.t. partitioned domain A.

Let T^* be a consistent successor theory of T (with an arbitrarily different theoretical superstructure), which is likewise causally normal w.r.t. partitioned domain A and which entails T's strong potential empirical success w.r.t. A in a T^*-dependent way.

Then T^* contains a theoretical expression $\tau^*(x)$ such that T and T^* together imply a *correspondence relation* of the form

(C): $A \rightarrow (\varphi(x) \leftrightarrow \tau^*(x))$

(*in words:* whenever a system x is exposed to the circumstances in one of the subdomains of A, then x satisfies the T-theoretical description φ iff x satisfies the T^*-theoretical description τ^*),

which implies that $\varphi(x)$ indirectly refers to the theoretical state of affairs described by $\tau^*(x)$, provided T^* is true.

Corollary 1. $BR(T) \cup T^*$ is consistent, and (C) follows already from $BR(T) \cup T^*$ (even from $BR(T) \cup BR(T^*)$; see proof step (14)).

Corollary 2. τ^* is unique in domain A modulo T^*-equivalence.

Remark. The theorem applies to all theoretical expressions φ of T which yield strong potential empirical success by way of bilateral reduction statements. We speak of *potential* success because the logical part of the theorem is independent of the factual success of the considered theories. If the potential success of T is preserved by T^* in a T^*-dependent way, and both theories are causally normal, the correspondence relation (C) follows. Corollary 1 tells us that this correspondence principle follows in a *non-trivial* way, from a certain part of T that is consistent with T^*. Corollary 2 informs us that if T^* contains several different theoretical descriptions $\tau_i{}^*$ which satisfy the correspondence theorem, then T^* entails that they are equivalent in domain A.

Proof of the correspondence theorem:

We assume the following bilateral reduction sentences follow from T:

$$T \Vdash BR(T), \text{ where } BR(T) := \{A_i \rightarrow (\varphi(x) \leftrightarrow R_i) : 1 \leq i \leq n\}, \qquad (6)$$

and because T is causally normal w.r.t. $A_1 \cup \ldots \cup A_n$, the A_is are expressed in terms of independent parameters and the R_is in terms of dependent parameters of T.

We assume that the strong potential empirical success of T, which follows from the bilateral reduction sentences in (6), is also entailed by T^*, and so the following must hold (recall footnote 6):

$$\forall i \neq j \in \{1, \ldots, n\} : T^* \Vdash (A_i \wedge R_i) \rightarrow (A_j \rightarrow R_j), \text{ and}$$
$$T^* \Vdash (A_i \wedge \neg R_i) \rightarrow (A_j \rightarrow \neg R_j). \qquad (7)$$

Because this strong potential empirical success entailed by T^* is T^*-dependent, (7) implies the following:

For every i there must exist T^*–theoretical descriptions $\tau_i^*(x)$ and $\mu_i^*(x)$ such that (8)

$$T^* \Vdash (A_i \wedge R_i) \rightarrow \tau_i^*(x), \text{ and}$$
$$\forall j \neq i,\ j \in \{1, \ldots, n\} : T^* \Vdash \tau_i^*(x) \rightarrow (A_j \rightarrow R_j),$$

and

$$T^* \Vdash (A_i \wedge \neg R_i) \rightarrow \mu_i^*(x), \text{ and}$$
$$\forall j \neq i,\ j \in \{1, \ldots, n\} : T^* \Vdash \mu_i^*(x) \rightarrow (A_j \rightarrow \neg R_j).$$

It is sufficient for our purpose to *choose one fixed i*, say $i = k$. We abbreviate $\tau_k{}^* = \tau^*$ and $\mu_k{}^* = \mu^*$. Recall footnote 8, which implies that the following proof remains essentially propositional.

One can see why the condition of *T^*-dependent* strong potential success is crucial. Without this condition it would be impossible to say *how* the empirical consequences in (7) are obtained from T^*. With this condition we gain two T^*-theoretical descriptions of x: one mediates the empirical consequences involving positive R_is, and the other one mediates the empirical consequences involving negative R_is.

From (8) it follows by propositional logic that:

$$T^* \Vdash (A_k \rightarrow (R_k \rightarrow \tau^*(x)),\ \forall j \neq k,\ j \in \{1, \ldots, n\} : T^* \Vdash A_j \rightarrow (\tau^*(x) \rightarrow R_j) \tag{9}$$
$$T^* \Vdash (A_k \rightarrow (\neg R_k \rightarrow \mu^*(x)),\ \forall j \neq k,\ j \in \{1, \ldots, n\} : T^* \Vdash A_j \rightarrow (\mu^*(x) \rightarrow \neg R_j).$$

So far we have two unilateral reduction sentences for *two different* T^*-theoretical descriptions, τ^* and μ^*, in the measurement conditions of φ. We want to derive a bilateral reduction sentence for a T^*-theoretical description that has the same form as the bilateral reduction sentence for φ. Now the condition of causal normality comes into play.

It follows from (9) by propositional logic that

$$T^* \Vdash (\neg\tau^*(x) \wedge \neg\mu^*(x)) \rightarrow \neg A_k,$$
$$\forall j \neq k,\ j \in \{1, \ldots, n\} :\ T^* \Vdash (\tau^*(x) \wedge \mu^*(x)) \rightarrow \neg A_j. \tag{10}$$

But since T^* is causally normal w.r.t. the given partition of A, A_j is an assertion described in terms of independent parameters, and so it follows that $\tau^*(x) \wedge \mu^*(x)$ and $\neg\tau^*(x) \wedge \neg\mu^*(x)$ must be T^*-inconsistent, for otherwise T^* could not be causally normal. Therefore we have

$$T^* \Vdash (\tau^*(x) \leftrightarrow \neg\mu^*(x)). \tag{11}$$

Together with (11), (9) gives us the following:

$$T^* \Vdash (A_k \to (R_k \leftrightarrow \tau^*(x)), \text{ as well as} \tag{12}$$
$$\forall j \neq k,\ j \in \{1, \ldots, n\} : T^* \Vdash A_j \to (\tau^*(x) \leftrightarrow R_j).$$

By summarizing the two statements in (12), we get

$$T^* \Vdash BR(T^*), \text{ where } BR(T^*) := \{A_i \to (\tau^*(x) \leftrightarrow R_i) : 1 \leq i \leq n\}. \tag{13}$$

In other words, T^* entails for τ^* the same bilateral reduction sentences as T entails for φ.

From BR(T) in (6) and BR(T^*) in (13), we can derive by propositional logic the intended correspondence relation:

$$\forall i \in \{1, \ldots, n\} : BR(T) \cup BR(T^*) \Vdash A_i \to (\varphi(x) \leftrightarrow \tau^*(x)), \tag{14}$$

and since $A := A_1 \vee \ldots \vee A_n$, (14) gives us

$$(C) : T \cup T^* \Vdash BR(T) \cup BR(T^*) \Vdash A \to (\varphi(x) \leftrightarrow \tau^*(x)).$$

Concerning Corollary 1: (C) entails the second conjunct of this corollary. For its first conjunct, we must show that BR(T) is consistent with T^*. *Proof by reductio ad absurdum*. If $T^* \cup BR(T)$ were inconsistent, then $T^* \Vdash \neg\wedge BR(T)$ and hence (i) $T^* \Vdash \vee\{\neg(A_i \to (\varphi(x) \leftrightarrow R_i)): 1 \leq i \leq n\}$ would hold.[9] Since $\neg(A \to B) \Vdash A$, we would obtain from (i) and propositional logic (ii): $T^* \Vdash A_1 \vee \ldots \vee A_n$. (If the anonymous time variable t is present, as described in footnote 6, we obtain, strictly speaking, $T \Vdash \exists t(A_1(x,t) \vee \ldots \vee A_n(x,t))$, which says that it follows from T that the system under consideration has been at least once in one of the circumstances A_i.) But since T is causally normal, this is *impossible*.

Concerning Corollary 2: Recall that we have defined $\tau^* = \tau^*_k$ for some fixed k. Since we can carry out the proof for every such k, we can obtain in step (13):

$$\forall k \in \{1, \ldots, n\},\ \forall i \in \{1, \ldots, n\} : T^* \Vdash A_i \to (\tau^*_k(x) \leftrightarrow R_i). \tag{15}$$

Step (15) implies by propositional logic:

$$\forall i \in \{1, \ldots, n\},\ \forall j \neq k \in \{1, \ldots, n\} : T^* \Vdash A_i \to (\tau^*_j(x) \leftrightarrow \tau^*_k(x)). \tag{16}$$

For example, $A_1 \to (\tau^*_1 \leftrightarrow R_1)$ and $A_1 \to (\tau^*_2 \leftrightarrow R_1)$ imply $A_1 \to (\tau^*_1 \leftrightarrow \tau^*_2)$, etc. Since $A := A_1 \vee \ldots \vee A_n$, (16) gives us by propositional logic

$$\forall j \neq k \in \{1, \ldots, n\} : T^* \Vdash A \to (\tau^*_j(x) \leftrightarrow \tau^*_k(x)). \tag{17}$$

[9] For S a finite set of statements, $\wedge$S denotes the conjunction and $\vee$S the disjunction of the elements of S.

So T^* entails that all the theoretical descriptions $\tau_i^*(x)$ $(1 \leq i \leq n)$ are equivalent in the domain A. ■

Against our proof of Corollary 1, the consistency of BR(T) with T^*, one may object that it relies on the logical properties of material implication, i.e., $\neg\forall x(A \rightarrow B)$ implies $\exists x(A \wedge \neg B)$. For a formulation with the help of nomological or counterfactual implications, this is not valid. But there are also other ways to establish the consistency of BR(T) with T^*. For example, the consistency of BR(T) $\cup$ T^* can be proved from the following condition:[10]

$$\forall x(\varphi(x) \leftrightarrow \tau^*(x)) \text{ is } T^*\text{–consistent.} \tag{18}$$

In almost all situations I can think of, (18) will be satisfied, because in the context of this condition the expression φ is deprived of all content which is not already contained in BR(T). In particular, (18) must hold whenever the following two stronger conditions are satisfied: (18a) φ's nonlogical terms are not contained in T^*, and (18b) φ's possible extensions are logically unrestricted (as defined in footnote 11).[11]

The correspondence theorem presupposes that T's theoretical description φ can be empirically characterized by means of *bilateral* reduction sentences. If φ were only characterized by *unilateral* reduction sentences of the two forms

$$A_i \rightarrow (\varphi(x) \rightarrow R_i), \text{ and}$$

$$A_k \rightarrow (R_k \rightarrow \varphi(x)),$$

then the proof of this theorem would not work. It is an *open question* whether under this condition a weaker version of the correspondence theorem can be established. However, I think that theories that do not provide bilateral but merely unilateral empirical characterizations of their theoretical concepts occur only in early stages of science. In particular, as soon as reduction sentences are formulated in terms of *quantitative* concepts, the difference between unilateral and bilateral reduction sentences *vanishes*, because functions are *right-unique*. More precisely, the following two unilateral reduction sentences

$$A_i \rightarrow \forall r(\varphi(x) = r \rightarrow f(x) = r), \text{ and } A_i \rightarrow \forall r(f(x) = r \rightarrow \varphi(x) = r)$$

are mutually logically equivalent, and are both equivalent with

$$A_i \rightarrow (\varphi(x) = f(x)),$$

[10] *Proof:* By (18) there exists a model M for $T^* \cup \forall x(\varphi(x) \leftrightarrow \tau^*(x))$. T^* entails BR(T^*), whence $T^* \cup \forall x(\varphi(x) \leftrightarrow \tau^*(x))$ entails BR(T), since BR(T) = BR(T^*)[φ/τ^*]. Thus M verifies T^* $\cup$ BR(T), and so T^* $\cup$ BR(T) is consistent.

[11] That φ's possible extensions are logically unrestricted means that for every domain D and $X \subseteq D^k$ (where φ has k free variables) there exists an interpretation I such that $I(\varphi) = X$. Given a model M = (D,I) of T^*, then by condition (18a + b) we can expand M to a model M′ for φ, which assigns to φ the same extension as to τ^*. This entails condition (18).

which is in turn logically equivalent with the bilateral reduction sentence

$$A_i \rightarrow \forall r(\varphi(x) = r \leftrightarrow f(x) = r).$$

At this point let us reflect on the *ontological interpretation* of the correspondence relation (C). Of course, (C) is *not* meant to say that whenever T's intended model is realized (phlogiston leaves the substance), also T*'s intended model is realized (electrons move to the bonding partner). This would be a strange scenario of 'causal overdetermination': two distinct causal scenarios were simultaneously realized, each sufficient to explain the observable behaviour. What (C) expresses is the possibility of a *φ–τ^*-reference shift:* instead of the reference assigned to φ in T's intended model (phlogiston leaves the substance), we can assign to φ the reference of τ^* in T*'s intended model (electrons move to the bonding partner), under *preservation* of the strong empirical success of T.

The claim of the correspondence theorem concerning the indirect reference of φ means (by our definition in Section 1) that the φ–τ^*-reference shift makes BR(T) indirectly true, i.e., true under the shifted interpretation (we identify T's restriction T_{restr} of Section 1 with BR(T)). The *model-theoretic proof of φ's indirect reference and BR(T)'s indirect truth* is straightforward: assume T is true in the intended (but 'non-real') model (D,I), T* is true in the intended (and 'real') model (D*,I*) and $I(\pi_i) = I^*(\pi_i)$ for all (shared) non-theoretical terms π_i of T and T*. This implies that $D \cap D^* \neq \emptyset$, and π_i's extensions are taken from $D \cap D^*$. The *φ–τ^*-shifted* interpretation I′ of BR(T) differs from I in that $I'(\varphi) = I^*(\tau^*)$. By the correspondence theorem, $T^* \Vdash BR(T^*)$, and so (D*,I*) verifies BR(T*), where BR(T*) results from BR(T) by replacing φ by τ^* (recall step (13) of the proof). Hence (D*,I′) verifies BR(T).

It is important that the expression φ, which yielded T's strong success, need not be a primitive term but may be *composite*, which leaves room for either an ontological underdetermination or the nonreference of T's primitive terms. In the case of phlogiston, the complex terms 'dephlogistication' and 'phlogistication' can be related to corresponding electron donations or acceptances, but not so for the primitive term 'phlogiston'. More generally, whenever T's expression φ corresponds to τ^* of T*, but the ontology of the old theory T concerning the entities involved in φ is *incompatible* with the contemporary theory T*, then it will be the case that φ is not a primitive but a *complex* expression of T, and T will contain certain theoretical assumptions about φ's *inner structure* or *composition*, which, from the viewpoint of T*, are false—for example, that 'dephlogistication' is a process in which a special substance different from ordinary matter, called 'phlogiston', leaves the combusted substance. While T has a *right* model about φ's *outer structure*, i.e., the causal relations between the complex entity φ and the empirical phenomena, it has a *wrong* model about φ's *inner structure*. This situation is *typical* even for most advanced contemporary theories. For example, we are confident that protons exist because they

are measurable common causes of a huge variety of BRs. But concerning the hypothesis about the *inner composition* of protons consisting of *three quarks*, things are different: physicists cannot measure quarks in isolation and, hence, are much more uncertain about their reality.

My notions of the outer and inner structure of a complex expression or entity φ reflect Worrall's distinction between 'structure' and 'content' in an ontologically unproblematic way, which I think can be defended against the objections of Psillos and Papineau (recall footnote 1): the 'structure' that is preserved is φ's *outer* structure, while the 'content' that is not preserved is φ's *inner* structure. Often, the preserved outer structure of a T-expression φ_1 does not only contain φ_1's relations to observable phenomena, but covers also φ_1's relation to other T-theoretical terms φ_2 for which a T*-correspondence can *also* be established. In this sense, the relation between dephlogistication and phlogistication as inverse chemical reactions, as described in footnote 7, is preserved in modern chemistry. To elaborate the notions of 'inner' and 'outer structure' in a more general way would be an important task for future work.

5 Further Historical Applications

We now illustrate the correspondence theorem with some more historical examples. First, let us see how phlogiston theory fits into the logical corset of the correspondence theorem. The phlogiston theory implies bilateral reduction sentences of the following form, for $1 \leq i \leq n$, $n \geq 2$:

> If an input substance x of kind X_i (e.g., a metal) is exposed to the influence of an input substance y of type Y_i (e.g. hydrochloric acid), then x gets dephlogisticated (or phlogisticated, resp.) iff the chemical reaction produces output substances $z = f(x,y)$ of type $Z_i = f(X_i, Y_i)$ (e.g., the metal dissolves and inflammable air evaporates).

Given that modern chemical oxidation and reduction theory entails the empirical consequences of these bilateral reduction sentences that make up the strong success of phlogiston theory, the correspondence theorem entails that the two theories together must imply a correspondence relation, saying:

> If x is an input substance of one of the kinds X_i and is exposed to the influence of chemical input substances y of type Y_i, then x gets dephlogisticated (or phlogisticated, resp.) during the reaction iff this reaction satisfies a certain theoretical description in terms of modern chemistry

—and we have found such a description, namely the following:

> ... iff during this reaction the atoms/molecules of x donate electrons to (or accepts electrons from, resp.) the atoms/molecules of y during the formation (or breaking, resp.) of a chemical bond.

Another example from Laudan's list is the *caloric theory of heat* (cf. Carrier [2004], pp. 151ff). According to this theory (which was, for example, believed by Lavoisier), every material substance contains some amount of *caloric*, and this amount is responsible for the heat or temperature of the substance—the more caloric it contains, the hotter it will be. Thereby caloric was assumed to be a substance consisting of weightless particles. While the particles of material substances *attract* each other in a substance-specific way, which is demonstrated by the forces of cohesion, the caloric particles *repel* each other, which is confirmed by thermal expansion, i.e., by the fact that (almost) all substances expand in their volume when their temperature and hence their amount of caloric increases. In the solid state of substances, the attractive forces among the material particles dominate the repulsive forces among the caloric particles, and this holds the solid substance together. In the fluid state, the repulsive forces between the caloric particles become stronger but not yet dominant. Finally, in the gaseous state these repulsive forces become completely dominant, so that the attractive forces between the material particles are negligible. These principles of the theory of caloric imply that the thermal expansion of gases, that is, the dependence of their volume on their temperature, should be entirely caused by the increase of caloric and, hence, should be *the same* in all gases, independent of their material nature. This was a prediction of a *novel* phenomenon, which was confirmed independently by John Dalton and Joseph-Louis Gay-Lussac in 1802 in their phenomenological gas law, which asserts that under constant pressure, the volume of any (sufficiently ideal) gas is proportional to its absolute temperature.

The modern theory of thermal expansion is based on the *kinetic theory of gases*. According to this theory, a special immaterial substance such as caloric does not exist; temperature is nothing but the mean kinetic energy of the molecules. In the gaseous state, the distance between the molecules of the gas is so large that the volumes of its molecules and the attractive forces between its molecules are negligible. Therefore gases with equal temperature under the same pressure will have the same volume.

In the case of caloric theory we enter the same situation as in phlogiston theory, namely that an empirical identification of caloric particles was impossible. The crucial theoretical expression of caloric theory, which was empirically measurable and did the relevant work, was *the amount* of caloric particles contained in a substance that *repel* each other. What corresponds to this theoretical expression in modern theory is the *mean kinetic energy* of the molecules. So we have the following correspondence relation:

Correspondence relations between the caloric theory and modern physical chemistry:

The amount of caloric particles in a substance X = the mean kinetic energy of X's molecules.

The repulsion force between the caloric particles in X = the expansion forces of X's molecules, which in the gaseous state correspond to the pressure of X.

These correspondence relations explain the strong success of the caloric theory in the described domain of application (though the caloric theory was less successful in other domains of application; cf. Psillos [1999], 115ff.).

The *quantitative* version of this correspondence follows from the famous *Maxwell–Boltzmann correspondence* between the (absolute) temperature T of a gas and the mean kinetic energy of its molecules:

$$A \rightarrow \left(T = \frac{2 \cdot N_A}{3 \cdot R} \cdot \frac{m \cdot \overline{v^2}}{2} \right). \tag{19}$$

Here A describes the empirical conditions for approximately ideal gases (high temperatures), $\frac{m \cdot \overline{v^2}}{2}$ is the mean kinetic energy of one gas molecule; R (Rydberg's constant) expresses the uniform rate of thermal expansion, and N_A is Avogadro's number. The standard derivation of (19) is remarkably similar to the proof of our general logical theorem: the laws for the average pressure of the gas molecules (which are derivable from the theory of mechanics) entail that the complex mechanical expression $\frac{2 \cdot N_A}{3 \cdot R} \cdot \frac{m \cdot \overline{v^2}}{2}$ has the *same measurement conditions* as the temperature T in the phenomenological gas law.[12] From (19) together with the quantitative law of caloric theory (20), which relates the amount of caloric χ with absolute temperature

$$A \rightarrow (\chi = k \cdot T) \tag{20}$$

(where k is a gas-independent proportionality constant), we obtain the following quantitative correspondence between caloric and mean kinetic energy:

$$A \rightarrow \left(\chi = \frac{2 \cdot k \cdot N_A}{3 \cdot R} \cdot \frac{m \cdot \overline{v^2}}{2} \right). \tag{21}$$

I merely indicate how my account applies to the *Fresnel–Maxwell case*. Here, the domain of application A are incident, reflected and refracted light beams

[12] Written as a measurement law for temperature, the phenomenological gas law has the form (i) $A \rightarrow (T = \frac{p \cdot V}{n \cdot R})$ (V volume, p pressure, n mole number). The mechanical law for the pressure of the gas molecules can be rewritten into the isomorphic form (ii) $A \rightarrow (\frac{2 \cdot N_A}{3 \cdot R} \cdot \frac{m \cdot \overline{v^2}}{2} = \frac{p \cdot V}{n \cdot R})$. (i) + (ii) entail the Maxwell–Boltzmann correspondence (19). (Cf. Barrow [1966], Chapter 2.2). The phenomenological gas law (i) contains the strong success shared by caloric theory and kinetic gas theory. Logically speaking, it unifies BR_is of the form $A_i \rightarrow (T = \frac{p \cdot V}{n \cdot R_i})$ for different *kinds* of gases *i* by the postulate of a *uniform* thermal expansion rate: $\forall i,j: R_i = R_j$, which was crucial for the entailment of novel predictions.

at the interface of two media, their angles and their intensities. Both theories entail the same non-theoretical equations relating these expressions. Worrall ([1989]) and Psillos ([1995]) have worked out (though they disagree on other points) that the non-theoretical notion of the intensity of light has a direct theoretical correlate in both theories: in Fresnel's mechanical wave theory it is directly proportional to the square of the oscillation velocity of the ether molecules (Psillos [1995], p. 36), and in Maxwell's electromagnetic theory of light it is directly proportional to the square of the oscillating strength of the electromagnetic field (cf. Young and Freedman [1996], p. 1036f). So we have a correspondence relation between the oscillation velocity of ether molecules in Fresnel's theory and the oscillation strength of the electromagnetic field in Maxwell's account. The application of my account to further examples (e.g., Kepler vs. Newton, classical vs. quantum mechanics) remains to be carried out in the future.

6 Discussion of the Correspondence Theorem: Objections and Replies

Objection 1: The non-realist (e.g., van Fraassen [2006], p. 298, second section) will also grant that the old theory T and the contemporary theory T^* are logically related to each other insofar as a certain class of correct empirical predictions, denoted by S (for 'success'), is logically entailed both by T and by T^*. Does your correspondence theorem say more than this?

Reply 1: Yes, in *many* ways. That the two theories T and T^* entail the same (true) empirical predictions S implies *nothing* about a logical relationship between the theoretical parts (axioms) of T and T^*, let alone a logical relation between some T-theoretical and T^*-theoretical expressions. In contrast, the correspondence theorem establishes an *equivalence* between the T-theoretical expression φ and some T^*-theoretical expression τ^* in the partitioned domain $A = A_1 \cup \ldots \cup A_n$ of T's success. This surprising result depends crucially on the satisfaction of the two explained requirements (1) and (2) concerning the predecessor theory T: (1) the class S is a case of strong (potential) success, which by definition means that S is a set of conditionals of the form $(A_i \wedge \pm R_i) \rightarrow (A_j \rightarrow \pm R_j)$ for $i \neq j \in \{1, \ldots, n\}$; and (2) this strong (potential) success S is *yielded* by a (possibly compound) theoretical expression φ of T by means of the bilateral reduction sentences BR(T): $(A_i \rightarrow (\varphi(x) \leftrightarrow R_i))$ $(1 \leq i \leq n)$, which follow from T. From the satisfaction of requirements (1) and (2), together with the satisfaction of the rather self-evident requirements (0) (T and T^* share their non-theoretical terms), (3) (T^* entails S in a T^*-dependent way) and (4) (T and T^* are causally normal), the correspondence theorem follows.

Objection 2: Requirements (1) and (2) express intricate *formal* conditions, but what is their *philosophical substance*?

Reply 2: Requirement (1) of strong (potential) success expresses the capability of T to produce *novel* predictions: from what was observed in one domain of application ($A_i \wedge R_i$), one may infer what will happen in another and potentially novel domain of application ($A_j \rightarrow R_j$). Requirement (2) expresses in its *causal* interpretation that the entity (or property) posited by φ figures as a *common cause* of the mutually correlated regularities or dispositional properties 'if A_i, then R_i' (i.e., $\varphi(x) \rightarrow (A_i \rightarrow R_i)$), in a way that allows φ's empirical measurement (i.e., $A_i \wedge R_i \rightarrow \varphi(x)$). In (Schurz [2008], Section 7.1) it is argued at length that it is exactly this *common cause* property that distinguishes *scientific* abductions to theoretical entities from *speculative* abductions: while typical speculations postulate for each new phenomenon a new kind of theoretical cause, science introduces new theoretical entities only if they figure as common causes of *several* intercorrelated phenomena. If φ were characterized by only *one* BR-sentence $A_1 \rightarrow (\varphi(x) \leftrightarrow R_1)$, the derivation of the correspondence principle would be impossible. What the *proof* of the correspondence theorem amounts to from the causal viewpoint is the following: if φ of T yields T's strong (potential) success by playing the role of a measurable common cause of the correlated dispositional properties 'if A_i, then R_i', and T^* entails T's success in a T^*-dependent way (where both T and T^* are causally normal), then also T^* must contain some expression τ^* whose designatum figures as a measurable *common cause* of the correlated dispositional properties. This is the reason why in the domain A, φ's reference can be understood as that entity/property which in T^* is denoted by τ^*.

Objection 3: What is the exact part of the old theory T which can be said to have reference within the ontology of T^*, or which is preserved in T^*? And whatever this 'part' may be, isn't its determination only possible *ex post*, i.e., when the new theory T^* is already known, by way of ad hoc manoeuvres? (this is van Fraassen's challenge in [2006], p. 290).

Reply 3: No. According to the correspondence theorem, the respective 'part' of T can be determined *in advance*, without knowing the new theory T^*. *Every* theoretical expression φ of a (causally normal) theory T which satisfies the requirements (1) and (2), *together* with the bilateral reduction sentences BR(φ,T) for φ, is such a 'part' of T. It corresponds to *some* T^*-expression τ^*, for *any* (causally normal) theory T^* which entails T's strong (potential) empirical success in a T^*-dependent way. Hence if the theory T^* is (approximately) true, then the 'part' of T expressed by the bilateral reduction statements BR(φ,T) and everything that follows from it is indirectly true in the sense explained in Section 1, i.e., true under the φ–τ^* *reference shift*, which assigns τ^*'s interpretation to φ. I regard this notion of 'indirect truth' as a scientifically important kind of 'partial' truth.

Let me compare the 'reference shift' of my correspondence theorem with Psillos' description of a reference shift. According to Psillos ([1999], p. 296), φ and τ^* denote the same entity iff (a) their putative reference plays the same causal role in a network of phenomena, and (b) the core causal description of τ^* takes up the kind-constitutive properties of the core causal description of φ. In my account, φ and τ^* can be understood to denote the same entity iff both φ and τ^* play the role of a measurable common cause of a set of empirical regularities as described by the same (or: isomorphic) bilateral reduction sentences BR(T) and BR(T*). So, instead of Psillos' 'same causal role' condition, I have the more precise 'isomorphic common cause' condition. However, I do not assume that the 'kind-constitutive properties' of the core causal description of φ in T are preserved in T*, since many of these kind-constitutive properties of φ will belong to φ's *inner structure* which—as I have argued—is typically not preserved in T*. In conclusion, Psillos' preservation-requirement for kind-constitutive properties seems to be rather problematic: contra Psillos ([1999], p. 158) I think that even the luminiferous ether had one kind-constitutive property which was *not* taken over by Maxwell's electromagnetic theory, namely its composition of ether molecules, whose displacement velocity is proportional to the amplitude of light.

The content-part of T which is preserved in T* is BR(T) and what follows from it in domain A. This is logically underpinned by the observation that the translation of φ into τ^* *preserves* the following kinds of *consequences*, where $S(\varphi)$ is a sentence about φ, and $S(\tau^*)$ is the analogous sentence about τ^* (obtained by substituting τ^* for φ)[13]:

(Content–preservation) If A, $BR(T) \Vdash S(\varphi)$, then A, $BR(T^*) \Vdash S(\tau^*)$.

However, the content part of T which is preserved in T* is not a simple *conjunctive* part of T's axioms. It is a 'structural' content-part of T. This contrasts with Psillos' 'theoretical constituents', which seem to be such conjunctive parts of the given theory (cf. [1999], p. 80f). Lyons ([2006]) has shown that, contra Psillos, many conjunctive parts of Kepler's theory T, which were used in the derivation of Kepler's area law, can in no way be said to be preserved in the ontology of modern physics. The preserved content-part of T in my account is not beset by Lyons' objections. The shifted interpretation of BR(T) within T*'s ontology presupposes forgetting T's hypotheses about φ's inner structure and considering only φ's relations to the observable phenomena. Properly speaking, the preserved content-part BR(T) has to be understood as a Ramsey-type

[13] The proof assumes that φ's possible extensions are logically unrestricted in the sense explained in footnote 11. *Proof by contraposition:* Assume A, $BR(T^*) \nVdash S(\tau^*)$ does *not* hold. Then there is a model M = (D,I) satisfying A and $BR(T^*)$ but falsifying $S(\tau^*)$. By our assumption, we can expand M to M′ for φ by assigning the extension $I(\tau^*)$ to φ. The model M′ verifies BR(T) and falsifies $S(\varphi)$. So, A, $BR(T) \nVdash S(\varphi)$ would not hold, too.

existential quantification in which one quantifies over φ as a whole: if BR(T) has the form S(φ), where φ has the form f(μ) (e.g., release-of-phlogiston), then what is preserved within T*'s ontology is not $\exists$XS(f(X)) but the logically weaker $\exists$XS(X). It is an open question that deserves future investigation whether this preservation of structural content-parts can be adequately captured by one of the existing accounts of *truthlikeness*.

Objection 4: Is the correspondence theorem analytic or synthetic? And if it is analytic, how can that be? For example, a 'witchcraft theory', which has got some empirical predictions right, cannot correspond to any entity posited by modern science—so it seems that the correspondence theorem can't hold for analytic reasons alone.

Reply 4: The correspondence theorem has the form 'if assumptions, then correspondence'. It holds on *analytic* reasons, because it is a *logical* theorem. But the truth of the 'assumptions'—which consist in the satisfaction of requirements 0 to 4—is, of course, a *synthetic* question. And my claim that these requirements are often (and typically) satisfied for scientific theories is an *empirical–historical discovery*. Therefore, my conclusion that usually (and typically) certain theoretical 'parts' of past theories are *preserved* in later theories depends on analytic argument as well as on empirical–historical discovery.

The worry about correspondence for 'witchcraft theories' is related to the following challenge of Laudan ([1984], p. 160), Psillos ([1999], p. 292) and van Fraassen ([2006], p. 301): a notion of correspondence would certainly be *trivial* if it allowed all sorts of outdated theories, even medieval astronomy or Aristotelian physics, to correspond 'in some way' to contemporary science. This complaint is correct, but it does not apply to my account, which is based on the condition that the outdated theory T satisfies the requirements (1) and (2). Aristotelian physics is a case in point. Van Fraassen remarks that 'Aristotelian science felt free to multiply theoretical terms for so-called occult properties' ([2006], p. 281). More specifically, Aristotelian physics did not explain different kinds of motion in terms of *common causes,* but posited a distinct cause for each kind of motion: 'earthly' sublunary bodies (when free of 'violent' forces) tend to fall down, because their 'natural place' is the centre of the earth; light sublunar entities (like flames) move upwards because their natural place is the heaven; etc. Logically speaking, the theoretical concept $\varphi_i(x)$:= 'attraction of x towards the natural place φ_i' is not characterized by a multitude of bilateral reduction sentences; rather each special theoretical term φ_i is characterized by just one bilateral reduction sentence of the form: $A_i \rightarrow (\varphi_i(x) \leftrightarrow R_i)$, where A_i describes the type of moved entity, φ_i the natural place of x, and R_i the direction of the movement. Hence conditions (1) and (2) do *not* apply to Aristotelian physics, and therefore correspondence to, e.g., the gravitation force of modern physics *cannot* be established. The same is true for most speculative theories of the medieval era.

On the other hand, *whenever* the outdated theory T satisfies requirements (1) and (2), correspondence to a contemporary theory T* (satisfying the other requirements) exists, *even if* the outdated theory uses terms with 'witchcraft-like' meaning-components. Here is an example (cf. Jammer [1957], Chapter 5): in his early astronomical writings Kepler stipulated a *force*, which holds the planets on their orbits around the sun, but he conceived this force as a kind of *soul* ('anima'). Later Kepler argued that this force is mathematically describable. Newton, who found out the correct mathematical description of the gravitational force, was aware that gravitation, being an action-at-a-distance, resists a mechanical explanation. Newton continued to be agnostic about the 'nature' of gravitation and emphasized that his theory is compatible with a material as well as a spiritual (non-material) conception of it (cf. Jammer [1957], p. 137). The message is this: *even if* gravitational force is interpreted as a spiritual (i.e., 'witchcraft-like') entity, this does not prevent its correspondence to the modern classical concept of force, as long as it is mathematically described in a way that yields correct predictions of movements of bodies in gravitational fields.

7 Consequences for Scientific Realism and Comparison with Other Positions

Does my correspondence theorem support scientific realism? If yes, which kind of scientific realism? Does it improve the justifications of scientific realism which have been proposed in philosophy of science? In this section, I will answer these questions. By way of comparing my account to other accounts, I will try to show that the answer to the first and the last question is positive.

7.1 Comparison with constructive empiricism

In van Fraassen's account, what is preserved in theory-succession is merely the entailment of strong empirical success. In contrast, my correspondence theorem entails that also on the theoretical level, a structural part of T's content (as specified above) is preserved in T*. Van Fraassen ([2006], p. 298) thinks that the mere preservation of empirical predictions from T to T* is sufficient to explain why the old theory T was successful. In my view, this is not enough for an explanation of T's success. My account explains why the theory T had strong empirical success as follows: because *that* expression of T which yielded T's success corresponds to a theoretical expression of T* that yields the same success within T*.

On the other hand, what my account has *in common* with van Fraassen's account is a sceptical attitude concerning the unrestricted Putnam–Boyd version of the *no miracles* argument, NMA for short (cf. van Fraassen [2006], p. 296). There are both strong *theoretical* and strong *empirical* arguments that

undermine the NMA. The major empirical argument against the NMA is Laudan's pessimistic meta-induction from which this paper started in Section 1. The major theoretical argument against the NMA in my view is the fact explained in Section 6 (replies 1 and 2): if the theory T does not satisfy requirements (1) and (2), but *speculates* for each kind of phenomenon a special cause, then there is simply *no way* to infer from its empirical success something about the truth-status of its theoretical part. Therefore I think the unrestricted version of the NMA is simply *false*. Only a restricted version of it is tenable. My account attempts to show *why* and under *which* restriction the NMA is defensible. This brings me to the next subsection.

7.2 Major difference from standard scientific realism

Most scientific realists base their arguments on some version of the NMA. In contrast, my account *in no place* presupposes the NMA or some other form of inference to the best explanation (IBE for short). My account is based on an *analytic* theorem. Together with empirical–historical evidence concerning the truth of the theorem's assumptions, my account establishes *independently* of the NMA or some other form of IBE as to *why* a specifically *restricted* version of the NMA can be defended. Herein I see the major difference and advantage of my account to scientific realism.

In Section 7.1, I have sketched two strong arguments against the reliability of the NMA in its unrestricted form. There are also other reasons why a justification of scientific realism that does not presuppose the NMA is desirable. For example, many authors have pointed out that the standard justification of the NMA by an IBE is *circular*, because the reliability of IBE is in turn justified by the NMA (cf. the discussion in Psillos [1999], pp. 81ff). Some NMA-adherents have replied that the 'rule-circularity' involved in the justification of NMA is a *nonvicious* kind of circularity. I strongly doubt this point: Salmon ([1957], p. 46) and Achinstein ([1974], p. 137) have convincingly demonstrated that absurd rules such as the rule of anti-induction, or the obviously invalid rule 'No F is G, Some Gs are Hs; therefore: All Fs are Hs' can be justified in a 'rule-circular' manner. Hence, 'rule-circularity' is as vicious as 'premise-circularity'. In conclusion, there are many independent reasons why a justification of scientific realism that does not presuppose the reliability of NMA is desirable—and my account attempts to offer such a justification.

7.3 From minimal realism and correspondence to scientific realism

How does my correspondence theorem justify scientific realism without presupposing the NMA? First of all, the correspondence theorem *alone* justifies only a *conditional realism: if* one assumes the (approximate) realistic truth of

the presently accepted theory T^*, then also outdated theories T satisfying the requirements (1) and (2) contain a (theoretico-structural) content-part which is indirectly true and hence partially true (since 'indirect truth' is an important case of 'partial truth'). This conditional realism *weakens* Laudan's pessimistic meta-induction. But conditional realism *alone* is not sufficient to justify scientific realism. For someone who, on independent epistemological grounds, does not believe that contemporary or future scientific theories are approximately true, this conditional realism cannot tell anything about the partial truth of earlier theories.

But the situation changes if one makes in addition the following assumption of *minimal realism* (MR):

(MR): The observed phenomena are *caused* by an external reality whose structure can *possibly* be represented in an approximate way by an ideal theory T^+, which is causally normal, entails the observed phenomena in a T^+-dependent way, and whose language is in reach of humans' logico-mathematical resources.

(MR) is a *minimal* realist assumption because it merely says that an approximately true theory describing the reality that causes the observed phenomena *is possible*—independent of whether humans will ever be able to find this theory. Together with (MR), the correspondence theorem *entails* that the abductive inference from the strong empirical success of theories to their partial realist truth is *justified*. For if (MR) is true, then there exists an approximately true *ideal theory* T^+, which need not be known to us and preserves all of the strong empirical success that our accepted theories have. So the correspondence theorem implies that every (theoretico-structural) content-part of our contemporary theories satisfying the requirements (1) and (2) corresponds to a content part of the ideal theory T^+, and hence is indirectly true. By this line of argumentation, the *abductive* inference from strong success to partial (approximate) truth is replaced by the *analytic* inference from (MR) plus strong success to partial (indirect) truth. In this way my account provides an *independent* justification of NMA and IBE.

7.4 Comparison with particular realistic positions

What kind of scientific realism is supported by (MR) plus the correspondence theorem, in comparison with other positions? This kind of scientific realism is weaker than the scientific realism defended by Putnam ([1975]), Boyd ([1984]) and Psillos ([1999]) in three respects. *First*, it does not assert partial realistic truth for all kinds of predictively successful theories, but only for theories satisfying requirements (1) and (2). *Second*, the partial truth is not asserted for some conjunctive part of the (axioms of the) theory, but for a part that

is 'structural' in nature (cf. reply 3 of Section 6). *Third*, the partial truth has been called *indirect* truth because it is obtained from a φ–τ^* reference shift, which treats the theoretical expression φ of T as primitive, abstracting from the theory's hypotheses about φ's inner structure. For this reason, the scientific realism supported by my account is compatible with a certain amount of *empirical underdetermination*, even in our most advanced theories (quantum mechanics is a case in point; cf. Ladyman [1998], pp. 418f).

On the other hand, the kind of scientific realism supported by my account is stronger than that of Worrall ([1989]). It shares with Worrall the 'structural' nature of that part of the theory that is asserted to be (indirectly) true, but does not assume that this structural part refers merely to a *mathematical* structure; it corresponds rather to a certain *real* structure among *real* entities or properties. My kind of scientific realism is also stronger than that of Carrier ([2004], pp. 154f), who merely assumes that strongly successful theories have got the classification of the empirical phenomena right, while I argue that also a certain content-part of their theoretical superstructure is (indirectly) true.

The fact that the kind of scientific realism supported by my account is weaker than that of Putnam, Boyd or Psillos may be seen as a disadvantage, but this disadvantage is the price of the two major advantages of my account. First, my justification of scientific realism does not presuppose the reliability of NMA or IBE. Second, in my account the 'structural' parts of a strongly successful theory can be specified *in advance* (independent of the successor theory) and with *logical precision*. These are my reasons for hoping that my account of scientific realism is better supported and more credible than those strong accounts of scientific realism that are based on dubious versions of the NMA.

Acknowledgements

For valuable comments I am indebted to Timothy Lyons (in particular for Section 6), Ioannis Votsis, Martin Carrier, Paul Hoyningen-Huene, James Ladyman, Theo Kuipers, Hannes Leitgeb, Johan van Benthem, Ulises Moulines and an unknown referee.

Philosophy Department
University of Duesseldorf
Universitaetsstrasse 1, Geb. 23.21
D-40225 Duesseldorf, Germany
schurz@phil-fak.uni-duesseldorf.de

References

Achinstein, P. [1974]: 'Self-Supporting Inductive Arguments', in R. Swinburne (*ed.*), *The Justification of Induction*, Oxford: Oxford University Press, pp. 134–8.

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

Barrow, G. M. [1966]: *Physical Chemistry*, New York: McGraw-Hill.

Boyd, R. [1984]: 'The Current Status of Scientific Realism', in J. Leplin (*ed.*), *Scientific Realism*, Berkeley: University of California Press, pp. 41–82.

Carnap, R. [1936]: 'Testability and Meaning (Part I)', *Philosophy of Science*, **3**, pp. 419–71.

Carrier, M. [2001]: 'Changing Laws and Shifting Concepts', in P. Hoyningen-Huene and H. Sankey (*eds*), *Incommensurability and Related Matters*, Dordrecht: Kluwer, pp. 65–90.

Carrier, M. [2004]: 'Experimental Success and the Revelation of Reality: The Miracle Argument for Scientific Realism', in M. Carrier, G. Küppers, J. Roggenhofer and P. Blanchard (*eds*), *Knowledge and the World: Challenges Beyond the Science Wars*, Heidelberg: Springer, pp. 137–61.

Hintikka, J. [1988]: 'On the Incommensurability of Theories', *Philosophy of Science*, **55**, pp. 25–38.

Hoyningen-Huene, P. [1993]: *Reconstructing Scientific Revolutions: Thomas Kuhn's Philosophy of Science*, Cambridge: Cambridge University Press.

Jammer, M. [1957]: *Concepts of Force*, Cambridge, MA: Harvard University Press.

Ladyman, J. [1998]: 'What Is Structural Realism?', *Studies in the History and Philosophy of Science*, **29**, pp. 409–24.

Ladyman, J., and Ross, D. [2007]: *Every Thing Must Go. Metaphysics Naturalized*, (with D. Spurrett and J. Collier), Oxford: Oxford University Press.

Laudan, L. [1981]: 'A Confutation of Convergent Realism', reprinted in D. Papineau (*ed.*), *The Philosophy of Science*, Oxford: Oxford University Press, pp. 107–38.

Laudan, L. [1984]: 'Discussion: Realism Without the Real', *Philosophy of Science*, **51**, pp. 156–62.

Laudan, L. and Leplin, J. [1991]: 'Empirical Equivalence and Underdetermination', *Journal of Philosophy*, **88**, pp. 449–72.

Lyons, T. D. [2006]: 'Scientific Realism and the Stratagemata de Divide et Impera', *British Journal for the Philosophy of Science*, **57**, pp. 537–60.

Oxtoby, D., Gillis, H. and Nachtrieb, N. [1999]: *Principles of Modern Chemistry*, Orlando, FL: Saunders College Publishing.

McCann, H. G. [1978]: *Chemistry Transformed: The Paradigm Shift from Phlogiston to Oxygen*, Norwood, NJ: Ablex Publishing Corporation.

Papineau, D. [1996]: 'Introduction', in D. Papineau (*ed.*), *The Philosophy of Science*, Oxford: Oxford University Press, pp. 1–20.

Psillos, S. [1995]: 'Is Structural Realism the Best of Both Worlds?', *Dialectica*, **49**, pp. 15–46.

Psillos, S. [1999]: *Scientific Realism. How Science Tracks Truth*, London and New York: Routledge.

Putnam, H. [1975]: 'What is Mathematical Truth?', in H. Putnam (*ed.*), *Mathematics, Matter and Method*, Cambridge: Cambridge University Press, pp. 60–78.

Salmon, W. C. [1957]: 'Should We Attempt to Justify Induction?', *Philosophical Studies*, **8**, pp. 45–7.

Schurz, G. [2008]: 'Patterns of Abduction', *Synthese*, **164**, pp. 201–34.

Sneed, J. D. [1971]: *The Logical Structure of Mathematical Physics*, Dordrecht: Reidel.

Van Fraassen, B. [1980]: *The Scientific Image*, Oxford: Clarendon Press.

Van Fraassen, B. [2006]: 'Structure: Its Shadow and Substance', *British Journal for the Philosophy of Science*, **57**, pp. 275–307.

Votsis, I. [2007]: 'Uninterpreted Equations and the Structure-Nature-Distinction', *Philosophical Inquiry*, **29**, pp. 57–71.

Worrall, J. [1989]: 'Structural Realism: The Best of Both Worlds?', *Dialectica*, **43**, pp. 99–124.

Young, H. D., and Freedman, R. A. [1996]: *University Physics*, 9th edition, Reading, MA: Addison-Wesley.