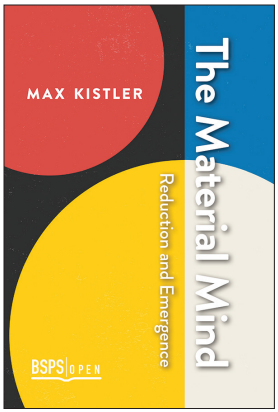
UNIVERSITY OF CALGARY
Press



THE MATERIAL MIND: REDUCTION AND EMERGENCE

Max Kistler

ISBN 978-1-77385-607-0

THIS BOOK IS AN OPEN ACCESS E-BOOK. It is an electronic version of a book that can be purchased in physical form through any bookseller or on-line retailer, or from our distributors. Please support this open access publication by requesting that your university purchase a print copy of this book, or by purchasing a copy yourself. If you have any questions, please contact us at ucpress@ucalgary.ca

Cover Art: The artwork on the cover of this book is not open access and falls under traditional copyright provisions; it cannot be reproduced in any way without written permission of the artists and their agents. The cover can be displayed as a complete cover image for the purposes of publicizing this work, but the artwork cannot be extracted from the context of the cover of this specific work without breaching the artist's copyright.

UNIVERSITY OF CALGARY
Press
press.ucalgary.ca

Acknowledgement: *We acknowledge the wording around open access used by Australian publisher,* **re.press**, *and thank them for giving us permission to adapt their wording to our policy* http://www.re-press.org

1

Unity of Science and Reduction

1. Introduction

Science aims to broaden and improve our knowledge of the world. Part of this knowledge consists of *descriptions* of things that exist, of events that happen, of processes that take place. But often we are not content with descriptions or with facts. We ask science to help us understand why things are as they are and why events and processes happen. There are two ways of providing us with such an understanding: by discovering the *properties* that objects, events, or processes possess and by knowing the *laws* that they obey by virtue of these properties. These two types of discovery go hand in hand: to hypothesize the existence of a property is to hypothesize the existence of a law (or a set of laws) that imposes constraints on the behaviour or evolution of what has the property. Making the hypothesis of the existence of a law means making progress in our understanding of the world in two ways. The law enables us to complete our knowledge of the properties implied by the law, and it enables us to understand the origin of the links between different things, facts, and events.[1] These links are manifested in regular associations of properties and regular successions of events. The link between properties and the laws in which they appear is more intimate than the empiricist tradition recognizes: possessing a natural property makes it necessary to obey the

1 Causey (1977, 17) counts laws, along with facts, among the objects of knowledge, whereas explanations and theories provide us with understanding. This is not necessarily incompatible with my assertion that laws already enable us to understand why regularities occur. Scientific understanding of the world is an iterative process, and what at a given moment is the object of knowledge can become the starting point for a new interrogation of its why.

laws in which it participates because these laws are constitutive of the identity of the property.[2]

But the discovery of properties and laws is only one step, albeit an essential one, toward a satisfactory understanding of nature. By postulating the existence of new properties that often are not directly observable, and new theoretical laws that these properties obey, theories deepen our understanding. According to the traditional view (shared by Duhem [1906] and logical empiricism), theories provide a unified understanding of a whole field of phenomena by explaining laws first established on the basis of experience and induction. The theory makes it possible to deduce laws previously discovered separately, in particular laws related to observable properties. These laws are called "experimental" (to refer to their origins in the experimental observation of regularities) or "phenomenological" (to point out that the properties concerned by these laws are at least partially observable) to distinguish them from theoretical laws.

A law gives us a unified understanding of a multitude of events. Hooke's Law states that the force exerted by a spring is proportional to its extension.[3] By making these two properties of extension and force appear to be linked by the law, the regularity of their association is understood as arising from a relationship between the properties themselves.[4] The singular events concerning springs can be explained on the basis of the hypothesis that they "fall under" this law or that they are "covered" by it.

The history of science has led not to the creation of a single theory but to a multitude of theories. Sometimes different theories deal with the same objects or phenomena. Gases are the subject of both thermodynamics and classical mechanics. Thermodynamics deals with the regular relationships between the macroscopic properties of gases, such as their temperature and

2 This thesis is developed in Kistler (2002a, 2005a). The apparent contingency of the laws is explained by the imperfection of the observed regularities. In terms of manifest properties, regularities are not perfect, such that there are exceptions. Albino crows are not black. The fall of an apple through the air is not uniformly accelerated. The Earth's orbit around the Sun is not a perfect ellipse. Laws impose constraints on properties that are not always manifest separately and that can be described as "capacities" or "dispositional properties." Only the results of their superpositions are manifest. The manifest properties resulting from these superpositions do not always show perfect regularity. I will develop this theme in Chapter 3.

3 This can be expressed concisely in a formula: $F = -kx$, where F represents the force exerted on the spring, k the spring constant, and x the extension of the spring.

4 This conception of laws is developed in Kistler (1999c, 2006d).

pressure, and the volume that they occupy. Classical mechanics, conversely, describes the regularities that govern the behaviour of the molecular components of the same gases. The need for understanding that drives scientific research prevents us from being content with simply juxtaposing the laws discovered by these two theories. After all, there is only one object with one behaviour. How is it that the laws discovered at the component level (in classical mechanics) and at the macroscopic gas level (in thermodynamics) do not contradict each other? The best way to deepen our understanding of this description of objects at several levels, by several theories, is to *reduce* one of the two theories to the other. Such a reduction, the logic of which we will study later, makes it possible to explain how an object can evolve both in accordance with laws that apply to the object as a whole and in accordance with laws that apply to its parts.

The overall aim of my inquiry is to assess the prospects of reducing cognitive psychology to neuroscience. My working hypothesis is that the conceptual problems that arise in this particular case are not fundamentally different from those that arise in the context of other reductions from one theory to another.

2. Deductive and Ontological Unification

One of the aims of science is to deepen our understanding of natural phenomena. When it is discovered that one theory can be reduced to another, this is an important step toward this goal: such a reduction shows that phenomena that have been explained by two independent theories have a common origin. Even if the reduction preserves the existence and some autonomy of the reduced theory, the reduction shows that the phenomena described by the reduced theory are not heterogeneous with respect to the phenomena described by the reducing theory. In this way, reduction gives rise to a "unification" of two previously disconnected domains of knowledge and explanation.

Accepting physicalism gives us a reason in principle to expect the discovery of reductions. According to materialism, everything that exists is material or composed solely of material constituents. *Physicalism* is a contemporary form of materialism, according to which everything that exists is composed exclusively of physical objects. A physical object is one whose properties are all physical, in the sense that they are properties whose identities are determined by the laws of physics. Fields, for example, are physical objects but do

not meet the traditional criteria of "material" entities. In this book, I accept physicalism as an empirical hypothesis justified by the success of science as a whole. If physicalism is true, then all theories are about physical objects. This means that

(1) all existing objects have only physical objects as parts and that

(2) all existing objects have properties of only two kinds:

  (a) physical properties and

  (b) properties determined by the physical properties of the object and the properties of its components.

It is useful to distinguish between *explanatory* and *ontological* unification. Each reduction gives rise to an explanatory unification. The Newtonian theory of gravitation, for example, provides a unique framework for explaining both the free fall of a massive body near the Earth's surface and the orbit of a planet around the Sun. Before the Newtonian reduction, these two explanations required the resources of two independent theories — the Galilean theory of free fall and the Keplerian theory of planetary motion — whereas only one was needed afterward. This simplification of the premises necessary to explain apparently heterogeneous phenomena deepens our understanding of phenomena because it makes them appear to belong to a single type (see Kitcher 1989). Reduction consists of showing that all of the constituent laws of the reduced theory derive (i.e., can be deduced) from the laws of the reducing theory, together with certain initial conditions.[5] For example, the reduction of the first Keplerian law of planetary motion (according to which planetary orbits have elliptical shapes) by the Newtonian law of gravitational attraction is based on three presuppositions. First, in the context of calculating their gravitational interactions, the Sun and the planets can be taken to be unextended points, with their masses situated at those points. Second, the force of gravitational attraction between two point masses is a central force that decreases as the inverse of the square of their distance. Third, the only force determining the orbit of planet m is the gravitational force between the Sun and m.

5 I will return to this derivation later in this chapter.

The fact that reductions always lead to explanatory unification is not controversial. However, it is difficult to explain the exact source of this deeper understanding. According to the tradition of logical empiricism, the derivation of reduced laws from the reducing theory must take the form of a deductive-nomological explanation. This is simply a consequence of the more general doctrine that any scientific explanation must take that form. In this model — thoroughly developed by Nagel (1961) — explanatory unification requires deductive unification. Causey (1977) and Hooker (1981) understand the *explanatory* unification achieved by a reduction in terms of *ontological* unification rather than just deductive unification; reduction often simplifies the ontology. Following Quine, we can consider that theories convey ontological commitment: if a theory is true, then all types of entities over which the axioms and theorems of the theory quantify exist.[6] Belief in the existence of entities of these types is justified to the same extent as belief in the truth of the theory. It makes sense to use the criterion of ontological commitment in a less restrictive form than Quine himself does: the mere fact that a scientific theory successfully introduces a property into the description of its models or into its axioms and theorems provides a reason (fallible of course) to believe in its existence.[7] Because of a reduction, the number of types of entity to whose existence one is committed according to the theories accepted decreases with the number of independent theories. Before the reduction of Mendelian genetics to molecular biology, the former included an ontological commitment to the existence of genes as a primitive type of entity. After the reduction, the ontology is simplified; genes are no longer considered a distinct type of entity. According to the new theory, their causal role is played by biological molecules, first and foremost DNA, and by a number of complex mechanisms that enable them to be replicated, to recombine in sexual reproduction, and to express themselves in the phenotype (see Mossio and Umerez 2014). Before the reduction of the temperature of a gas to the average kinetic energy of the molecules that make it up, gases and their properties, such as pressure and temperature, were fundamental entities in whose existence it

6 "We may be said to countenance such and such an entity if and only if we regard the range of our variables as including such an entity. To *be* is to be a value of a variable" (Quine 1976, 199). See also Quine (1939).

7 It is not necessary, as Quine's original criterion suggests, for the axioms of the theory, or for the description of its models, to quantify over the predicates that express these properties. See Kistler (2012, 2016, 2020).

was reasonable to believe, insofar as the thermodynamics of gases was taken to be true. After the reduction of thermodynamics to classical mechanics, the ontology is simplified by the elimination of the gas as a fundamental type of entity, the fundamental entity now being the set of molecules. We will see that it is sometimes difficult — and often controversial — to judge whether such an elimination of a fundamental type of entity is equivalent to its elimination altogether or whether the new theory, having achieved reductive unification, is still ontologically committed to its existence as a "derived" entity.

A reduction achieves its aim of unifying the representation of the world by bringing together two theories that had distinct domains before reduction. This unification is achieved by providing principles that allow conclusions to be drawn about the objects of the reduced theory, based on premises formulated in the language of the reducing theory. From a premise concerning the average energy of the molecules contained in a sample of gas, the reduction of thermodynamics to classical mechanics allows us to draw a conclusion about the temperature and pressure of this sample. However, the latter concepts apply to macroscopic objects and the former concepts to microscopic objects, one smaller than the other by several orders of magnitude. Similarly, the reduction of elementary learning to neurophysiology makes it possible to draw conclusions about an animal's cognitive state of conditioning from information about the state of the animal's microscopic components. For example, it is possible to conclude from a premise that relates to the change in the conformation of Ca^{2+} channels in certain presynaptic axonal endings of nerve cells in an individual of the species *Aplysia californica* that this individual is in a state of habituation or, conversely, of sensitization in regard to the siphon withdrawal reflex following stimulation of its tail.[8] This might seem to be surprising given that the premise concerns microscopic objects (membrane proteins), whereas the conclusion concerns a disposition to the behaviour of a macroscopic animal. How does the reduction bridge the distance between the domains of such disparate objects of discourse?

8 I will develop this example of the reduction of cognition to neurophysiology later in this chapter, in section 7.

3. The Deductive-Nomological Model of Reduction

Let us begin by examining the form that this question takes within the context of the now classic theory of reduction between theories proposed by Nagel (1961). His analysis of reduction presupposes the framework of the deductive-nomological (D-N) approach to scientific explanation that — since its original proposal by Hempel (1942) and Hempel and Oppenheim (1948) — was supposed to cover not only the explanation of particular facts but also the explanation of laws. According to Nagel, the reduction of a theory, called "secondary," to a more fundamental theory, called "primary," is a scientific explanation in the sense of the D-N model. According to that model, this explanation takes the form of a deduction of the laws of the secondary theory from the laws of the primary theory (Nagel 1961, 338).

Nagel distinguishes "homogeneous" and "heterogeneous" reductions. In a homogeneous reduction, the primary (reducing) and secondary (reduced) theories share the same vocabulary for describing objects in their respective domains. In a heterogeneous reduction, the secondary theory contains primitive descriptive terms that do not belong to the vocabulary, primitive or derived, of the primary theory. The reduction of Galileo's laws of the free fall of objects near the Earth's surface to Newtonian laws of mechanics and gravitation is an example of homogeneous reduction: "Although the two classes of motions are clearly distinct, no concepts are required for describing motions in one area that are not also employed in the other" (Nagel 1961, 339). The reducing theory, like the reduced theory, studies the movements of macroscopic bodies.

Yet the reduction of the macroscopic thermodynamics of gases to classical mechanics is a case of *heterogeneous reduction*. Temperature is a fundamental concept required to describe the objects of the secondary science (thermodynamics), but it is not part of the conceptual repertoire of the primary theory (classical mechanics), which describes the movements of the molecules that make up gases and their interactions. The heterogeneity of the descriptive vocabulary and conceptual apparatus of the primary and secondary theories is at the origin of an "acute sense of mystification" (Nagel 1961, 340) when the reduction of one to the other has been achieved. How is it possible to deduce laws that bear on macroscopic objects and describe links between their macroscopic properties, such as temperature and pressure, from laws that deal with an entirely different domain of objects, smaller by several

orders of magnitude? How is it possible to establish a reductive link between these theories when the objects belonging to their respective domains do not share the relevant properties (individual molecules can possess neither temperature nor pressure)? From the point of view of logic, such an explanation by reduction seems to be impossible:

> If the laws of the secondary science contain terms that do not occur in the theoretical assumptions of the primary discipline [i.e., if the reduction is heterogeneous], [then] . . . the logical derivation of the former from the latter is *prima facie* impossible. The claim that the derivation is impossible is based on the familiar logical canon that . . . no term can appear in the conclusion of a formal demonstration unless the term also appears in the premises. (Nagel 1961, 352–53)

The existence of heterogeneous reductions is not controversial, so they must be possible. According to Nagel, the reduction of heterogeneous theories appears to be impossible only insofar as their logical reconstruction omits an essential premise: the logical possibility of such a reduction presupposes the introduction of "assumptions of some kind . . . which postulate suitable relations between whatever is signified by 'A' [a term of the secondary science absent from the vocabulary of the primary science] and traits represented by theoretical terms already present in the primary science" (Nagel 1961, 353–54). Such an assumption is necessary in particular for the term "temperature," absent from the reductive theory, which applies to the molecules of a gas. Nagel calls this condition for the possibility of a reduction the "condition of connectability [sic]" (1961, 354). He puts forward two theses of great importance for my purposes.

First, Nagel departs from earlier work on reduction that stipulated that these *linkages* between concepts of the reduced and reductive theories must take the form of universal statements of biconditional form.[9] A statement of "biconditional" form indicates a necessary and sufficient condition, whereas a statement of "conditional" form indicates only a sufficient condition. "If it's raining, I'll take my umbrella" (conditional form with "if" or "if . . . then

9 Including his own earlier work (Nagel 1951) as well as Woodger (1952) and Kemeny and Oppenheim (1956).

. . .") indicates that the fact that it is raining is *sufficient* for my taking my umbrella. But this statement does not say that rain is a *necessary* condition: it is compatible with the fact that I take my umbrella even when it is sunny. Conversely, the statement in biconditional form (with "if and only if . . .") "I'll take my umbrella if and only if it rains" excludes the possibility that I will take my umbrella when it is sunny. The biconditional form, indicated by the expression "if and only if," indicates that the condition (the expression following the "if") is both sufficient and necessary for me to take my umbrella. In the expression "if and only if," "if" expresses the fact that it is sufficient; "only if" expresses the fact that it is necessary.

With regard to the link between the temperature of a gas and the average kinetic energy of the molecules of which it is composed, statement (B) in biconditional form expresses the thesis that the average kinetic energy is a necessary and sufficient condition for the corresponding temperature:

> (B) (for "biconditional") For any sample of ideal gas x, x has temperature T *if and only if* the molecules making up x have average kinetic energy $E_{kin}(T)$ proportional to T.[10]

However, Nagel (1961, 355n5) explains that the existence of conditions of biconditional form, as in (B), is not necessary for reduction. In order to recover the laws of the secondary theory from the laws of the primary theory, it is sufficient to suppose that there is a *conditional* dependence of the macroscopic property on a microscopic property. We will see that this weakening of the conditions imposed on reductions leads to a deep modification of the concept of reduction.

> (C) (for "conditional") For any sample of ideal gas x, *if* the molecules making up x have an average kinetic energy E_{kin}, *then* x has a temperature $T(E_{kin})$ proportional to E_{kin}.

If the relationship is conditional (and not biconditional), then the fact that the molecules have the mean energy E_{kin} is *sufficient but not necessary*

10 This proportionality is expressed in the following formula: $E_{kin} = \overline{e_{\text{kin}}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$, where m is the mass of a molecule, E_{kin} is the mean kinetic energy of a molecule, k is Boltzmann's constant, and $\overline{v^2}$ is the mean square velocity of the molecules. I will come back to this reduction in detail later.

for the gas to have T. This means that it cannot be ruled out that something could have a temperature T for a reason other than having molecules with the mean energy E_{kin}.

The quantities of energy and temperature are linked by a numerical equality $E_{kin} = {}^{3}/_{2}\, kT$. It might therefore seem to be obvious that the dependence between T and E_{kin}, which appear on both sides of the identity symbol =, must be symmetrical, such as the biconditional relation (B), which stipulates that each of T and E_{kin} is necessary and sufficient for the other. However, this appearance is ungrounded. The equation $E_{kin} = {}^{3}/_{2}\, kT$ is neutral with regard to the question of the nature of the dependence between the properties E_{kin} and T themselves. The equation simply expresses the fact that the *numerical values* of these quantities are proportional. It would never occur to anyone to suppose that temperature is identical to kinetic energy, multiplied by *2/(3k)*, for the simple reason that it makes no sense to talk about the multiplication of properties.

An important reason for not requiring biconditional but only conditional links is that it allows the model to be applied to the reduction of *multi-realizable* properties. This is the case with temperature. The state of the microscopic components of an object can vary while its temperature remains the same. Moreover, temperature is a property shared by objects of very different compositions. Gases, and other bodies composed of atoms or molecules, have a temperature by virtue of the kinetic energy of the atoms or molecules of which they are composed. But there are objects that are not composed of molecules yet have a temperature: plasma and radiation. The fact that cognitive properties are multi-realizable (i.e., they can exist in animals of different species thanks to different neurophysiological bases) plays a crucial role in Fodor's (1974) argument for the irreducibility of cognitive properties. This argument is aimed more generally at establishing the irreducibility of the laws of what are known as the "special sciences" (i.e., sciences whose fields of application are more restricted than that of physics, which applies to any spatio-temporal object). However, as we will now see, Fodor's argument depends on the distinction between the biconditional form and the conditional form: multi-realizable properties are irreducible only if we require that a reduction presupposes the discovery of a *biconditional* linking principle between the reducing and reduced properties but not if we admit that in order to reduce it is sufficient to discover a *conditional* relation.

The anti-reductionist argument presupposes that linking principles are biconditional laws: "Bridge laws express symmetrical relations" (Fodor 1974, 129). The crucial point is the thesis that there can be no biconditional law between a multi-realizable property and the different properties realizing it. Let us assume that a given psychological property can be realized, at the neurophysiological level, in principle by an infinite set of structures. In this case, there can be no biconditional law linking the psychological property to neurophysiology since its neurophysiological term would be an open disjunction, with an indeterminate number of terms.

Let us take a closer look at the question of whether the form of the linking statements required for a reduction must necessarily be biconditional or whether a reduction can be achieved with conditional linking statements. The answer to this question depends on the answer to another question: does the reduction require the *derivability* of the laws of the reduced theory, or would a weaker relation of *connectability* be sufficient? The derivability of a law means that it is possible to deduce the law from the reducing theory. Yet, to ensure connectability, it is sufficient for there to be laws that establish a link between properties belonging to the reducing theory and properties belonging to the reduced theory. The existence of conditional bridge laws satisfies what Nagel calls "the condition of connectability" (1961, 354).

> A linking principle of conditional form is a law of nature of the form
>
> All N_1 are P,

where N_1 is a predicate of the reducing theory (e.g., neurophysiological) and P is a predicate of the reduced theory (e.g., psychological).[11] In what follows, the capital letter N represents neurophysiological properties, and the capital letter P represents psychological properties, and N^* and P^* refer to the same properties instantiated at a later time.

11 I will use this non-formal expression. The conditional form appears explicitly in the logical form of the statement: it is a universally quantified conditional: $(x)\ (Nx_1 \rightarrow Px)$, where $\rightarrow$ represents the conditional "if . . . then . . . ," which means "for any object x, if x is N_1, then x is P."

Why are laws according to which it is sufficient to have property N_1 in order also to have property P not sufficient to deduce logically a law of the reduced theory from laws of the reducing theory?[12]

Let us assume that "all N_1 are N_1*" and "all N_2 are N_2* et cetera are neurophysiological laws. In Figure 1.1, these laws are represented by the horizontal arrows at the bottom. Let us suppose that the linking principles also have a conditional form: "All N_1 are P," "all N_2 are P," "all N_1* are P*," and so on. These linking principles correspond to the dotted lines (in vertical or oblique direction) in Figure 1.1, which indicate that a neurophysiological property N is sufficient for a psychological property P.

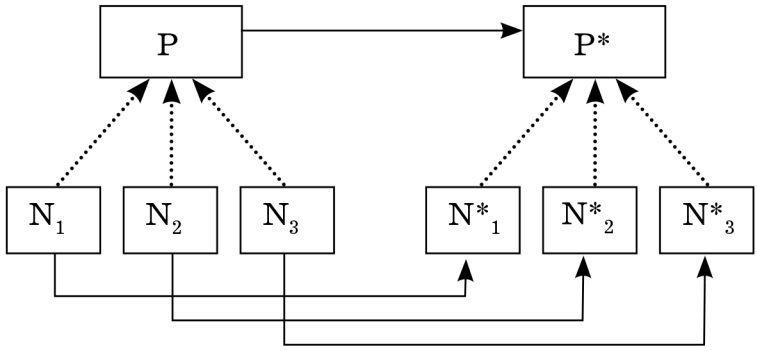


Figure 1.1 Reduction of laws involving multi-realizable properties. Modified from Fodor (1974, 139).

The psychological law "all P are P*" is represented by the top-level horizontal arrow. The question of reduction is to know under which conditions it is possible to deduce logically the psychological law from the neurophysiological laws and the linking principles. It appears that such a deduction is possible only if there are also "top-down" laws corresponding to arrows pointing downward: they would be laws stipulating that it is sufficient to have a certain psychological property P in order to have a certain neurophysiological property N_i. If there is such a "downward" law, for example "all P are N_5," then we can deduce, by transitivity, that

12 For the reasoning that follows, see Marras (2002, 248 ff.).

> All the P are N_5 (downward linking law), all N_5 are N_5* (reducing law), all N_5* are P* (upward linking law), therefore all P are P* (reduced psychological law).

Multiple realizability corresponds precisely to the absence of such downward laws. If the psychological property P can be realized, in different individuals, by different neurophysiological properties, N_1, N_2, N_3, et cetera, then having P is not sufficient for any one of them in particular.

Richardson is right to note that bridge laws of conditional form are sufficient for connectability, but he is wrong to say that such laws are also sufficient for derivability and therefore for reduction in the strongest sense. As he puts it, "derivability, with its explanatory parsimony, is adequately accounted for . . . if only we find sufficient conditions at a lower level of organization capable of accounting for phenomena initially dealt with at a higher level; and this . . . requires no more than a mapping *from* lower *to* higher level types and *not* a mapping from higher to lower level types" (1979, 548).

Nevertheless, Richardson expresses an important truth: when we discover the microscopic properties and mechanisms, for example biochemical, that underlie a given biological property, we consider that this discovery makes it possible to give a reductive explanation of the phenomenon even if, in other organisms, the microscopic properties and/or the mechanisms are different. Take, for example, the reductive explanation of the "signal" contained in a protein that enables the molecule to reach its destination within the cell. According to the "signal hypothesis," each protein synthesized in the cytoplasm by ribosome and RNA complexes possesses a property that determines its path to its functional destination: this property is the signal that enables the protein to be oriented. It appears that very different microscopic properties can "play the role" of such a signal in different organisms and for different proteins in a given organism. For example, "signal sequences [of amino acids with a protein] for insertion into the ER [endoplasmic reticulum] . . . may vary over 200% in length, apparently show diverse physical chemical interaction with membrane lipids . . . , may or may not be cleaved in serving their function depending on the signal sequence involved, and are sometimes species-specific in their functioning" (Kincaid 1990, 581). The discovery of each of the mechanisms allowing a microproperty underlying the signalling property of a protein to direct the protein toward its destination yields a reductive explanation. In this sense, it is correct to say, with Richardson, that

"alternative (possible or actual) mechanisms . . . do not prevent reduction" (1979, 549).

But we must be aware that the reduction achieved by the discovery of microproperties and the associated microscopic mechanisms is not a reduction in the sense of the logical derivability of the higher-level law, in the sense of Nagel. The tension between Richardson's statement that I just quoted and Kincaid's thesis that "such biochemical diversity underlying biological unity is the root obstacle to reduction" (1990, 583) is the result of different interpretations of reduction. Kincaid speaks of reduction in the sense of logical derivation, whereas Richardson's thesis can be defended if the word *reduction* is given a weaker meaning.[13] Insofar as it is sufficiently enlightening to derive, for one or another of the types of system possessing the reduced property, a law structurally equivalent to the reduced law $Sx \rightarrow S^*x$, of the form $P_ix \rightarrow P_i^*x$, from the reducing theory that correctly describes that type of system possessing *S*, we can consider that such a discovery constitutes a "reductive explanation" though not in the Nagelian sense of derivability.

In this sense, Ausonio Marras observes that it is legitimate to say that biological properties realized in different ways in different types of organisms can nevertheless be reduced, provided that we weaken the Nagelian conditions for reducibility, so that "we take the essential core of the reduction to be *not* the derivation of the *actual laws* of the target theory from the laws of the base theory, but merely the derivation of the *images* of such laws under appropriate boundary conditions" (Marras 2002, 249).[14]

In such a weakened conception of reduction, the fact that an indeterminate number of physical properties N_i underlie a mental property *P* does not prevent the reduction of *P*. Multiple realizability is compatible with the possibility of reducing the multi-realizable property to different realizing properties. This change of perspective has several important consequences. First, as Kim has pointed out, such reductions are only "local": "If each of the psychological kinds posited in a psychological theory has a physical realization for a fixed species, the theory can be 'locally reduced' to the physical theory of that species" (1992a, 19; 1993b, 328). Second, the concept of local reduction

13 Richardson (1979) himself does not seem to be aware of this: he wrongly claims that conditional bridge laws are sufficient for Nagelian derivability.

14 In this context, law A is called the "image" of law B if A and B are different but analogous in the sense that A shares (part of) the structure of B.

forces us to abandon the thesis (defended by Causey 1977) according to which the discovery of a reduction necessarily takes the form of the discovery of *the identity* of properties.[15] Conversely, the fact that there is no law of biconditional form on which the reduction is based constitutes a reason to deny that there is a (unique) reducing property identical to the reduced property.

Awareness of the possibility of local reductions is certainly an important step toward anchoring the mind in matter by showing that the multi-realizability of psychological properties does not present an insurmountable obstacle. There is a weaker model of reduction that does not require the deduction of the reduced theory from the reducing theory but only the discovery of local theories that apply to one or more realizations. The fact that the local reducing law $P_i x \rightarrow P_i{}^* x$ can be seen as reducing the law $Sx \rightarrow S^* x$ makes local reduction compatible with an important feature of reductions as they occur in the history of science. Most reductions are accompanied by *corrections* to the reduced theory. According to the model developed by Schaffner and others, what is deduced from the reducing theory is not the old theory that is the subject of the reduction but a new theory that resembles it.[16]

In Schaffner's (1967) terms, the theory $T_R{}^*$ derived from T_B (which stands for "base theory"; it is the reducing theory) must be in a relationship of "close similarity" to the original theory T_R that was to be reduced; the numerical predictions made from $T_R{}^*$ must be "very close" to those made from T_R; moreover, between the theory T_R to be reduced and the theory $T_R{}^*$ actually derivable from T_B, there must be a "strong analogy" or "positive analogy" (Schaffner 1967, 144). The abandonment of the deducibility requirement and its replacement by the requirement of the deducibility of a theory *analogous* to the reduced theory make Schaffner's conception compatible with multi-realizability. In the case of a multi-realizable property, there are several reducing theories, each serving as a basis for the deduction of a theory analogous to the reduced theory, without the different theories thus obtained being identical to each other.

15 According to Esfeld and Sachse (2011), the functional reduction of higher-level properties does justice to the existence of special sciences even though these higher-level properties are *locally* identical to physical structures.

16 The model developed by Schaffner (1967, 1993) to account for reductions that do not obey the Nagelian requirement of derivability was taken up by Hooker (1981), Churchland (1985), and Bickle (1998).

4. The Model of Reduction by Analogy

Historically, the new reduction model has been constructed to take account of the fact, noted by many authors, that reductions do not preserve the details of the theories that existed prior to the reductions. On the contrary, an important motivation for the search for a reduction is the corrective modification that the new theory T_B imposes on the old theory T_R, in terms of both observable predictions and theoretical assertions. This observation can lead to two conclusions. According to Popper (1957), Feyerabend (1962), and Kuhn (1962), the fact that "falsifications" or "paradigm shifts" lead to the adoption of a new theory incompatible with the old theory shows that it is not appropriate to speak of a reduction. Rather, it is the "replacement" or "elimination" of the old theory in favour of a radically different new theory. If T_R and T_B are incompatible or even incommensurable, one cannot be reduced to the other insofar as a reduction consists of justifying the old theory on new grounds. Among the cases traditionally referred to as "reductions," it is exceptional for the reduction to lead to the retention of the reduced theory in its precise form.

Popper (1972, 198–200) and Feyerabend (1962, 46–48) show this with the example of the corrective reduction of the Galilean law of free fall.[17] According to Galileo, a projectile launched from the surface of the Earth moves along a parabola. If its initial velocity is zero, then its free fall is a rectilinear and uniformly accelerated motion toward the centre of the Earth. However, this Galilean law cannot be derived as it stands from the Newtonian laws of motion. Newton showed that the trajectory of a projectile is elliptical (in the case of a spherically symmetrical attractor) and never strictly parabolic. However, the Newtonian theory also helps to explain the success of the Galilean law despite its falsity: when the total length of the projectile's trajectory is small compared with the Earth's radius, the parabolic shape is a good approximation to the elliptical trajectory.

The same conclusion can be drawn for Kepler's laws of motion of the planets around the Sun. Kepler's third law states that the ratio of the cube

17 Glymour (1970, 345) and Sklar (1993, 335) point out that, in order to derive Galileo's law, it is necessary to make the counterfactual hypotheses that there are no forces acting on the falling body other than gravitation (no friction in particular) and that the Earth is perfectly spherical. Popper (1972, 200) notes that Galileo's law can be deduced within the framework of Newtonian theory only if a false premise is added: the radius of the Earth is infinite. He adds that this premise is not only de facto false but also without sense since it has absurd consequences in Newtonian theory.

of the planet's mean distance from the Sun, *a*, to the square of the planet's period (i.e., the duration of one revolution around the Sun), *T*, is a constant (i.e., the same for all planets).

(K) $a^3 / T^2 = k$ (where k is a constant)

In Newtonian theory, we can only derive the following law, according to which this ratio, for a system composed of two point masses, is proportional to the sum of their masses m_1 and m_2.

(N) $a^3 / T^2 = k(m_1 + m_2)$ (where k is a constant)

From a Newtonian perspective, Kepler's original law is false for two reasons. First, the law (N) only applies to a system of two bodies and not when there are several planets that also influence each other. Not only is it false that there is only one planet, but also, if there were, then Kepler's law would be meaningless: its content is a regularity in the behaviour of all planets. If there were only one planet, then there would be no point in trying to establish any regularity between the planets. Second, (K) would be true (i.e., $m_1 + m_2$ would be a constant for all the planets) only if the masses of all the planets were the same or negligible compared with the mass, m_1, of the Sun.

According to Schaffner (1967), Churchland (1979, 1985), Hooker (1981), and Bickle (1998), in typical cases of reduction, such as that of the law of free fall or Kepler's third law, T_R is not derivable from T_B in any formal sense, and the primitive terms of T_R have no equivalents (nomologically co-extensional terms) in the language of T_B. In the situation that results from a "corrective" or "approximate" reduction, the new theory T_B can typically explain why the old theory was able to fulfill its explanatory and predictive role, although it is now considered to be false. Placing T_R and T_B in parallel allows us to understand in what sense T_R is an "approximation" of T_B. In some cases, one can indicate fictitious situations in which T_R can be deduced from T_B. One can then "obtain T_R from T_B, deductively: if one conjoins to T_B, certain contrary to fact premises . . . , one can obtain T_R" (Schaffner 1967, 138; variables modified).

In general, reduction leads to a modification of the reduced theory. For example, the reduction of the psychological theory of learning by the

neurobiological theory of Kandel (which I will present later in this chapter) has led to a change in the conception of the different types of learning: "Available evidence suggests that classical conditioning and sensitization are not fundamentally different, as is frequently thought, but rather the cellular mechanism of conditioning appears to be an elaboration of the mechanism of sensitization" (Hawkins and Kandel 1984, 389). In other words, "neurobiology may have discovered that simple and associative learning are not as different as psychology has supposed" (Gold and Stoljar 1999, 864). In a similar way, molecular biology's reductive explanation of the biological concept of the gene has led to its modification without, however, eliminating it. The old concept of the gene has been split into three different concepts corresponding to different criteria of gene identity.[18]

Kemeny and Oppenheim (1956) had already observed that the fact that the reducing theory generally corrects the reduced theory, and the fact that the laws derived from the reducing theory are therefore generally incompatible with the laws of the reduced theory, make it impossible to satisfy the requirement imposed on reduction by Nagel (1951) and Woodger (1952) that there be biconditionals linking the vocabularies of the reduced and reducing theories.[19] According to Kemeny and Oppenheim, "any actual example has to be stretched considerably if it is to exemplify connections by means of biconditionals, and most examples will under no circumstances fall under this pattern" (1956, 13). Their article anticipates the central thesis of Schaffner's theory: "We might suggest that it is some modification T_R^* of T_R that is actually reduced to T_B" (17; symbols modified). However, they refrain from

18 Genes can be thought of as units of recombination: in this sense, a gene is a "recon." They can also be seen as units of mutation: in this sense, a gene is a "muton." But what corresponds most closely to the traditional functional concept of a gene is what enables hereditary traits to be transmitted from one generation to the next: in this sense, a gene is a "cistron" (Kitcher 1982; Rosenberg 1985). In the words of Endicott, "'cistron' is a corrected image of the Mendelian gene (a term in T_R^*, and hence a term supposedly [according to the CHB model, where CHB stands for Churchland, Hooker, Bickle] formulated within the idiom of T_B). Yet it was not created from molecular genetics (T_B) *ex nihilo*, but from the pressure of the original Mendelian theory (T_R) to find a structure with the function of a gene. So *co*-evolved terms within T_B or rather its subset T_R^* are by their very nature dually constrained by the rationales and conceptual resources grounded at both levels. In a word, they are theoretical hybrids." (1998, 65)

19 Wimsatt points out that, insofar as one can reconstruct the reduction of an old theory — now considered false — from a new theory that corrects it, as a deductive argument whose form is valid, "there had better be an equivocation somewhere!" (1976a, 218).

developing this idea because they consider that "such a T_R* is not usually formed, and it may be very difficult to formulate it" (17).[20]

Instead of judging, as Kemeny and Oppenheim do, that correctives are too complex to be subject to formal analysis, and instead of judging, like Popper, Feyerabend, and Kuhn do, that they are not really reductions since they refute the old theory T_R (reduction requires that it be justified by being deducible from a more fundamental theory T_B), several authors — including Putnam (1965), Hempel (1965a), and Schaffner (1967) — have tried to construct a more sophisticated concept of reduction, which makes it possible to account for the fact that the reductions found in the history of science do not preserve the reduced theory but generally correct it.

Schaffner proposes a "general reduction paradigm" (1967, 144) supposed to apply both to reformative reductions that correct the reduced theory and to conservative reductions that preserve it. According to this model, T_B reduces T_R if it is possible to derive from T_B a "corrected" theory T_R* that "bears a close similarity" to T_R and that produces quantitative predictions that are "very close" to those of T_R (144). Schaffner rejects Nagel's (1961) thesis that binding principles of the conditional — rather than biconditional — form are sufficient to accomplish a reduction. His justification for taking up the earlier criterion formulated by Nagel (1951) and Woodger (1952) is that, following Feigl (1958), the most plausible interpretation of linking statements is that they express "synthetic identities" (145) and that an identity statement is a fortiori of biconditional form. He distinguishes between the association and subsequent identification of objects in the domain of theories T_B and T_R

20 The model of reduction that they propose in exchange does not take account of a direct relationship between the *theories*: for Kemeny and Oppenheim, it is impossible to find a link between the reduced and reducing theories that accounts for the reduction. The only condition that it is possible to express concerns the *observable consequences* of these theories. In stronger reduction models, Kemeny and Oppenheim's condition will be considered necessary but not sufficient. A reduction that satisfies only the minimal condition of Kemeny and Oppenheim does not establish any link between the theories themselves; it is therefore inappropriate to speak of an intertheoretical reduction. In a case in which there is no intertheoretical reduction but replacement (or "elimination"), the requirement of Kemeny and Oppenheim gives precise meaning to the idea that the new theory explains the entire domain of phenomena that the old theory explained. In Schaffner's terms, it is a matter of "reduction as explanation of the experimental domain of the replaced theory. Though in this latter case we do not have intertheoretic reduction, we do maintain the 'branch' reduction" (1992, 320) of some science conceived in terms of the domain of phenomena it explains. Also see Schaffner (1993, 423, 431).

and the association and subsequent identification of the properties expressed by their predicates, stating that "it is possible to set up a one-to-one correspondence representing synthetic identity between individuals or groups of individuals of T_B and T_R*" (144; symbols modified).[21]

This is a major change in the conception of reduction. For Nagel, the reduced and reducing theories possess and retain their own domain of individuals and properties, with the linking principles allowing the deductive integration of the laws of T_R into the theoretical framework of T_B. The linking principles merely express the existence of dependencies that form the basis of the reductive inferences that move from one domain to the other. There is a big step to be taken between the hypothesis that the temperature of a macroscopic gas *depends on* the kinetic energy of the microscopic molecules that make it up and the hypothesis of *the identity* of these two domains, even if it is "synthetic" (i.e., known a posteriori). Then it is just one step further to accept the idea that the deduction that corresponds to a Nagelian reduction, between T_B and T_R*, is in fact an *intratheoretical* deduction that belongs entirely to the reducing theory.[22] It is then the link of analogy between T_R* and the old theory T_R that corresponds to the reduction relation.

21 Schaffner adds that the reduction functions that associate individual constants and predicates are "in general . . . interpretable as expressing referential identity" (1967, 144). See also Schaffner (1976, 618). The transition from function to identity is much less straightforward than Schaffner presents it to be. The existence of an association function between the properties described by T_R and the properties described by T_B is compatible with the thesis of the emergence of the former, according to interaction laws. Schaffner sets as a condition for reduction that "T_R* (entities) = function [T_B (entities)]" and that "T_R* (predicates) = function [T_B (predicates)]" (1976, 618). In the case of a multi-realizable property, there is such a regular association function. But this condition is much more general (or weaker) than the "referential identity" condition that Schaffner seems to consider as equivalent. It corresponds to the case in which the function is identity.

22 This step has been taken by Churchland, Hooker, and Bickle. See, for example, Churchland (1985, 11) and Bickle (1998, 108).

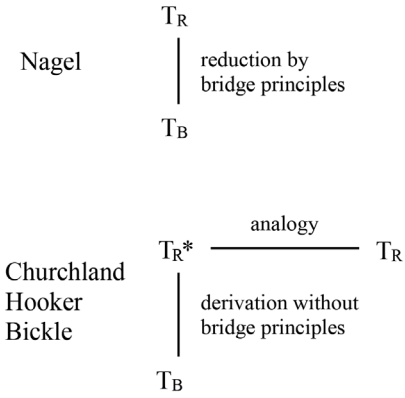


Figure 1.2 Two reduction models, differing with respect to historical change. In Nagel's model, the old theory T_R is deduced from the new theory T_B using linking principles. In the CHB (Churchland-Hooker-Bickle) model, this deduction is carried out without any use of linking principles, using only the conceptual resources of T_B, and leads to a new theory T_R*, analogous to the reduced theory T_R. T_R* is of the same theoretical level as T_R but improves it.

This conception aims to assimilate microreduction, the subject of Nagel's model, to the reduction between successive theories dealing with objects of the same size.[23] In the sciences, the term "reduction" is often used to characterize the relationship between a new theory, such as special relativity, and an older theory that it replaces while dealing with the same objects, in this case classical Newtonian mechanics. In this sense, the term "reduction" refers to the relationship between a new theory, T_R*, and an older theory, T_R, which strictly speaking is false and which T_R* replaces. In general, T_R can be recovered from T_R* by giving certain parameters counterfactual values (e.g., an infinite speed of light in the equations of relativistic mechanics). In this context, it is said that "T_R* reduces to T_R." In this sense, we can say that relativistic mechanics "reduces to" classical Newtonian mechanics "in the limit of small speeds" (i.e., in the limit in which the speeds under consideration are much smaller than the speed *c* of light). For example, in relativistic mechanics, the

23 On this second concept of reduction, see Glymour (1970), Batterman (1995, 2002), and Rueger (2000b, 2001, 2004); on the comparison between the two concepts, see Nickles (1973) and Wimsatt (1976a, 215 ff.).

momentum p of an object is equal to $p = \frac{mv}{\sqrt{1-(v^2/c^2)}}$, where m indicates the mass of the object (defined in a frame in which it is at rest), v its velocity, and c the speed of light. It is commonly said in science that this equation "reduces" to the classical momentum equation, $p = mv$, "in the limit of small speeds": that is, in situations in which $v << c$ (the object's speed v is much smaller than the speed of light c).[24] In this type of case, the reduction links two theories, T_R and T_R^*, which have the same field of application. The theory T_R^* replaces the old T_R because it allows its errors to be corrected while reproducing its successes: T_R^* is empirically stronger or simpler, or both, than T_R. But the two theories deal with the same objects: unlike in microreduction, in which the reducing theory is concerned with parts of the objects that are the subject of the reduced theory, this is an "intralevel" or "domain-preserving" reduction (Nickles 1973, 181).

Schaffner's model of reduction (and its variants developed by Churchland, Hooker, and Bickle) aims to assimilate "intralevel" reduction to microreduction or "domain-combining" reduction (Nickles 1973, 181), where the domains of the reduced and reducing theories are at different levels of the micro-macro scale.[25] This seems to be difficult to conceive since the first reduction model requires that the objects of both theories belong to the same domain (and the same level of the micro-macro scale), whereas the second model is supposed to cover reductions between theories describing objects of different sizes, where the domain of the reducing theory relates to objects that are *parts* of the objects in the domain of the reduced theory. How could "interlevel" microreduction be assimilated into "intralevel" reduction?

---

24 With $\gamma = 1/\sqrt{1-(v/c)^2}$, γm is the relativistic mass (i.e., the mass in a reference frame where the object is at speed v). If $v<<c$, then $\gamma \approx 1$, and $m(v) \approx m$, so that $p \approx mv$.

25 Rueger (2004) attempts to achieve this assimilation by another means. He sees the microdescriptions and macrodescriptions of a system as two descriptions of the same level (i.e., the macrolevel). The microdescription involves the attribution of a microstructural property in terms of a variable on the microscopic scale. Both are causally efficacious, the macroproperty being a "part" of the "micro"property (which in fact is a macroproperty that takes the microstructure into account). The least that we can say is that the thesis that one property can be "part" of another, in the mereological sense, needs to be justified. In the absence of such a justification, it is the *description* of the macroproperty that is "part" (in the sense of the terms of a conjunction) of the description of the microproperty, where the latter is written in the form of a Taylor series development of the solution of the equation that determines the microproperty.

Schaffner's idea — later taken up by Churchland (1979, 1985), Hooker (1981), and Bickle (1998) — is that the reduction of T_R to T_B comprises two distinct stages. The first stage involves the construction, within T_B and under certain conditions C, of a theory T_R^* that replaces the old theory to be reduced T_R. This first step is supposed to bridge the distance between the microscopic theory T_B and a theory of macroscopic objects, *without* resorting to Nagelian linking or "bridge" principles. It therefore respects the stricter conditions imposed by an a priori reduction model: the higher-level theory is deduced solely from the conceptual resources of the reducing theory, without recourse to linking principles or other conceptual resources foreign to the microscopic theory T_B. I propose to call such a conception of reduction "conceptual reduction." It is only at the second stage that the CHB model of reduction (as noted above, I will use this acronym to refer to the model elaborated by Churchland, Hooker, and Bickle) makes use of "linking principles" between terms of the old theory and terms of the new theory. Between T_R and T_R^*, Schaffner says, there must be a "positive analogy" (1967, 144). In his development of Schaffner's conception, Hooker argues that the existence of an "analog relation" between the theory T_R^* derived from the base theory T_B and the theory to be reduced, T_R, "warrants claiming (some kind of) reduction relation, R, between T_R and T_B" (1981, 49).[26]

Let us assume the existence of two stages, the first of which consists of crossing the distance between the microscopic theory T_B and the macroscopic theory T_R^* by means of a deduction that exploits only the resources of T_B. Insofar as T_R^* is conceived as a theory of the same level as T_R that corrects T_R, the relationship between T_R^* and T_R resembles an intralevel reduction in the sense of Nickles (1973) and Wimsatt (1976a). The controversial thesis required to justify this assimilation concerns the first stage of the reduction. The CHB model sees the derivation of T_R^* from T_B as a deduction internal to T_B. According to this model, it is not necessary to use Nagelian "linking principles" to bridge the distance between the microdomain of T_B and the macrodomain of T_R^*. In this sense, the derivation of T_R^* from T_B is therefore

26 In Churchland's words, "a reduction consists in the deduction, within T_B, not of T_R itself, but rather of a roughly equipotent *image* of T_R, an image still expressed in the vocabulary proper to T_B" (1985, 10; symbols modified). Bickle (1998) gives a detailed account of the conception of reduction developed by Schaffner, Hooker, and Churchland.

intratheoretical and interlevel.[27] To take account of the fact that the reduced theory T_R is often only analogous to a particular case of application of T_B, the derivation of T_R^* uses "limiting assumptions" and "boundary conditions" as premises in addition to T_B. These are the types of assumptions regularly used in intralevel reductions. To give an example of a limiting assumption, we can think of the fact that, to find equations structurally analogous (T_R^*) to the equations of classical mechanics (T_R) from relativistic mechanics (T_B), we have to use the assumption that the speeds of the objects to which the equations are supposed to apply are very slow compared with the speed of light. To give an example of boundary conditions, in the derivation of certain equations of statistical mechanics (T_R^*) analogous to the thermodynamic equations (T_R) for a gas, it is assumed that the number of molecules in the gas remains constant and that the gas remains confined to a constant volume.

Before going any further, I will examine in the next section whether the model of conceptual reduction suggested by Schaffner, Churchland, Hooker, and Bickle applies to a paradigmatic case of reduction.

5. The Reduction of Thermodynamics to Classical Mechanics

The controversial condition of the CHB model requires that it be possible to deduce, from T_B alone, the laws of T_R^* without employing linking laws that appeal to T_R concepts. In order to evaluate the thesis that this model adequately represents historical cases of scientific reductions, and to avoid begging the question, I will examine whether the CHB model can account for a paradigmatic case of a successful reduction, that of thermodynamics to classical mechanics. Bickle (1998) argues that the CHB model passes this test: he tries to show that it is able, in particular, to account for the reduction of the ideal gas law to classical mechanics. According to Bickle, it is possible to derive, within classical mechanics, the following "analog structure" of the ideal gas law:

27 The crucial thesis is that no "linking principle" or "bridge law" is needed to derive the "image" T_R^* from T_B (the reducing theory), an image whose isomorphism with T_R (the reduced theory) justifies the claim that T_B reduces T_R. As Churchland puts it, "the correspondence rules play no part whatever in the deduction. They show up only later, and not necessarily as material-mode statements, but as mere ordered pairs: <Ax, Jx>, <Bx, Kx>, . . . " (1985, 10).

$$(*)\quad \frac{Nm\overline{v^2}}{3lwh}\cdot lwh=\frac{2N}{3}\cdot\frac{1}{2}m\overline{v^2},$$

where N denotes the number of molecules in a sample of an "ideal" gas, m the mass of a molecule, v (the absolute value of) the velocity of a molecule, $\overline{v^2}$ the average square velocities of the molecules taken over all of the molecules and over time, and l, w, and h the length, width, and height of the container enclosing the gas.

The question is whether, as Bickle maintains, the derivation of an equation formally analogous to the ideal gas law can be obtained within classical mechanics alone or whether this deduction requires "linking" principles that describe the dependence of certain systemic properties of a macroscopic object (such as $\overline{v^2}$ [the average of the squares of the velocities of all molecules]) on the microscopic properties of its components (e.g., the velocities v of the individual molecules).

The evaluation of the thesis that the CHB model is adequate for the analysis of this particular case of reduction is of some importance. It is a paradigmatic case because of its relative simplicity compared with reductions of chemistry to physics, or of psychology to neurophysiology, which I will consider later. The relative simplicity of its mathematical derivation justifies my consideration of this case as a touchstone: if the conditions imposed by a given model of reduction are too strong for this case to be considered a successful reduction, then the model is inadequate. It is plausible to assume that more complex reductions satisfy these conditions even less.[28]

In an elementary presentation of this reduction, consider an ideal gas of N molecules, each with a mass m, contained in a volume V of a box whose sides define the directions x, y, and z of the Cartesian reference frame. The number of molecules per unit volume is $\rho = N/V$. The fundamental idea behind the reduction is that the pressure of the gas on the walls of the box results from the force exerted by all of the impacts of the molecules on the wall. The aim is to calculate the number of molecules that strike a given surface A during a given period of time Δt and then to multiply it by the force exerted by each of

28 In Krüger's words, it can be assumed that the reduction of thermodynamics to classical mechanics "will mark something like an upper bound to the strength or the completeness one is likely to achieve in reduction in general" (1989, 373).

these impacts on the wall. As a first approximation, we simply assume "that each molecule moves with the same speed, equal to its average speed $\bar{v}$" (Reif 1967, 40).

This simplification contains an innocent part that is relatively easy to abandon as well as a substantial part: the average taken over the speeds of all molecules at a given instant also represents the average over a long time given the dynamics of the molecules. This average only corresponds to a real property of the system when it is in equilibrium. Equilibrium is characterized by the fact that, despite the inevitable changes in the state of motion of the particles, the overall distribution of the speeds of all the molecules is approximately constant (although it undergoes fluctuations around this constant distribution). Once this assumption has been made, the innocent part of the simplification consists of calculating the pressure, not on the basis of the Maxwell-Boltzmann distribution, which indicates the proportion of molecules with a given speed, but on the basis of the overall average speed, taken over all of the molecules. Insofar as this average over the molecules necessarily exists (the number of molecules being finite), this will not change the result of the calculation.[29]

One gets the number of molecules striking the surface A over a period of time Δt by assuming that 1/6 of the molecules move approximately in the direction toward A, which corresponds to the x axis. One in three molecules has a velocity almost parallel to the x axis, and one in two of those molecules is moving toward A, while the other is moving away from A. Next one notes that all of the molecules that hit A in an interval Δt must be located in a cylinder (a fictitious construction by the theorist) with face A and length $\bar{v}\Delta t$. $\bar{v}\Delta t$ is the distance travelled by a molecule with speed $\bar{v}$ during the time interval Δt. The molecules contained in this cylinder, and only those molecules, will reach A during the interval Δt. Based on the assumption that the density of molecules is everywhere $\rho = N/V$, the number of molecules that hit A during the interval Δt is therefore $\frac{1}{6}\varrho\bar{v}A\Delta t$. To calculate the force exerted on A during a collision of one molecule with A, one again uses the assumption that the

29 Richet (2001, 315–16) shows this by calculating the pressure, not simply by multiplying the number of molecules per volume by their average velocity, but by taking the integral of the product of the velocity and the number of molecules with that velocity: $v_x n(v_x)$, where v_x denotes a particular velocity and $n(v_x)$ the number of molecules with that velocity, over all possible velocities, from zero to infinity, using the Maxwell-Boltzmann distribution for the function $n(v_x)$.

gas is in equilibrium. It is assumed that the kinetic energy of such a molecule must remain unchanged during the shock. "This must be true, at least on the average, since the gas is in equilibrium. . . . The *magnitude* of the momentum of the molecule must then, on the average, also remain unchanged" (Reif 1967, 41). Since the force exerted by the wall is equal to the force exerted by the molecule (Newton's third law), and the force exerted by the molecule is equal to its change of the momentum (Newton's second law), it is sufficient to calculate the change of the momentum of a molecule travelling in a perpendicular direction to the wall (the x direction), given that the modulus of this momentum remains unchanged: this is the case for a molecule that rebounds after an elastic shock, so that the change in (the x component of) its momentum is $-2m\bar{v}$. For the average force (this is the average over both molecules and time) exerted by the molecules on the wall in the x direction, one obtains

$$\bar{F} \approx (2m\bar{v})\left(\frac{1}{6}\rho\bar{v}A\right) = \frac{1}{3}\rho m\bar{v}^2 A$$

The average pressure is the average force exerted by the molecules on the wall in the x direction, per unit area of the wall:

$$(1)\quad \bar{p} \approx (2m\bar{v})\left(\tfrac{1}{6}\rho\bar{v}\right) = \tfrac{1}{3}\rho m\bar{v}^2$$

The crucial question is this: does this calculation allow us to deduce, from the concepts and principles of classical mechanics alone, which is the reductive microtheory describing the behaviour of microscopic particles, an expression that describes a macroproperty of the macrosystem of which these molecules are the components? This is the fundamental condition of the CHB model, which applies, according to its proponents, to the case of the reduction of thermodynamics: no linking principle is necessary. The first part of the reduction consists of deducing, using only the concepts and principles of the reducing theory, an analogous image of the reduced theory. This analogous image is supposed to have two properties: it is entirely formulated in the language of T_B (mechanics), and it nevertheless describes global (or systemic) properties in the domain of the reduced theory T_R^* (thermodynamics). The second part of this statement is true but not the first part. It is true that v is a term belonging to the vocabulary of the reduced theory. It is reasonable to argue that this is also the case for the average speed of molecules at a given instant. In favour of this hypothesis, it can be argued that this average speed

is always the speed of a microscopic object, albeit a virtual object: there is as much reason to say this as there is to say that the centre of gravity of a system of massive bodies, albeit virtual, is a mechanical object, at the same level as the massive bodies that make up the system.

Now the velocity $\bar{v}$ in (1) is an average obtained by neglecting the fluctuations of individual molecules. It is only on this condition that it is possible to conceive of the expression on the right in (1) as designating a macroscopic property of the gas. One can only reduce the temperature of the gas to the kinetic energy of the molecules if one defines this kinetic energy on the basis of such an average.

$e_{kin} = \frac{1}{2}mv^2$ (where e_{kin} denotes the kinetic energy of a molecule, v its velocity, and m its mass)

is a microscopic property that can be attributed to the molecular components of the gas. But the quantity that forms the basis for reducing temperature as a macroscopic quantity is

$$\overline{e_{\text{kin}}} = E_{kin} = \frac{1}{2}m\overline{v^2}$$

To be able to consider that an expression refers to a macroscopic property of the gas, it is necessary to assume that the averages over time correspond to real properties; in short, it is necessary to assume that the system is in equilibrium. "The temperature and pressure of a gas have only a statistical significance. An isolated atom has no temperature and pressure. . . . Temperature and pressure can be defined only when the number of atoms is large enough that their values are time independent" (Richet 2001, 316).

The conceptual transition from the microscopic mechanical domain to the macroscopic thermodynamic domain therefore corresponds to the transition from a system made up of a set of molecules whose average speed and kinetic energy over the macroscopic sample and over time can be calculated, to a single object with the stable global properties p and T. In order to describe and explain the behaviour of a macrosystem in thermodynamic terms, it must be assumed that it actually possesses these properties. For the averages of microscopic quantities over time to correspond to real properties of the system, the system must be in equilibrium. If this is not the case, the average $\overline{e_{kin}}$ over time does not give rise to a temperature, and the average $\bar{v}$ over

time does not give rise to a pressure: the system does not possess these latter properties, and it does not obey the laws that define them, such as the "zeroth law of thermodynamics," which defines temperature in terms of equilibrium.

One might suspect that the need to resort to the equilibrium hypothesis is merely an artifact of the simplified presentation of reduction in textbooks. This is not the case. In the presentation of the reduction of thermodynamics by Gibbs (1902) — which still constitutes the most important theoretical model — the conceptual leap between the consideration of a system as composed of a set of microcomponents and the consideration of a thermodynamic macrosystem is clearly apparent. It is impossible to know the exact state of the molecules. Even at equilibrium, the real values of the mean pressure at a given moment, and of the average kinetic energy $\overline{e_{kin}}$, taken over all of the molecules at a given moment, are not strictly identical to their mean values taken over time but fluctuate around this mean value. Yet the macroscopic properties T and p by definition are properties that characterize systems at equilibrium: by definition, they are independent of time. In Gibbs's approach, the values of (macroscopic) thermodynamic quantities are calculated from the fiction of the "set" of all possible microsystems given the macroscopic constraints imposed. This construction ensures that the values of these quantities remain stable over time. This means that thermodynamic properties, such as temperature, are not calculated directly from the state of the particular underlying microscopic system. They are calculated from the fiction of all the systems obeying the same constraints. Insofar as the system is in equilibrium, it really does have a macroscopic temperature and pressure. These properties are not fictitious. They are real properties whose conception is irreducible to the framework of microscopic mechanics alone. The conceptual transition corresponding to Gibbs's derivation of these quantities from the description of the system in terms of the microproperties of its components cannot be accomplished with the conceptual resources of the reducing theory alone. The use of the Gibbsian concept of an "ensemble" of systems (or some other concept that cannot be reduced to the conceptual apparatus of microscopic mechanics) is indispensable.[30]

30 Nagel is aware of the indispensable nature of statistical premises, themselves irreducible to mechanics, in the reduction of thermodynamics: "It is one thing to say that thermodynamics is reducible to mechanics when the latter counts among its recognized postulates assumptions (including statistical ones) about molecules and their modes of action; it is quite a different thing to

Regarding pressure, Sklar expresses this point as follows:

> There is, for a particular sample of gas at equilibrium, the actual momentum transferred by the molecules impinging on a wall of the box in a short time, and there is its average value per unit area per unit of that time. On the other hand, there is the quantity calculated for an ensemble of similarly constituted systems. . . . Whereas the former sort of pressure, the feature of the individual system, will be expected to fluctuate, the latter kinds of ensemble quantities, quantities defined by the macroscopic characterization of the system and the chosen probability distribution over the ensemble, will, of course, not. Here, fluctuation will show up as assimiliated into the ensemble description by the calculation of averages or most probable values of quantities, but the averages themselves are not the sort of things to fluctuate. (1993, 349–50)

The "orthodox" procedure for reducing thermodynamic quantities to mechanics involves Gibbs's concept of ensemble. Of course, the fact that the derivation of thermodynamic concepts via the ensemble concept is not purely mechanical does not mean that there is no other way of deriving them that does not require recourse to concepts that do not belong to mechanics.[31] However, other attempts to reduce thermodynamics have encountered the same difficulty. For example, it has been suggested that the concept of probability should be avoided when moving from the description of a system in microscopic terms to its description in macroscopic terms, by deducing macroscopic laws from mechanical laws and initial conditions; it is plausible to think that the non-existence of macroscopic systems that break macroscopic laws while obeying microscopic mechanical laws (e.g., an isolated system whose entropy decreases "spontaneously" without being compensated by an increase in entropy in another system) can be explained by the fact that

---

claim that thermodynamics is reducible to a science of mechanics that does not countenance such assumptions" (1961, 362). According to my analysis, the first claim is justified but not the second claim. However, the reduction would only satisfy CHB's conditions if the second statement were true.

31 Krüger (1989) briefly presents three other approaches. None of them uses only mechanical conceptual resources.

this would require exceptional initial conditions. Clearly, we cannot, in practical terms, specify such initial conditions since we can neither observe nor describe the state of motion of 10^{23} particles at a given moment.[32] But let us put aside this difficulty; it might be only an epistemic one that does not affect the ontological aspect of the question. Now, even if we ignore the problem of the number of factors that form part of the initial conditions, there is no reason to think that the initial conditions that characterize systems obeying macroscopic laws can be rendered in microscopic terms. On the contrary, it seems to be plausible that the only commonality of these conditions is the *macroscopic* property of characterizing systems that obey macroscopic laws.[33] The specification of initial conditions, if it were practicable, is therefore possible only by using macroscopic concepts and vocabulary. Consequently, a reduction of thermodynamics along this path would not be able to deduce it from the conceptual resources of microscopic mechanics alone.

It can be concluded provisionally that in the present state of science the CHB model does not apply to the reduction of thermodynamics to mechanics. Its principles cannot be derived from mechanical resources alone. As Krüger concludes, "notions like equilibrium and temperature (and thereby entropy) must be given physical meaning on a basis of more than just mechanics" (1989, 382).[34]

---

32 There are approximately 10^{23} atoms or molecules in 1g of matter. More precisely, the "Avogadro number," defined as the number of carbon atoms in 12g of the 12 of carbon, is approximately equal to $6 \cdot 10^{23}$.

33 "There is, as far as I am aware, no indication that there is a non-trivial or non-question-begging property in the language of mechanics that would be common to all microstates which behave normally, but absent from all those which do not" (Krüger 1989, 379).

34 This is independent of the interpretation of the concept of ensemble. Gibbs himself interprets it as an expression of our ignorance of the detailed microscopic state: "The laws of thermodynamics . . . express the laws of mechanics for these systems, as they appear to beings who do not possess sufficient fineness of perception to be able to appreciate quantities of the order of magnitude of those belonging to the individual particles, and who cannot repeat their experiments often enough to obtain other than the most probable results" (1902, vii–viii; cited by Krüger 1989, 380). For Gibbs, we have to construct ensembles because we do not have access to the individual properties of all the microscopic particles that make up a macroscopic system. Einstein, who developed the ensemble approach independently of Gibbs, interprets ensembles in a different way. For Einstein, the ensemble describes the distribution of energies in a collection of systems in contact with a "heat bath" (i.e., an infinitely large reservoir of heat that has a fixed temperature) (1902, para. 5; 1903, paras. 3–4; Krüger 1989, 382). In Einstein's interpretation, a fictitious infinite ensemble serves as the basis for assigning thermodynamic quantities to a real individual system. Irrespective of the

One of the attractions of the CHB model is its promise to account for reduction without having to postulate linking principles that give rise to the suspicion of making the reduction obscure. As long as these binding principles themselves are not reduced (by derivation from T_B), the higher-level theory is only incompletely reduced but remains partly mysterious. In Bickle's words, "one advantage of the H-C [Hooker-Churchland] account is that it avoids having to specify the logical status of cross-theoretic identity statements" (1992, 223). It avoids this problem because, "if the deductive part of a reduction has no gap to bridge between the language or the ontology of premise and conclusion, then the nonexistence of lawlike connections between reduced and reducing concepts or kinds is of no consequence" (Bickle 1998, 108). We have seen that the CHB model, far from circumventing the problem, puts forward a hypothesis to solve it: it is the hypothesis that the relevant "identity statements" can be derived within T_B (or *approximately* derived, insofar as T_R*, an approximation of T_R, is strictly derived). However, I have shown that this assumption is false in the case of the reduction of thermodynamics.

Bickle's thesis that the CHB model of reduction does not need linking hypotheses is crucial in his defence of reductionism against various anti-reductionist arguments. It is important to show that Bickle's failure to refute these arguments does not refute reductionism. The "synthetic model of reduction," which I will introduce in the next section, makes it possible to answer them without the thesis (which, as we have seen, is mistaken) according to which T_R* can be derived from T_B without linking statements.

1. Davidson (1970) justifies the autonomy of psychology by the absence of strict psychophysical laws. This argument presupposes that the reduction of psychology to neurophysiology requires the discovery of psychophysical laws that can play the role of the linking principles in Nagel's model. However, Bickle claims that "the impossibility of psychophysical laws is irrelevant to the new thesis of mind-brain reductionism and the novel account of intertheoretic reduction underwriting it" "since an H-C [Hooker-Churchland] reduction nowhere requires bridge laws" (1992, 218, 224). This defence of reductionism against Davidson's argument invites two objections.

First, the absence of linking principles in the CHB model does not stand up to the test of the analysis of a paradigmatic case of reduction. This analysis

interpretation chosen, the need to make use of a fictitious infinite ensemble shows the inadequacy of strictly mechanical concepts for deriving thermodynamic properties and their laws.

shows that at least some reductions require linking principles. The analysis of the reduction of long-term memory will show (in section 7) that it involves laws of composition that play the role of linking principles. Davidson's thesis, according to which the difference between the conditions of attribution of physical and psychological states precludes the existence of psychophysical laws, cannot therefore be true in general. It is still possible that this thesis is true of a (very important) part of our psychological states: it is possible that there are no linking principles that directly connect *intentional* states, such as propositional attitudes of believing and desiring, to states of the brain. However, such intentional states may be related nomologically to other mental states that *are* reducible to brain states.

Second, as Endicott has pointed out, the CHB model itself contains linking principles between T_R* and T_R: in the case of relatively "retentive" reductions, where the corrections that T_R* makes to T_R are modest, the reduction justifies identities between objects and properties described by the theories T_R* and T_R (see Churchland 1979, 83; 1985, 11). Furthermore, "property identity guarantees nomic coextension. So bridge laws exist within the new-wave account" (Endicott 1998, 68): that is, in the CHB model of reduction.

2. Bickle sees scientific theories as sets of models rather than sets of statements.[35] In logic, a "model" of a statement (or set of statements) is an interpretation in which the statement (or set of statements) is true. An "interpretation," in the logical sense of the term, associates an object with each singular expression and a set of objects with each predicate. The "semantic conception" of scientific theories consists of conceiving of theories not as sets of statements but as sets of models: the structured sets of objects that make the theory true. According to Bickle, adopting the semantic conception makes it possible to analyze the relationship between reduced and reductive theory without resorting to linking principles. Although this assertion is undeniably true in a literal sense, it seems to be rather superficial. Indeed, in the case of "homogeneous ORLs" (Bickle 1998, 78), where ORL stands for ontological reductive link, the basic sets of the models of T_R constitute a subset of the basic sets of the models of T_B. For example, the point masses of classical collision mechanics (T_R) are identical to the point masses of Newtonian

35 In other words, Bickle adopts the "semantic conception" of scientific theories, an important alternative to the traditional "syntactic conception," according to which theories are sets of statements.

particle mechanics (T_B).[36] Even if, at a fundamental level, theories are sets of models and not sets of statements, they *imply* statements. The assertion (part of the requirements of the CHB model) that the entities designated by the statements of T_R are identical to the entities designated by the statements of T_B can be presented as a consequence of inclusive relations of the basic sets of the respective models of T_R and T_B; this does not prevent the fact that, as soon as such an identity is expressed in a statement, it is a linking statement. In other words, the CHB model still contains statements of intertheoretical connection, even if they occupy the place of consequences and not that of fundamental postulates.[37]

The attempt to identify interlevel reduction with corrective intralevel reduction therefore fails, at least in this paradigmatic case, for the reason that the deduction of $T_R{}^*$ from T_B cannot be conceived of as intralevel within T_B. Another observation diminishes the plausibility of this assimilation. The corrective modification of the reduced theory, although it is a regular effect of reduction, is not always the unique aim of interlevel reduction.[38] Another important aim is explanatory unification: a reduction aims to show what macroscopic properties such as temperature *consist of*. It is perfectly possible for this goal to be achieved by an interlevel reduction that justifies the high-level theory as it is, whereas the only reason for an intralevel reduction is to improve the reduced theory. Interlevel reduction provides information about the detailed nature of the reduced properties, without necessarily showing that the reduced theory is wrong: reduction justifies the notion of temperature as

36 See Balzer, Moulines, and Sneed (1987, 255–67); Bickle (1998, 78).

37 Schaffner had already shown that the "semantic" conception of reduction first proposed by Suppes (1967), conceiving of it as a relation of isomorphism between models of theories rather than between their statements, "is a special case of a Nagel type of reduction" (1993, 430; Schaffner 1967, 145). The isomorphism of models is not sufficient for a reduction because theories can relate to domains that have isomorphic models without being connected. Schaffner (1967, 145) mentions heat theory and hydrodynamics, which share their formal structures, without one being reducible to the other. For this reason, Bickle adds the condition of the existence of ORLs. With this condition, the semantic conception of reduction becomes equivalent to the syntactic conception in terms of linking statements. Furthermore, Endicott (1998, 71) points out that the only innovative element of the CHB model compared with Schaffner's model — namely, the requirement that $T_R{}^*$ be inferred from the conceptual resources of T_B alone, cannot be expressed in the semantic conception, because the notion of inference in accordance with rules makes sense only in the context of a conception in which theories are expressed in statements.

38 "Unlike the evolutionary intralevel cases, the reduced theory in interlevel situations does not stand in need of technical correction in every case" (McCauley 1996, 30).

a stable (time-independent) property of macroscopic systems at equilibrium, but it also provides the information that this stable property emerges against a background of microproperties subject to fluctuations and themselves therefore stable only on average.

The lesson to be learned from the analysis of the reduction of thermodynamics is that macroscopic thermodynamic properties cannot be *identified* with microscopic properties of their components. It is not enough to assume that gases are composed entirely of molecules to be able to reconstruct the macroscopic properties of gases from the microscopic properties of their components. "A too naive application of the notion of identificatory reduction would be misleading, because 'temperature' in many of its uses in statistical mechanics refers not to an instanced property of a particular system sample at a time, but, rather, to some feature of a probability distribution over systems of a specified type" (Sklar 1993, 352).[39]

6. The Synthetic Model of Reduction

We have seen that the paradigmatic reduction of thermodynamics to classical mechanics is possible only on the basis of concepts and principles that go beyond those available in the reducing theory. In what follows, I propose a "synthetic" model that takes this into account. This model, sketched in Figure 1.3, retains elements of each of Nagel's model and the CHB model.

39 In Chapter 4, we will see that we can characterize the relationship between a macroproperty such as temperature and the microproperty that allows its reduction more adequately using the concept of emergence. The properties of a macroscopic object composed of microscopic parts are determined by laws of composition. When the law of composition is non-linear, macroproperties can emerge that are qualitatively different from the microproperties that determine them.

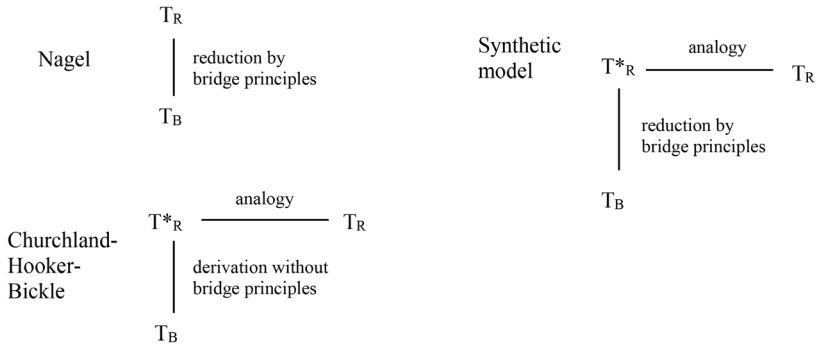


Figure 1.3 Three reduction models, differing with respect to bridge principles and historical change.In Nagel's model, the old theory T_R is deduced from the new theory T_B by means of linking principles. In the CHB (Churchland-Hooker-Bickle) model, this deduction is carried out without any use of linking principles, using only the conceptual resources of T_B, and it leads to a new theory T_R*, analogous to the reduced theory T_R. T_R* is at the same theoretical level as T_R but is superior to it. In the synthetic model (inspired by Schaffner), the reduction of T_R involves deducing a theory T_R* that corrects T_B, starting from T_B *using linking principles*.

The synthetic model retains Nagel's thesis that the linking principles (or bridge laws) cannot be deduced without prior knowledge of the phenomena and laws of the level of the reduced theory (T_R* and T_R). The derivation of T_R* from T_B presupposes the a posteriori discovery of interaction laws as well as concepts and laws specific to the transition between levels. My model also retains the observation of the CHB model that the result of such a reductive deduction does not in general coincide exactly with the old theory T_R but that it yields corrections. In short, the synthetic model takes account of the fact that reductions generally lead to a correction of the reduced theories without requiring the reduced theory to be absorbed by the reducing theory. In general, a reduction consists of the discovery of a nomological explanation of structures and laws governing the evolution of these structures, which uses conceptual resources from the reducing theory T_B and the reduced theory T_R. Indeed, as we saw in the previous section, the reduction of thermodynamic laws uses, in addition to the principles of the reducing theory, statistical principles foreign to microscopic theory, in particular Gibbs's notion of an ensemble.

Schaffner's model does not explicitly require the derivation of T_R* to use only the conceptual resources of T_B. In this respect, it resembles the synthetic

model that provides the possibility of deriving T_R* using conceptual resources specific to T_R. Schaffner takes up the Nagelian distinction between "homogeneous" and "heterogeneous" reductions. In the case of the former, "all the primitive terms . . . appearing in the *corrected* secondary theory T_R* appear in the primary theory" (1967, 144; variables modified), whereas in the case of the latter the primitive terms of T_R* are associated only with the terms of T_B, where these associations are subject to a certain number of conditions. But these conditions do not limit the conceptual resources that can be used in the construction of T_R* (i.e., resources from T_R as well as those from T_B).[40] Schaffner's GRR model ("general model of reduction-replacement") requires only the existence of a function that associates the predicates of T_B (possibly corrected to T_B*[41]) with the predicates of T_R*. The last condition can be satisfied if the properties of T_R* are determined, as a function of the properties of T_B*, by non-causal laws[42] of composition, according to my synthetic model.

This is an important difference, and Schaffner's model partially escapes my criticism of the CHB thesis, according to which the resources of T_B alone are sufficient to derive T_R*. Nevertheless, the conditions that Schaffner imposes on reduction are too strong: the paradigmatic case of the reduction of thermodynamics does not satisfy them because Schaffner imposes that, even in heterogeneous reductions, the terms of T_R* must have the same reference as the expressions constructed in T_B associated with them during the reduction. The association of T_R* terms referring to individuals, with expressions formed in T_B, takes the form of a "one-to-one correspondence representing synthetic identity between individuals . . . in T_B and T_R*," and "the primitive predicates of T_R* . . . are . . . associated with an open statement in T_B," so that the reduction function that accomplishes this association is "in general . . . interpretable as expressing referential identity" (Schaffner 1967, 144; variables

40 Endicott correctly observes that "Schaffner does not *require* that T_R* . . . be constructed out of the [conceptual] resources of a higher-level T_R" (1998, 58n14). But Schaffner's model does not require T_R* either to be constructed merely from the conceptual resources of T_B.

41 The GRR model takes account of the fact that fruitful reductions often also lead to a modification of the reducing theory. The reduction consists of deducing a corrected theory T_R* from a corrected reducing theory T_B*. See Schaffner (1993, 427–29).

42 The assimilation of the relation of simultaneous non-causal determination to non-simultaneous causality is a major source of confusion in the philosophy of reduction and mental causation. I will explain in Chapter 5 that this confusion plays a key role in Kim's argument against the causal efficacy of macroscopic properties.

modified).[43] However, this thesis is questionable at least with regard to predicates. We saw in the previous section that, in the case of the reduction of the thermodynamic properties of pressure *p* and temperature *T*, there is no expression in the vocabulary of T_B that shares their reference.

7. The Reduction of Cognitive Phenomena by Neurophysiology: Elimination or Co-Evolution?

According to my working hypothesis, reduction is one of the forms of unification of scientific knowledge. Therefore, it will be instructive to compare my analysis of the reduction of thermodynamic quantities in the physical sciences with a case of reduction in the cognitive sciences.

Recent research in neuroscience has uncovered the neural bases of several basic forms of learning: sensitization, habituation, and classical conditioning. Sensitization is a form of learning in which an animal learns to react differently to a stimulus: repetition of the stimulus leads to reinforcement of the behavioural response. Habituation, conversely, reduces the strength of the animal's reaction to a stimulus repeatedly presented. These forms of learning are non-associative in that they involve a single stimulus and a single response. In contrast, classical conditioning (CC) is a form of associative learning in which an animal learns to react with a behavioural response R to a stimulus (CS for *conditioned stimulus*) that is neutral before learning (i.e., that does not provoke any behavioural response), by associating this CS with another stimulus, known as the *unconditioned stimulus* (US), which provoked response R before the learning session. It can be innate that the US triggers R.

The reduction of these elementary forms of learning is well advanced in science.[44] The results presented by Hawkins and Kandel (1984) enable us to understand the neurophysiological mechanisms by which conditioning and

43 True, the requirement that the connection between the predicates of the theories T_R* and T_B (or T_B*) must be established by a hypothesis of "synthetic identity" does not appear in the formal conditions of the GRR model (Schaffner 1993, 429). Nevertheless, Schaffner keeps maintaining that "reduction functions" establish that entities designated by predicates of T_R* are identical to entities designated by predicates of T_B*. "Reduction functions link entities and predicates of reduced and reducing extensionally via an imputed relation of synthetic identity" (Schaffner 1993, 440).

44 There are other cases of detailed neurobiological reductions of cognitive capacities, such as the phenomenon of colour opposition, the basis of which has been discovered in neurons known as *colour opponent cells*. See Hardin (1988); Zeki (1993); Gold and Stoljar (1999). I will examine this reduction in Chapter 4.

learning occur: in a situation in which the animal is confronted with the CS, it reacts with the learned behaviour R.

The importance of reducing these forms of learning would be even greater if it turned out that sensitization, habituation, and classical conditioning form a "learning alphabet" whose combination makes it possible to explain more complex forms of learning. Based on the discovery that the microscopic processes underlying classical conditioning (a type of associative learning) are variants of the processes underlying sensitization (a type of non-associative learning), Kandel makes the hypothesis that "more complex forms of learning can be built up from the molecular components of simpler forms. By this means a variety of distinct forms of behavioral modifications could be achieved by a small set of molecular mechanisms" (1995, 680; see also Hawkins and Kandel 1984). This perspective allows us to answer the objection that the reduction of elementary learning is of very limited significance for psychology, insofar as this reduction has been achieved only in the case of a rather primitive creature, the species *Aplysia californica*: a naked-bodied gastropod mollusc of the genus *Aplysia* (some of whose species are also known as "sea hare") whose cognitive capacities are viewed as rudimentary. A great deal of research on the physiological basis of learning has been carried out using the species *Aplysia californica*, which lends itself to this type of study because of the simplicity of its neural system and the large size of its neurons. However, the primitive forms of learning studied in *Aplysia* are shared by much more complex animals and even by humans. We therefore accept the widespread view that Hawkins and Kandel's discovery is a paradigmatic case for cognitive neuroscience, which aims to discover the neurophysiology underlying cognitive abilities.[45]

Let us start with a form of sensitization that has been studied at the cellular level in *Aplysia*.[46] *Aplysia* has an innate reflex that consists of retracting the siphon inside the parapod and the gill in the mantle (R) as a consequence of a threatening stimulus (US: a tap on the mantle or siphon).[47] This gill-withdrawal reflex is present before sensitization. Sensitization is a process that

45 See Churchland (1986, 369); Bickle (1998); Gold and Stoljar (1999).

46 See Hawkins and Kandel (1984, 377–78); Hall (1992, 474 ff.); Kandel (1995, 671–76).

47 A stimulus is called unconditional (US) if it triggers a behavioural response without any prior learning, as happens with innate reflexes. A stimulus is said to be conditional for a given response R if the CS triggers R only after learning by classical conditioning, whereby the animal learns to associate the CS with a US and therefore with the response R triggered by the US.

leads to the reinforcement of this reflex. It is triggered by a noxious stimulus ($US_2 = N$) to another part of the body, in this case a shock to the tail or head. In psychological terms, the state provoked by *N* can be interpreted as a state of alert that reinforces all defensive behaviours (Hawkins and Kandel 1984, 381).

Here is a simplified explanation of sensitization. The neurons that transmit information about *N* make synapses on facilitator interneurons. These interneurons make synapses on the axon ending of the sensory neuron that transmits information about the US, precisely at the point where this axon has a synapse with the motor neuron responsible for triggering R. The mechanism of sensitization is presynaptic facilitation: the US-R synaptic connection is modified by the interneuron so that R responds more strongly to the US. This is achieved by a modification of the dispositions of the molecular parts of this synapse.[48] Sensitization is reduced in two stages. In the first stage, the cognitive process is reduced to a neurophysiological process of synapse modification. In the second stage, the central stage of the neuronal mechanism is in turn reduced to molecular processes.

Several molecular changes occur in parallel. Stimulation of the interneuron (stimulated by *N*) causes, via a biochemical mechanism, the closure of a number of K^+ channels in the US-R presynaptic axonal ending. When an action potential triggered by a US arrives at the axonal ending, open K^+ channels tend to bring the potential difference across the axonal membrane back to its equilibrium value, whereas closed K^+ channels increase the depolarization that determines the strength of the action potential. Stimulation by *N* thus leads to "changing the conformation of the channel and decreasing the K^+ current," which "prolongs the action potential, increases the influx of Ca^{2+}, and thus augments transmitter release" (Kandel 1995, 673) into the US-R intersynaptic channel.

The reductive explanation of classical conditioning also proceeds by highlighting microscopic changes that result in a modified state of the neurons involved, which can be characterized alternatively in categorical terms or in dispositional terms. Classical conditioning leads to establishing or reinforcing an animal's disposition to react by R to the perception of a CS to which it reacted little or not at all before learning. For this conditioning to take place,

48 In Chapter 3, we will see that cognitive properties can be described both categorically, by describing them in themselves, in abstraction from their role or function, and dispositionally, by identifying them through their causes and effects.

the sensory neurons originating from the CS must be stimulated in a precise temporal interval that precedes the stimulation of the same axonal bulb by the US.

The first stimulation by the CS puts the presynaptic axonal bulb in a state of greater receptivity to successive stimulation by the unconditional stimulus. The molecular basis of this greater receptivity consists of a change in the conformation of adenyl cyclase, a molecule involved in the mechanism leading from stimulation by an action potential to release of the transmitter into the intersynaptic cleft.[49]

Each reduction of a cognitive capacity by the discovery of an underlying microscopic mechanism relies on laws of interaction between the microscopic parts of the system. The overall determination resulting from these laws can be considered to be based on a single "law of composition" (Broad 1925, 63) specific to this type of system. This law determines, in a non-causal way, that the complex system possesses the overall property *M* because its parts (the various ion channels, transcription activators, etc.) possess certain properties: the property of a fraction of the K^+ channels to be closed (P_1), the property of the transcription activators to be phosphorylated (P_2), and so on. The reduction shows how the global cognitive property *M* nomologically depends[50] on microscopic properties $P_1 \ldots P_n$ of parts of the system. In what follows, I will sketch another example of reductionist explanation in cognitive neuroscience before using these examples to evaluate the synthetic model of reduction introduced above.

The discovery of the mechanism underlying the acquisition of long-term memory is another example of a well-advanced psychophysiological reduction. Experimental research in cognitive science has shown that the transformation of short-term memory into long-term memory requires (most of the time) repetition of the stimulus and can be prevented by "retrograde interference": that is "distractions introduced after the initial items have been learned and stored in short-term memory" (Bickle 2003, 47).

---

49 The conformational change of the molecule "enhances its ability to synthesize cAMP in response to serotonin released in the US pathway" (Kandel 1995, 679). cAMP is short for cyclic adenosine monophosphate, one of the molecules involved in the mechanism of long-term memory consolidation. See Kandel (2000, 1254).

50 It is equivalent to say that *M* depends nomologically on $P_1 \ldots P_n$ and to say that $P_1 \ldots P_n$ determine nomologically *M*.

Several molecular mechanisms underlie the development of long-term memory. The first mechanism corresponding to early long-term potentiation (E-LTP) gives rise to a synaptic modification that persists for approximately one to three hours (see Bickle 2003, 63–67). Through a cascade of biochemical interactions involving numerous molecules, the strong depolarization of the post-synaptic membrane leads — through the reception of molecules of the neurotransmitter glutamate from the synaptic cleft and after numerous intermediate steps involving other channel proteins in the membrane as well as cAMP molecules (secondary messengers) — to a change in the conformation of two species of receptor molecules, called AMPA and NMDA. The consequence of this conformational change is that the channels with which these molecules are associated remain in an "open" state. In this state, the conductivity of the AMPA receptor for Na^+ ions, for example, is almost tripled. This means that, if a new stimulation reaches the post-synaptic neuron while it is in the long-term potentiated state, it will produce an enhanced post-synaptic excitatory potential (EPSP), increasing the likelihood that the overall depolarization at the axon neck of the post-synaptic cell will be strong enough for it to send out an action potential in response to stimulation by the pre-synaptic cell.

The second phase of long-term potentiation (L-LTP) is triggered by repeated stimulation of the post-synaptic cell. It leads to a much longer-lasting change in the structure of the post-synaptic neuron. L-LTP essentially involves gene expression. It is triggered by a product of the post-synaptic causal chain, the catalytic molecules PKA, which migrate to the nucleus of the cell, where they trigger expression of the *uch* gene (see Bickle 2003, 67–71). This gene transcribes the protein hydrolase ubiquitin, which triggers the transcription of other proteins that cause the growth of new dendritic spines and hence the formation of new synapses. The end result of this mechanism underlying long-term memory is an increase in the number of synapses between two neurons. This, in turn, increases the likelihood that the post-synaptic neuron will send out an action potential when stimulated, however weakly, during the L-LTP period, which can last for days or weeks.

The discovery of this molecular mechanism underlying the formation of long-term memory provides a reductive explanation of a number of properties of long-term memory first highlighted at the cognitive level. I will mention just two of them.

First, research in experimental psychology conducted at the end of the nineteenth century established that there is a linear dependence between the number of times that a stimulus is repeated during conditioning and the length of time that it is retained in long-term memory.[51] The description of the underlying microscopic processes makes it possible to explain this dependence, insofar as the brain event triggered by the repetition of the stimulus produces the state underlying memory consolidation. Stimulus repetition leads to enhanced activation in presynaptic axons, triggering the cascade of biochemical events described above, which gives rise to structural changes in synapse configuration. This explains the neuron's increased disposition to respond to similar stimuli.

Second, it has also been known since the end of the nineteenth century that brain trauma suffered after the initial phase of learning can prevent the fixation of memories related to the period immediately preceding the shock. Experimental work on this phenomenon, known as "retrograde interference," shows in particular that, if an animal is given an electric shock between twenty seconds and fifteen minutes after having undergone an experience whose memory is stored in short-term memory, that memory will not get fixed in long-term memory.[52] Furthermore, the retrograde amnesia that concussion victims suffer is explained at the neuronal level by the fact that the trauma interrupts one of the biochemical stages of the process leading to L-LTP.

We have seen with the example of thermodynamics that its reduction to classical physics does not make the reduced theory superfluous. First, the discovery of the laws that determine macroscopic properties and processes presupposes prior knowledge of macroscopic phenomena and laws. Second, the deduction of macroscopic phenomena necessarily involves concepts and principles that cannot be deduced a priori from the principles and laws governing microscopic phenomena alone. The reductive explanation, therefore, leads to a unification of knowledge and the enrichment of both theories rather than the elimination of the reduced theory.

51 The work of Ebbinghaus is presented in Squire and Kandel (1999, 130–32) and Bickle (2003, 47). However, repetition of the stimulus is not necessary. Depending on the biological species and the object of learning, certain experiences can be sufficient, without ever being repeated, to induce a stable long-term memory.

52 After Ebbinghaus and Müller and Pilzecker at the end of the nineteenth century, this research was taken up again by Duncan (1949). See Bickle (2003, 112).

Similarly, neurophysiological reduction leads to a deeper understanding of cognitive phenomena, such as learning, but does not render their psychological descriptions superfluous. Specifically cognitive concepts are indispensable for understanding certain properties of learning by conditioning. Consequently, a cognitive concept such as having learned to react with a behavioural response to the perception of the conditioned stimulus is neither eliminated when the underlying microprocesses have been discovered nor identified with any concept applying to those microprocesses. The cognitive concepts describing learning at the cognitive level remain indispensable insofar as they are necessary for the very description of the underlying microscopic mechanism. According to Rescorla (1988), such a mechanism cannot be understood in terms of a "low-level mechanical process in which the control over a response is passed from one stimulus to another" (152). Indeed, the mechanism underlying learning becomes comprehensible only insofar as we appeal to the notion of information (see Gold and Stoljar 1999, 31). Rescorla shows that there are two ways of conceiving of learning. First is the "reflex tradition in which Pavlov worked and within which many early behaviourists thought" (152). In this research tradition, conditioning is interpreted "as a kind of low-level mechanical process" (152). Second is "the associative tradition," which "sees conditioning as the learning that results from exposure to relations among events in the environment," where "the information that one stimulus gives about another" is crucial (152). Rescorla presents the transition from the old to the new theory of learning by conditioning as the replacement of a theory whose concepts are located at the physical level by an authentically psychological theory, constructed using the concept of information. This interpretation seems to be too simple: it is also a matter of replacing a crude theory with a more sophisticated one. However, the crucial point for my analysis of reductive explanation is the fact that the improvement of the theory of learning is the result not of the discovery of neurophysiological mechanisms but of the use of the cognitive concept of information.

As theories of the mechanism of learning, both theories can be expressed in cognitive vocabulary. According to the crude theory, as expressed in textbooks from the 1980s, conditioning is "a form of learning in which a neutral stimulus, when paired repeatedly with an unconditioned stimulus, eventually comes to evoke the original response" (Gardner 1982, 594, cited by Rescorla 1988, 151). For conditioning to be effective, the presentation of the CS and the US must follow a precise protocol. The CS must be presented within a

well-defined interval before the US is presented, an interval known as the interstimulus interval (ISI); optimal classical conditioning requires an ISI of half a second (see Hawkins and Kandel 1984, 379). The more sophisticated theory presented by Rescorla shows that the concept of *contiguity* between the unconditioned stimulus and the conditioned stimulus is too crude: contiguity is neither necessary nor sufficient for conditioning to occur.

The fact that it is not *sufficient* is demonstrated by Rescorla's experiments on rats. In these experiments, rats were exposed to two salient events: a tone lasting two minutes and a weak electric shock applied to the grid on which they were standing. Two experimental protocols were compared. In the first, there was no temporal correlation between the two events, so tones contained no information about shocks. Some shocks occurred in the presence of a tone, others in its absence. In the second, the rats were exposed to tones that had the same distribution as in the first protocol; however, they were exposed to shocks only in the presence of tones, and there were no shocks in the absence of tones. Both protocols satisfied the condition of contiguity between the US (the shock) and the CS (the tone), but only the rats that followed the second experimental protocol developed an association between the tone and the shock. Rescorla explains this result by the fact that, in the second protocol but not in the first, the tone contained information about the shock. In the first protocol, there was as much probability of a shock when there was a tone as when there was no tone; in the second protocol, shocks occurred only during tones. The two learning situations "share the same contiguity of the tone [the CS] with the US, but they differ in the amount of information that the tone gives about the US" (Rescorla 1988, 152). In the first group, the presence of the CS informed the animal of the presence of the US, which explains the creation of a conditioned response to the CS previously adequate only for the US. Conversely, in the second group, the CS contained no information about the US because there was as much US in the absence of CS as in its presence, which explains why the animal did not develop a conditioned response to the CS.

Contiguity is not *necessary* for learning either. In a variant of the experiment described above, starting from the protocol in which the US was present simultaneously with the CS and in the absence of the CS, all US contiguous with CS were removed. Since there was no more contiguity between CS and US, the simple theory of learning by contiguity predicts that no learning will occur. However, what the animals in fact learn in this situation is that the

CS is a reliable indicator of *the absence of* the US. They learn by developing a conditioning in which the CS acts as a conditional inhibitor (see Rescorla 1988, 153).

Rescorla takes the old Pavlovian theory of the reflex to be a "mechanistic" theory that draws its conceptual resources exclusively from the neuronal level and the cognitivist theory of learning to be a theory that makes an irreducible use of the concept of information. Now this reconstruction is as questionable as Bickle's (2003) eliminativism, which claims that the discovery of the underlying biochemical processes renders the use of psychological concepts superfluous. It seems to be more appropriate to interpret the difference between behaviourist and cognitivist theories of learning by conditioning in terms of how fine a distinction they draw. In fact, each of these theories (or each of these variants of the theory) has a corresponding reducing theory at the neurophysiological and molecular levels. Kandel's theory identifies both a molecular mechanism that underlies learning as a function of the contiguity between the CS and the US and — albeit more hypothetically — a mechanism (actually two mechanisms) underlying the *absence* of learning in a situation in which the US occurs without the CS in the intervals between simultaneous (contiguous) presentations of both the US and the CS. I have already presented the outline of a molecular explanation for the creation of an association between the CS and R originally triggered by the US: it is a variant of sensitization that requires a precise sequence in the order and temporal interval between the presentations of the CS and the US. Experimental research with *Aplysia* has shown that, if the CS is presented about half a second before the US, Ca^{2+} channels are opened when the US signal arrives, increasing the signal transmitted by the synapse to the motor neuron (see Kandel 1995, 677). But Hawkins and Kandel (1984) also proposed two molecular mechanisms that reductively explain the phenomenon discovered by Rescorla (1968) and Kamin (1969) described above: learning does not occur when there are isolated occurrences of the US in addition to contiguous presentations of the CS and the US.

Rescorla and Wagner (1972) suggested that this situation is equivalent to *blocking*. According to them, the reinforcement of a complex CS *AX* — composed of simple stimuli *A* and *X* — presented just before the US depends on the total strength that the two components *A* and *X* possessed prior to learning. The phenomenon of blocking consists of the fact that "prior conditioning of *A* reduces the degree to which reinforcement of an *AX* compound

increments the associative strength of *X*" (Rescorla and Wagner 1972, 77). In cognitive terms, their theory explains this by the fact (which emerges from their equations) that the variation of the associative strength of the CS *X*, ΔV_X, is proportional to $\lambda - V_{AX} = \lambda - (V_A + V_X)$ (i.e., the difference between the maximum strength λ, which depends on the US used, and the total associative strength V_{AX} of all the stimuli present). Therefore, if V_A is already close to λ because of its previous conditioning, then the presence of *AX* before the US will not significantly increase either *A*'s or *X*'s associative strength. In this case, ΔV_X is close to 0, because V_A is close to λ, whereas V_X is close to 0 because *X* has not been involved in previous conditioning.

Hawkins and Kandel offer a molecular reduction of this blocking phenomenon. During conditioning of *A* (called CS_1 in Hawkins and Kandel 1984), the facilitative interneurons trigger the response more and more exclusively following the presence of *A*, at the expense of its triggering by the presence — immediately afterward — of the US. This is explained by accommodation and recurrent inhibition (Hawkins and Kandel 1984, 385). In the end, the US no longer elicits R, the response being monopolized by *A*. When the complex stimulus *AX* (CS_1 CS_2 in Hawkins and Kandel 1984, 386) appears at the second step of the type of learning that manifests blocking, the presence of stimulus *X* (CS_2) is not followed by the activation of facilitator neurons, which would be necessary for the conditioning of *X*.

Hawkins and Kandel propose to follow Rescorla and Wagner's hypothesis that the situation outlined above, in which isolated presentations of the US alternate with presentations of the CS in contiguity with the US, constitutes a variant of the blocking situation. In the molecular reducing theory, the absence of learning in this situation is explained by the hypothesis that an intermittently presented US produces "conditioning to background stimuli," which "cause continuous excitation of facilitator neurons, rendering them insensitive to the US" (Hawkins and Kandel 1984, 388, 387).

The explanation of the phenomena of blocking and the absence of learning, when there is no reliable correlation between the CS and the US, can therefore be completed by highlighting the underlying neurophysiological processes. This does not mean that cognitive explanations become superfluous. The complex properties of learning by conditioning shown by Rescorla and others can be understood only in cognitive terms, not in purely neurophysiological terms. The identification of the underlying biochemical mechanisms leads one to justify, and sometimes modify, their cognitive

explanation. As with any reduction, its fruitfulness is measured by its capacity to induce modifications in the reduced theory as well as in the reducing theory. The discovery of a new phenomenon at the psychological level might require the modification of neuronal and molecular theory, but in the same way aspects of conditioning first discovered at the molecular level might require the modification of psychological theory.

According to some, the discovery of the mechanisms underlying learning by conditioning is part of an evolution that leads, in the long run, to the replacement of psychological theories of learning, and of the psychological concepts involved, by purely neurophysiological theories and concepts.[53] However, the discovery of a reductive explanation of a psychological phenomenon does not lead to its "elimination" as a psychological phenomenon. On the contrary, the discovery of the processes underlying the phenomena of learning strengthens our reasons for believing that these phenomena exist. This is particularly clear when the discovery of the underlying neurophysiological mechanism makes it possible to explain a phenomenon in the precise form attributed to it by the psychological theory subject to reduction. This is illustrated by the mechanism underlying classical conditioning, which explains why the simple contiguity of the CS and the US is neither necessary nor sufficient for conditioning. The elimination of psychological concepts is not justified either when the discovery of the underlying neurophysiological mechanism leads to a correction of the psychological theory: it continues to use the psychological concepts of the CS and the US.

If the reduction of a psychological theory does not lead to the elimination of the psychological phenomenon, then we might be tempted to conclude that it leads to its *identification* with the underlying neurophysiological mechanism (see Causey 1977; Churchland and Churchland 1994). The reduction shows that the cognitive capacity — that of learning to react (with R) to the CS as if it were the US — is identical to a microscopic property, in this case the property of being in a state of sensitization of the pre-synaptic termination of the sensory neuron (originating from the CS), which has synapses with the motor neuron leading to R or with interneurons leading to R. In a similar way, it might be said that the exercise of the capacity is identical to the unfolding of the underlying mechanism.

53 This is the thesis of Bickle (2003). Gold and Stoljar (1999) call it the "radical doctrine of the neuron" (after Barlow 1972).

It is one of the central theses of this book that such an identification is justified neither with regard to the relationship between the temperature of a gas and the kinetic energy of the molecules of which it is composed, nor with regard to the relationship between water and the H_2O molecules of which it is composed, nor finally with regard to the relationship between the conditioning process and the underlying microscopic processes. At first sight, it might seem that the reduction of the (macroscopic) property of being water shows that it is *identical* to the (microscopic) property of being H_2O, in the same way that it might seem that the reduction of the (macroscopic) property of having the temperature T shows that it is identical to the property of being composed of molecules having an average kinetic energy E_{kin}. In the first case, this appearance is the result of an ambiguity in the expression "is H_2O." The property of being a molecule of H_2O is microscopic and can only belong to molecules. Yet the property of *being made up of* H_2O molecules is macroscopic. The reduction of the property of being water shows that this property is identical to the second, which is macroscopic, but not to the first, which is microscopic. Similarly, the reduction of the macroscopic property of having temperature T does not lead to the identification of this property with the microscopic property (of the individual molecules) of having the average kinetic energy E_{kin} or with the property of being a set of molecules whose average molecular energy is E_{kin}: many sets of molecules have an average kinetic energy without having a temperature because they do not interact with each other (see Kistler 1999c). The reduction of the macroscopic property of having temperature T leads to its identification with the macroscopic property of having microscopic components whose interaction allows them to exchange energy and whose average kinetic energy is E_{kin}. The details of the reduction show how temperature is determined by the microscopic properties of the components of the object that has the temperature and by the interactions among these components. In the same way, the reduction of cognitive properties and processes — such as the disposition to learn by conditioning, learning by conditioning itself, and the state of being conditioned in a given way — does not lead to their identification with microscopic properties and processes. Such a neurocognitive reduction shows that the cognitive property of an organism is identical to the property of having parts articulated in a given way so that the neurophysiological properties of the parts and their articulation determine — in a nomological way — the cognitive property of the organism. For example, the reduction of an organism's cognitive state of

having learned an association between the CS and the US shows that this is the property of having neurophysiological parts articulated in a certain way. The reductive explanation shows that the neural properties of certain parts of the organism determine the overall property of the organism.

The anti-reductionist conclusion of Gold and Stoljar that "the claim that Kandel's model is a reduction of classical conditioning . . . cannot be sustained" (1999, 825) is based on a conception of reduction as identification. Yet, when we construe reduction as the demonstration of a non-causal relationship of determination[54] of the global properties of a complex system by the properties of its parts and their interactions, it can be argued both that Kandel's model succeeds in reducing[55] certain forms of classical conditioning and that "the concept of synaptic change cannot capture the concept of information or surprise" (Gold and Stoljar 1999, 825). The concepts of information and surprise belong to the cognitive level and apply to the organism, whereas the concept of synaptic change belongs to the neurophysiological level and applies to parts of the organism that play a key role in the reductive explanation of conditioning. The reductive explanation shows in detail how each episode of learning unfolds. Concepts such as information and surprise used in Rescorla's (1988) theory of conditioning can explain, from an evolutionary perspective, why the conditioning process obeys the laws that I have outlined above. To use Dretske's (1988) distinction, reduction allows us to understand the mechanism of conditioning in terms of its "triggering cause," whereas we need to use the notion of information to give a teleological and functional explanation of this mechanism in terms of its "structuring cause." Natural selection helps to explain the appearance of this learning mechanism during evolution: animals capable of conditioning can adapt to their environments because conditioning enables them to act in ways appropriate to the presence of the US even before it is perceived, insofar as the CS objectively contains the information that the US will occur.

Nagel imposes "non-formal" conditions for the success of a reduction.[56] In the ideal case, a reduction induces new hypotheses and research direc-

54 The determination of the properties of a complex system by the properties of its parts and their relationships cannot be causal because it does not extend in time: it is a form of simultaneous determination, whereas causes must precede their effects. I will return to this point in Chapter 4.

55 More precisely, it is a reductive hypothesis that leaves open the question of its truth.

56 "For a reduction to mark a significant intellectual advance, it is not enough that previously established laws of the secondary science be represented within the theory of the primary discipline.

tions, both in the reduced macrotheory and in the reducing microtheory. Many "conservative" reductions — which do not lead to the elimination of the reduced theory — have had the effect of inspiring improvements in both the reduced theory and the reducing theory. However, this situation is conceivable only if each of the two theories is situated within and explains a proper domain of phenomena. The CHB reduction model seems to exclude this possibility: insofar as it is possible to construct an adequate description of macrophenomena within the microtheory, the macrotheory loses the autonomy necessary to inspire new hypotheses or corrections to the hypotheses of the microtheory. Depending on the quality of the analogy between T_R^* and T_R, the old macrotheory is eliminated (if the analogy is bad) or preserved approximately (if it is good). But in the latter case what is retained, strictly speaking, is T_R^*, which has no autonomy in relation to T_B. Insofar as T_R differs from T_R^*, it is false, but false theories cannot inspire corrections to correct theories.

My analysis of the reduction of thermodynamics shows that there are reasons to abandon the requirement that T_R^* be derivable from T_B without linking assumptions. According to the synthetic model of reduction introduced above, the deduction of T_R^* from T_B is not a logical derivation but involves non-analytical laws. If we assume that the macroproperties are determined by the microproperties and their interactions by virtue of laws of nature, and not only by virtue of logical and mathematical rules of calculation, then it seems to be legitimate and necessary to pursue the development of theories at both levels in parallel in order to improve the reduced theory, the reductive theory, and our knowledge of the laws of composition used in reduction. To describe this situation, Robert McCauley introduced the notion of "co-evolution" of theories that deal with the same phenomena but at different levels. He distinguished three variants. Only one of them, "co-evolution$_p$," gives rise to what he called "explanatory pluralism" (1996, 27), in which theories at different levels influence each other. This typically leads to the emergence of a new "interfield theory"[57] that forges a synthesis of the reduced theory, the reducing theory, and the links of determination between them.

The theory must also be fertile in usable suggestions for developing the secondary science, and must yield theorems referring to the latter's subject matter which augment or correct its currently accepted body of laws" (Nagel 1961, 360). Also see McCauley (1981); Enc (1983).

57 This concept was introduced by Maull (1977) and Darden and Maull (1977).

In contrast, the CHB model only allows for the possibility of co-evolution$_M$, in which the reduced theory is justified by its derivation from T_B, but preserves no conceptual independence from T_B, and co-evolution$_S$, in which the "reduced" theory is eliminated. We have seen that, in the cases that we have examined, the conditions for reduction of the co-evolution model$_M$ are too strong. Co-evolution$_S$, which according to Paul and Patricia Churchland is the most appropriate model for reducing psychology to neuroscience, appears from this perspective to be the result of a "category mistake" (McCauley 1996, 34). It seems to be plausible for one theory to eliminate another only when these "theories compete for the same logical space" (Endicott 1998, 59)[58] — that is, seek to account for the same phenomena. Now a microtheory that reduces a macrotheory does not meet this condition, or it could meet it only if the reduction conformed to the CHB model. If the first stage of the reduction consisted of a derivation of T_R^* without recourse to linking hypotheses, then the microtheory would cover, through its implication of T_R^*, the same domain of phenomena as T_R. However, insofar as the resources of T_B alone are not sufficient to cover the macrophenomena in the domain of T_R and T_R^*, T_B and T_R (as well as T_B and T_R^*) are not in competition. This removes the plausibility of the idea that T_B can eliminate T_R, even when there is a reduction.

The correction of theories in the course of their unification or reductive integration can be reciprocal (Schaffner 1993, 427–29). It is not always only the reduced theory that is corrected, as suggested by the CHB model: in the case of reductions that have been achieved, the higher-level reduced theory suggests as many avenues of research in the microtheory as the latter suggests in the former.

As we saw above, Rescorla (1968) established at the psychological level that learning by classical conditioning is impossible when, between perceptions in which the CS appears to be associated with the US, the US appears alone, unaccompanied by the CS. This phenomenon, first discovered at the cognitive level, prompted Hawkins and Kandel (1984) to look for a cellular mechanism that could provide a reductive explanation. The fact that the reduced theory still contains elements that can suggest research at the level of the reducing theory, even after the reduction has been completed, undermines the CHB model. According to that model, the reduced theory does

58 This analysis is inspired by Wimsatt's (1976a, 222) comparison between reduction between levels and reduction within a level.

not retain sufficient conceptual autonomy to be able to inspire research at the level of the reducing theory. If the psychology of learning lost all conceptual independence as a result of its reduction, then how could it be a fruitful source of research in neuroscience? Endicott takes this reasoning a step further: insofar as the reducing theory is influenced by constraints "from above" (i.e., from the reduced theory), "the basic reducing theory becomes permeated with high-level concepts and concerns" (1998, 65; see also Gold and Stoljar 1999; van Eck, Looren de Jong, and Schouten 2006).

Functional concepts from molecular biology — such as signal sequence (a sequence of amino acids containing a protein, which has the function of directing the protein to its destination), antibody, secondary messenger, and receptor protein — can be "incorporated in an *integrated interlevel* theory" (Kincaid 1990, 590). However, these concepts cannot be "reduced" to molecular biology (in the sense of being replacable in principle by concepts from the latter[59]) because the explanation of the mechanisms underlying the exercise of these functions requires the use of other macroscopic concepts of cell biology. It is possible in principle, for example, to specify the molecular composition of any antibody. But there are no molecular-level properties common to all antibodies, of which there are millions.[60] The only property that they have in common is the functional property of establishing a bond with an antigen so that this bond triggers an immune reaction. Identifying the underlying mechanism offers no prospect of eliminating the concept of antibody, which alone makes it possible to express a regularity at the macroscopic level, invisible from the point of view of the multitude of underlying microscopic mechanisms.

With regard to research on the mechanisms underlying vision, Bechtel concludes that "there is no basis for assuming that one can provide a complete account of the functioning of the mechanism in terms of the parts alone. The behaviour of the mechanism depends not just on the parts but how they are organized and the context in which they are situated" (2009, 559–60).

59 Kincaid presupposes "the root notion of reduction — that one theory can do all the work of or replace another" (1990, 590).

60 This observation is reminiscent of the one that I made earlier about the failed attempt to reduce thermodynamics to mechanics without appealing to probability but by indicating the initial conditions that characterize systems whose behaviour is in accordance with macroscopic laws. It turns out that these initial conditions can be specified only in terms of macroscopic concepts of thermodynamics: these initial conditions characterize systems that conform to thermodynamics.

To understand the neurophysiological mechanism of vision in terms of the articulation of its component parts, we need to analyze the function of vision as a whole in an animal's interaction with its environment. If we were to try to understand vision merely from the perspective of its neurophysiological mechanism, then we would tend to forget that the function of vision is to inform the cognitive subject about its environment.[61]

8. Conclusion

The reduction between two theories that study the same domain of phenomena at different levels is a major conceptual tool for understanding the process of the unification of science. The rise of cognitive neuroscience is just the latest episode, albeit a particularly spectacular one, in the process of unifying domains of knowledge concerning different scientific theories. The interpretation of this unification, which merges the formerly separate sciences of psychology and neuroscience into a single theory, is of particular importance insofar as it concerns psychology. There is a long tradition of claiming the autonomy and irreducibility of psychology. The prospect of the reduction of psychology gives rise to particularly intense fears and hopes. Given the importance that we attach to our minds, we might fear that such a reduction would reduce us to the level of mere assemblies of cells and thus risk undermining our moral dignity. But we can also hold out hope that we will finally gain understanding and explanation of the mysteries surrounding our minds, such as the origin of mental illnesses, their dependence on certain brain dysfunctions and new ways of curing them, and the function and significance of sleep and dreams. One of the aims of this book is to show that the prospect of reducing psychology to the neurosciences appears to be dramatic and worrying only when viewed under particular interpretations of what a reduction is. Others are compatible with the intuition of the autonomy of psychology and with the existence of a mind or, more precisely, with the existence of cognitive and mental properties distinct from the neurophysiological properties of our brains.

61 The fact that perception depends as much on the neurophysiological mechanism as on the interaction of the cognitive subject with its environment led Clark and Chalmers (1998) to the "extended mind" hypothesis. According to this hypothesis, cognition is extended beyond the cognitive subject to include the environment. See Clark and Chalmers (1998); O'Regan and Noë (2001); Clark (2008).

In the tradition of logical empiricism, the reduction of one domain of phenomena to another is conceived of as an explanatory relationship between the theories that cover these domains of phenomena. The reduction from one theory to another consists of a deductive-nomological explanation: each of the axioms and principles of the reduced theory is deduced from premises taken from the reducing theory. In his now canonical presentation of this conception, Ernest Nagel introduces the distinction between homogeneous and heterogeneous reductions. The reduction of psychology to neuroscience belongs to the category of heterogeneous reductions. In such cases, the reduced theory contains concepts that do not appear in the reducing theory: neurophysiology, for example, knows nothing about motivation, perceptual discrimination, or iconic memory. The debate on the interpretation of heterogeneous reductions hinges on the status of linking statements, or "bridge laws," that must be introduced if we hope to find a deductive explanation of the axioms of the reduced theory, on the basis of the reducing theory. These linking statements are neither metalinguistic and analytical nor, in general, identity statements. In the following chapters, I will examine three other important hypotheses regarding linking statements. According to the hypothesis of conceptual reductionism analyzed in Chapter 2, linking statements can be deduced a priori from the reducing theory alone. According to the hypothesis of classical emergentism analyzed in Chapter 4, linking statements are primitive and inexplicable "transordinal" laws. According to functionalist reductionism, examined in Chapters 3 and 5, linking statements define the conceptual relationship between a functional role and what occupies that role.

I have suggested that linking statements are non-causal laws of a particular type, which I propose to call "composition laws." These are laws that determine the global properties of a system according to the laws that govern the properties of its parts and their interaction. We will return to this concept in Chapters 3 and 4. I have motivated and illustrated it here with two examples: the reduction of temperature to mechanics and the reduction of learning and memory to neurophysiology.

Nagel's model has been mostly criticized for overlooking the fact that historical reductions are only rarely conservative. In general, the reduction of a theory is accompanied by its correction. The main motivation for seeking a reduction is the hope of improving the existing theory. However, insofar as a reducing theory corrects the higher-level theory, as it was before the

reduction, it is impossible for the latter to be logically deducible from the reducing theory. We have seen that there are reductions in which the theory T_R^* that can be deduced from the reducing theory T_B is not identical but only structurally analogous to the reduced theory T_R. However, I have questioned the thesis of Nagel's critics according to which it is possible to deduce T_R^* from T_B without using bridge laws. In particular, I have shown that the reductionist explanation of T_R^* from T_B is not an intratheoretical reduction, as predicted by the model put forward by Churchland, Hooker, and Bickle. When we examine historical cases of reduction, it turns out that the reduced theory T_R^* is not derived from the assumptions of T_B alone. In the case of the reduction of thermodynamics to classical physics, we have seen that the use of bridge hypotheses is indispensable. Similarly, the reduction of memory fixation and learning by conditioning is intelligible only within the conceptual framework of the reduced theory T_R (or T_R^*). The reduced theory is the starting point for the reduction and guides the search for a reducing theory. To take account of the a posteriori nature of the discovery of the laws of composition between the levels of T_B and T_R^*, I have suggested a two-part model of reduction. The first or "interlevel" part of this synthetic model of reduction corresponds to the composition laws, discovered on the basis of prior knowledge of T_R and T_B. The second or "intralevel" part corresponds to the demonstration that there is a structural analogy between the theory T_R^* deduced from T_B, thanks to the laws of composition, and the reduced theory T_R.

The thesis that the theory T_R^* (deduced from T_B and the composition laws) might be only structurally analogous to the reduced theory T_R, without being strictly identical to it, allows us to explain the possibility of reducing multi-realizable properties. In their case, the T_R theory undergoes a separate reduction in each type of system to which T_R applies. The theories $T_R^*{}_1$, $T_R^*{}_2$, and so on, deduced from different reducing theories T_{B1}, T_{B2}, and so on, and specific to different types of systems, are all analogous to each other and to the reduced theory T_R, without being identical either to each other or to T_R. This is all the more important in the case of psychology: for example, the physiological diversity of the different animal species to which the theory of learning applies provides the main reason for holding the latter to be irreducible to neurophysiology. Given that the structural similarity of the reducing theories — specific to each species — to the reduced theory is sufficient for a reduction, the diversity of neurophysiological substrates is no longer a reason for holding psychology to be irreducible.

The fact that general psychology remains different from — albeit structurally analogous to — species-specific theories also helps to explain why it retains a certain autonomy, even once it has been reduced. This autonomy is essential to explain the fact that discoveries made at the level of the reduced theory often inspire modifications in reducing theories. The observations made at each level with the help of concepts specific to that level are indispensable constraints on the development of interlevel theories. Cognitive neuroscience is such a theory, where items of knowledge obtained separately at the psychological and neurophysiological levels influence and illuminate each other mutually.