philosophia naturalis

Journal for the Philosophy of Nature

Reduction in the Philosophy of Mind

Raphael van Riel and Albert Newen (eds.)

KLOSTERMANN

Band/Volume 47-48/2010-11

Heft/Issue 1-2

philosophia naturalis

Journal for the Philosophy of Nature

47-48 / 2010-11 / 1-2

KLOSTERMANN

The articles are indexed in *The Philosopher's Index* and *Mathematical Reviews.*

Abonnenten der Printausgabe können über Ingentaconnect auf die Online-Ausgabe der Zeitschrift zugreifen: *www.ingentaconnect.com*

Zurückliegende Jahrgänge sind mit einer Sperrfrist von fünf Jahren für die Abonnenten von *www.digizeitschriften.de* zugänglich.

Satz: Mirjam Loch, Frankfurt am Main / Druck: KM-Druck, Groß-Umstadt.
Gedruckt auf alterungsbeständigem Papier ♾ ISO 9706.

ISBN 978-3-465-04126-9
ISSN 0031-8027

Preface

Under a certain interpretation the whole modern philosophy of mind can be conceived as being concerned with issues of reductionism or anti-reductionism – either by trying to free the community from its Cartesian heritage or by trying to reinstall it in modern terms; debates on type- and token-theories, on functionalism, on supervenience, on mechanistic explanations of the mind, on consciousness and on phenomenality bear directly upon or are explicitly concerned with issues of reductionism and anti-reductionism, and it seems that these are not the only ones. Consequently, the questions of (i) what reduction consists in and (ii) whether or not reductionism is true figured among the most prominent in the philosophy of mind, but also in related areas like metaphysics and philosophy of science, in the early second half of the last century. Due to relatively recent developments in the neurosciences, which were enthusiastically described as pushing us towards a natural science of the mind, they have undergone some sort of revival in the past decades. Moreover, recent interpretations of models of reduction seem to suggest that the alleged problems for reductionism, as, for example, posed by arguments which are based on the assumption of the multiple realizability of mental kinds, do not affect the reductionist's claims at all – identification of mental kinds with disjunctive kinds or with contextualized kinds which are seemingly not (relevantly) multiply realizable form two attempts to reconcile the fact of multiple realizability with reductionism.

Simplifying, one can distinguish three steps within the debate: From (reductionist) type-identity theory to (non-reductionist though still materialist) token-identity theory and, finally, to more recent debates about the compatibility of token-identity theory with reductionism. The latter have been inspired by the development of new interpretations of type-identity theory and functional reduction. This strand of the discussion was paralleled by a number of more specialized debates, such as debates about the reducibility of phenomenal aspects of consciousness, intentionality, natural kinds, reductive and mechanistic explanation, and mental causation. This volume contains papers of the former as well as of the latter kind. In the first part, the papers of Michael Esfeld, Douglas

Kutach, Christian Sachse and Robert Van Gulick discuss rival *models of reduction*. Some of these papers draw conclusions from closely related fields, like the philosophy of biology and the philosophy of science in general. The papers of the second part are concerned with aspects of the *explanatory link* between the reduced and the reducing level. Markus Eronen discusses the relation between functional reduction and mechanistic explanation, whereas Raphael van Riel focuses on the relation between identity-based reduction and reductive explanation. Phenomenality is one of the most demanding issues in the context of reductive explanation. Albert Newen suggests a theory of phenomenal concepts, analyzing them in terms of mental files, which allows us to account for the knowledge argument from an antecedent physicalist's perspective.

Michael Esfeld focuses on the relation between identity-based reduction, functionalism, and eliminativism. He argues in favour of a causal-functional theory of properties including the physical ones and a conception of properties as tropes or modes in the sense of particular ways that objects are. Esfeld gives an idea of how these premises open up a version of functionalism according to which the properties on which the special sciences focus are identical with configurations of physical properties and thereby causally efficacious without there being any threat of eliminativism.

In his closely related paper, Christian Sachse argues in favour of a reductionist strategy in the philosophy of biology in order to maintain the unity of science. Basically, Sachse describes a conservative, non-eliminativist reductionist strategy which is based on the theoretical possibility of constructing functionally defined sub-concepts in biology that are nomologically coextensive with physical descriptions. This *theoretical* link between biology and physics makes it possible to understand the original and operational biological concepts as abstractions from these biological sub-concepts. Against this background, Sachse argues that biological kinds are natural ones and that biology adumbrates laws and explanations of different degrees of abstraction.

Douglas Kutach proposes an interpretation of reduction which fits into the more general framework of 'Empirical Fundamentalism', whose signature feature is the extensive use of a distinction between fundamental and derivative reality. Within the framework of Empirical Fundamentalism, derivative reality is treated as an abstraction from fundamental reality. Kutach examines how one can understand reduction

and supervenience in terms of abstraction. On this basis, the introduced machinery is applied to understand paradigmatic examples of allegedly reductive relations, like the relation between water and H2O and mental states and brain states.

Van Gulick develops an account he calls "teleo-pragmatic functionalism". He discusses what that view implies about the nature of cognition, theories and understanding and thus about the limits on our ability to explain the mental and its relation to the non-mental. It is argued that teleo-pragmatic functionalism leads naturally to a version of non-reductive physicalism that combines theoretical pluralism with a strongly contextualist and pragmatic view of theories and models. This account is then judged non-reductionist at the theoretical and conceptual level, and, at the same time, thoroughly and robustly physicalist in its ontology.

Thus, these four essays deal with general conceptions of models of reduction and reductionist strategies in science. The essays contained in the second part of this volume discuss issues related to the explanatory link between reduced and reducing level.

Markus Eronen argues that the functional model fails as an account of reduction due to problems related to three key concepts: functionalization, realization and causation. Suggesting a revision of the model which accommodates these problems, Eronen then argues that functional reduction collapses into what has been described as mechanistic explanation. Hence, instead of analyzing reduction in philosophy of mind in terms of functional reduction, Eronen concludes that it should be analyzed in terms of mechanistic explanation.

Albert Newen focuses on reductive strategies concerning phenomenal experiences, thus focussing on what is usually regarded as lying at the core of the *explanatory gap* problem. Building on the idea that we can account for phenomenal aspects of experience without being committed to the claim that there is a *sui generis* kind of qualia, and building, secondly, on the idea that problems posed by arguments against the reducibility of qualia can be overcome adopting a phenomenal concept strategy, he develops an account of phenomenal concepts. Phenomenal concepts are analyzed as theory-based concepts, in the sense that possession of a phenomenal concept requires possession of at least a mini-theory. Phenomenal concepts are then spelled out using the idea of mental files and the account is applied to the *knowledge argument.*

Raphael van Riel discusses the relation between identity-based reduction and reductive explanation. He argues (against Jaegwon Kim) that identity-based reduction is perfectly compatible with corresponding reductive explanations. From the discussion of this point, van Riel draws the conclusion that identity-based reduction is partly defined via conceptual contents under which reduced and reducing kind, property, or phenomenon are presented.

This volume has its roots in an international workshop entitled "Reductionism, Explanation and Metaphors in the Philosophy of Mind". The workshop was held as a satellite event of the GAP-Conference – the tri-annual conference of the *German Society for Analytic Philosophy* (*Deutsche Gesellschaft für Analytische Philosophie*) in Bremen, in September 2009. Some selected contributions of the meeting together with further submitted articles constitute the special issue. The papers have been evaluated by a peer-review process. Concerning the meeting we would like to express our thanks to the local organizers for their assistance, to the GAP for their generous funding, and to the speakers and the audience for stimulating talks and interesting discussions. The organization of the meeting and the preparation of the edition were managed at the University of Bochum with important organizational support by Josua Faller, Lara Kirfel, Sebastian Lorenz, and Robert Schütze. Finally, we would like to thank the management of the journal, with a special thanks to Andreas Bartels and Carsten Seck.

Raphael van Riel and Albert Newen

Part I Models of Reduction

Michael Esfeld

Causal Properties and Conservative Reduction[1]

Abstract

The paper argues in favour of (1) a causal-functional theory of all properties including the physical ones and (2) a conception of properties as tropes or modes in the sense of particular ways that objects are. It shows how these premises open up a version of functionalism according to which the properties on which the special sciences focus are identical with configurations of physical properties and thereby causally efficacious without there being any threat of eliminativism.

Zusammenfassung

Der Artikel argumentiert (1) für eine kausal-funktionale Theorie aller Eigenschaften einschließlich der fundamentalen physikalischen Eigenschaften und (2) für eine Konzeption von Eigenschaften als Tropen oder Modi im Sinne der jeweiligen Weisen, wie die Objekte existieren. Er zeigt auf, wie diese Prämissen es ermöglichen, eine Version des Funktionalismus zu vertreten, gemäß der die Eigenschaften, von denen die Einzelwissenschaften handeln, mit Konfigurationen physikalischer Eigenschaften identisch sind und dadurch kausal wirksam sind, ohne dass die Gefahr einer eliminativistischen Konsequenz droht.

1. The dilemma of functionalism

Since the 1970s, functionalism has been the standard position with respect to the special sciences such as, in particular, biology and psychology. Functionalism seems to be able both to show how the special sciences and the entities they deal with are related to fundamental physics and to pay heed to their specific character. Since the 1990s, however,

philosophia naturalis 47-48 / 2010-11 / 1-2

serious doubts have been raised as to whether functionalism can really achieve that target. The aim of this paper is to set out a version of functionalism that starts from the idea that all properties in the world are functional in the sense of causal properties and that avoids on this basis the pitfalls of both epiphenomenalism and eliminativism. I first point out the dilemma of epiphenomenalism and eliminativism into which the standard versions of role and realizer functionalism run (this section) and then argue in favour of the causal theory of properties (section 2). On this basis, I draw the consequence of a conservative ontological reductionism (section 3). Finally, I maintain that ontological and epistemological reductionism stand or fall together (section 4).

Mainstream functionalism as conceived notably by Hilary Putnam (1967/1975) and Jerry Fodor (1974) regards the properties on which the special sciences focus as causal roles that are realized by configurations of physical tokens (*role functionalism*). For instance, a gene type consists in a causal role, such as to produce proteins of a certain type, and that role can be realized by different types of DNA configurations in certain molecular contexts. Role functionalism thus binds the property types on which the special sciences focus to the physical domain through realization and establishes at the same time their specific character through multiple realization, which prevents them from being identical with physical types.

Nonetheless, role functionalism faces the same problem as the theory of non-physical, emergent properties, namely that it is not intelligible how the role properties can be causally efficacious. According to role functionalism, the functional properties that are the subject matter of the special sciences are causal roles, defined by a characteristic pattern of effects each. However, the role properties as such are not causally efficacious. It is the properties that carry out the role, the physical properties, that bring about the effects in question. In other words, the presence of functional role properties indicates that there are other properties, the physical realizer properties, that produce certain effects. Hence, if one conceives the functional properties that are the subject matter of the special sciences as role properties, being distinct from realizer properties, one faces the consequence that these functional properties are epiphenomenal. Only the physical realizer properties are causally efficacious (cf. Block, 1990).

One may attempt to avoid this consequence by invoking systemat-

ic overdetermination (see notably Bennett, 2003; Loewer, 2007). The idea is that both the role properties and the realizer properties figure in the relevant causal relations so that one and the same effect – such as the production of certain proteins – is overdetermined by two causes that are not identical with one another, such as genes and molecular sequences of bases. There is systematic overdetermination because any physical effect of a role property is caused at the same time by the realizer property in question – and given supervenience, there is a sufficient physical cause for any effect whatsoever that a role property causes. The proponents of this idea maintain that such systematic overdetermination is acceptable since the role properties strongly supervene on the realizer properties. Strong supervenience is to say that the existence of the physical realizer properties is a metaphysically sufficient condition for the existence of the functional role properties in question.

Strong supervenience is sufficient to make certain counterfactual propositions true. Suppose that in a situation in which a physical effect p_2 has both a sufficient mental cause m_1 and a sufficient physical cause p_1, the proposition "If p_1 had not occurred, p_2 would not have occurred either" is a true counterfactual. Then, by strong supervenience, the proposition "If m_1 had not occurred, p_2 would not have occurred either" is a true counterfactual as well, since there is no possible world in which there is a counterpart of p_1 without there being also a counterpart of m_1. However, we need an argument why that latter counterfactual should express a causal relation. In other words, we need an argument why strong supervenience on its own should exclude epiphenomenalism, that is to say, should exclude that a property which strongly supervenes on another property can be epiphenomenal. The mere fact of a property strongly supervening on another property cannot constitute a sufficient reason for claiming that the supervenient property also causes – some – of the effects that the subvenient property has so that these effects are overdetermined (see Esfeld, 2010). Such a claim would amount to simply stipulating that properties which strongly supervene on other properties cannot be epiphenomenal, without offering any argument.

Since Jaegwon Kim (Kim, 1998) has laid stress on the problem of the causal efficacy of the functional properties on which the special sciences focus, the reductionist version of functionalism, the *realizer functionalism* that goes back mainly to David Lewis (see Lewis, 1994, for a summary), has gained in popularity and is to a certain extent favoured by

Kim himself (Kim, 1998, and 2005). Like role functionalism, realizer functionalism starts from the functional descriptions that the special sciences use. However, whereas role functionalism conceives these descriptions as being about causal role properties that are second order properties, realizer functionalism takes them to refer to the physical realizer properties. It conceives the descriptions of the world that the special sciences offer as functional descriptions, being fulfilled or realized by in the last resort configurations of microphysical tokens. In his metaphysics of Humean supervenience, Lewis (1986a, introduction) regards the fundamental physical properties as perfectly natural, purely qualitative and intrinsic properties that occur at space-time points. Everything else that there is in the world supervenes on the distribution of the fundamental physical properties in space-time in the sense that it is a feature of that distribution. There is nothing over and above that distribution, it being sufficient to make true all the truths about the world. In particular, the descriptions (theories, laws) that the special sciences offer are true in virtue of certain features of contingent regular co-occurrence or counterfactual dependence in the distribution of the fundamental physical, purely qualitative properties as a whole. In a nutshell, there are no functional properties of the special sciences, but only categorical properties of physics. Nonetheless, the functional descriptions of the special sciences are true. The distribution of the categorical properties in space-time makes them true. Lewis' realizer functionalism thus provides for truthmakers of the functional descriptions of the special sciences without being committed to recognizing functional properties that exist in the world.

Consequently, multiple realization is construed as an epistemological rather than an ontological feature. Multiple realization means that the functional descriptions of the special sciences refer multiply in the following sense: descriptions of one and the same type of a special science refer to physical configurations that come under different physical types due to their differences in composition. These configurations of categorical, physical properties all make true descriptions of a certain functional type, without having anything physical in common that distinguishes them from all the other physical configurations that do not make a functional description of the type in question true.

By way of consequence, it is, however, questionable how these functional descriptions can have a scientific quality. All the effects that the

physical configurations that make these descriptions true can have possess complete physical causes and can therefore be explained exclusively in physical terms. If the functional descriptions of the special sciences are not about genuinely functional properties, but refer to configurations of categorical, physical properties and if these configurations do not have anything physical in common that distinguishes them from all the other configurations, then it is not intelligible what contribution to a scientific account of the world the functional descriptions of the special sciences could make that is not already provided for by physics.

Conceiving multiple realization in that manner as multiple reference can therefore also be received as an invitation to construct on the basis of a fundamental and universal physical theory several specific physical theories that map the special science theory in question for a particular domain or species in which there is uniform realization – e.g. one specific theory that maps the psychological theory of pain in the case of humans, another specific theory that maps the psychological theory of pain in the case of octopuses, etc. (see Lewis, 1980, and Kim, 1998, 93–95, as well as Bickle, 1998). Each of these physical theories then takes the place of the special science theory for a particular domain or species. Hence, there is then nothing left in which the scientific quality of the theories of the special sciences could consist.

In sum, realizer functionalism adopts an eliminativist attitude with respect to the functional properties in which the special sciences deal. The descriptions (laws, theories) of these sciences are nevertheless true, being made true by physical configurations. However, this eliminativist attitude as regards the functional properties leads to the consequence that – due to multiple realization or, more precisely, multiple reference – the scientific quality of the special sciences is also eliminated. Only purely physical descriptions and theories that apply to particular, physically defined groups each and that are integrated into an encompassing physical theory belong in the last resort to a scientific account of the world.

Functionalism as it stands thus faces a dilemma between epiphenomenalism (role functionalism as non-reductive physicalism) and eliminativism with respect to the scientific quality of the special sciences (realizer functionalism as reductive functionalism). Nonetheless, there is no other position visible that could take the place of functionalism. The two main non-functionalist alternative positions of either admitting

non-physical emergent properties or of retreating to a physicalism that recognizes only physics obviously run into the mentioned dilemma as well – epiphenomenalism in the first case, eliminativism in the second one. This diagnosis suggests the following conclusion: functionalism still is an attractive position. However, it has to be spelled out in another manner than in the standard versions of role and realizer functionalism.

This paper takes the mentioned dilemma as a motivation for drawing the following two conclusions:

– It is wrong-headed to conceive an opposition between functional and physical properties (or functional and physical descriptions, respectively).
– It is wrong-headed to build anti-reductionist arguments on multiple realization.

On this basis, the paper seeks to make a first step towards overcoming that dilemma by proposing that all properties, including the physical ones, are functional properties in the broad sense of causal properties, and showing how this view of properties paves the way for a conservative reductionism.

2. Causal properties

Leaving aside the issue of qualia, let us take for granted that the special sciences' descriptions are functional descriptions and that, accordingly, the properties in which they trade are functional ones. Functional properties are causal properties. What they are consists notably in certain effects that they produce. Let us focus on the effects – the forward-looking causal features in Sydney Shoemaker's terms (see e.g. Shoemaker, 2007, chapter 2) – and let us leave open whether it is essential to a functional property to possess a certain causal history, thus simplifying the argumentation. In being causal properties, functional properties are dispositions: they consist in the disposition to produce certain effects. However, it does not follow that they are a subclass of dispositional properties; following Stephen Mumford (1998, chapter 9), one can rather maintain that all dispositional properties are causal-functional ones.

In contemporary philosophy, the causal theory of properties has been developed mainly in opposition to what is known today as Humean

metaphysics (see notably Shoemaker, 1980, for that opposition). According to Humean metaphysics, properties are pure qualities whose essence is independent of the causal relations in which properties of a given type stand in a given world. Their essence thus is purely qualitative, being a primitive suchness, known as *quiddity* (that term has been introduced in the contemporary debate by Black, 2000). Consequently, that essence cannot be known; that consequence is known as *humility* (see e.g. Locke, 2009).

Quidditism is in a certain sense analogous to haecceitism. A haecceitistic difference is a difference between possible worlds which consists only in the fact that there are different individuals in two worlds, without there being any qualitative difference between the worlds in question. In other words, a haecceitistic difference is a difference between individuals which has the consequence that worlds have to be recognized as different, although they are indiscernible. If one maintains that the essence of properties is a primitive suchness (a quiddity), a similar consequence ensues: one is in this case committed to recognizing worlds as different that are identical with respect to all causal and nomological relations, but that differ in the purely qualitative essence of the properties that exist in them. A quidditistic difference thus is a qualitative difference between worlds due to which worlds have to be recognized as different, although they are indiscernible.

Consequently, there always automatically is multiple realizability or the possibility of multiple reference. Functional descriptions of one and the same type can be made true by configurations of intrinsic properties whose primitive suchness is entirely different, as long as the relations among these properties are such that they fulfil the functional descriptions in question against the background of the whole distribution of intrinsic and categorical properties in the world in question. Thus, for instance, the intrinsic properties that make true the fundamental physical description "negative elementary charge" in the real world can make true the fundamental physical description "mass *x*" in another possible world, and *vice versa*. In other words, properties of one and the same type – possessing the same primitive suchness – can in one world fulfil the charge role and in another world the mass role. That is why one cannot deduce the nature of the realizer from the role.

David Lewis, the main proponent of Humean metaphysics in contemporary philosophy, has endorsed the consequences of quidditism and

humility in one of his last papers (published in 2009). However, the discussion in the last decade has made increasingly clear that notably the commitment to quidditism is highly objectionable, in particular for a metaphysics that takes itself to be close to empiricism and inspired by science (see already Black, 2000; for a contrary view see Noonan, 2010): one has to acknowledge an essence of properties that is a pure quality and thus a primitive suchness, being detached from all causal and nomological relations and hence detached from anything of which it is possible to gain knowledge. This consequence is a good motivation for the causal theory of properties that ties the essence of properties to the causal relations in which they stand.

One can sum up the central claim of the latter in the following manner: *in being certain qualities, properties are causal, namely powers to produce certain specific effects*. Take charge for example. This is a fundamental physical property that can occur at space-time points. Charge is a certain quality that is distinct from, for instance, mass. Insofar as charge is a qualitative property, it is a power that manifests itself in certain causal relations, namely the power to generate an electromagnetic field, resulting in the attraction of opposite-charged and the repulsion of like-charged objects.

This view of properties is only coherent on the condition that one considers the qualitative and the causal character of properties as identical. These are not even different aspects of properties, but exactly one and the same. The position of C. B. Martin (1997) and John Heil (2003, chapter 11) is often read as a double aspect theory of properties – properties having a qualitative and a causal aspect – and is thus seen as standing in opposition to the position of Sydney Shoemaker (1980) and Alexander Bird (2007a). However, Heil (2009, 178) says with respect to Martin's last position that he finally conceived properties as "powerful qualities". Against that background, there is no substantial disagreement between the views of Martin and Heil on the one hand and Shoemaker (1980) and Bird (2007a) on the other.

If, by contrast, one interprets Martin and Heil as holding a double aspect theory, there are two obvious objections: what is the relationship between the qualitative and the causal aspect of a property? How can the objection of quidditism be avoided as regards the purely qualitative aspect of properties? There is only one reasonable position in this context, namely the one that conceives properties as being causal in being

certain qualities: properties that are purely causal without being certain qualities would be pure potentialities instead of being real, actual properties (cf. the objection of Armstrong 1999, section 4). And properties that are certain qualities without these qualities being certain causal powers would be quiddities, committing us to recognize worlds that are indiscernible as being qualitatively different nonetheless.

Properties that are causal powers in being certain qualities are dispositions. If this theory applies to all properties, the fundamental physical properties are dispositions as well. This view therefore implies that dispositions do not presuppose a categorical basis. However, if one conceives dispositional properties as being certain actual qualities, it is no problem that there are no underlying categorical bases. One would run into a problem if and only if one conceived dispositions as pure potentialities, which are not actual properties as such, but presuppose a categorical basis (cf. Bird, 2007b, 519–523).

If one regards properties as causal in being certain qualities, one can nevertheless maintain that properties are in a certain sense intrinsic. It is a fact about an object in itself, independently of other objects, that it has certain powers in having certain qualities. This fact is independent of whether the object in question is alone or accompanied by other objects (cf. the definition of intrinsic properties by Langton and Lewis 1998). Charge, for instance, may be an intrinsic and qualitative property inhering in an object and at the same time a causal property, since the qualitative nature of this property consists in generating an electromagnetic field, resulting in the attraction of opposite-charged and the repulsion of like-charged objects.

Even if properties are causal, the relata of causal relations can be objects or events. Causal relations obtain between objects or events in virtue of their properties. That the properties are causal is to say the following: insofar as an object or event has certain properties, it has certain powers. The causal relations that consist in the production of these effects are metaphysically necessary in the following sense: in any possible world in which there are properties of the types in question, there are also causal relations of these types. Thus, in any possible world in which there is charge, charged objects generate an electromagnetic field, resulting in the attraction of opposite-charged and the repulsion of like-charged objects. In short, one can make a sound metaphysical case for all properties being functional properties in the sense of causal prop-

erties that is independent of the debate about the mentioned dilemma of functionalism. This is important in view of the project to apply the causal theory of properties in order to overcome the mentioned dilemma of functionalism.

We have to mention briefly another issue in the metaphysics of properties: *let us conceive properties as particulars (tropes or modes) instead of universals*. I prefer the term "mode" to the term "trope", since it does not suggest a commitment to the bundle view of objects, expressing instead the idea that properties are the ways in which the objects are (cf. Heil, 2003, chapter 13, as well as Strawson, 2008). Again, there are independent arguments available for that conception. Suffice it here to hint at the following two ones: (1) The instantiation relation that is supposed to hold between universals and particulars is unclear. If properties exist as universals beyond the empirical world, it is unclear how what there is in the empirical world can participate in them (see already Plato, *Parmenides*, 130e–133a). If properties exist as universals in the empirical objects, it is not intelligible how numerically one and the same universal can be instantiated in many different objects. (2) As far as the fundamental physical properties are concerned, one can maintain that there is no need to posit universals in order to account for the similarities among them, since these similarities amount to qualitative identities. This qualitative identity is the basis for all further significant, objective similarities in the world.

For instance, all modes of negative elementary charge in the world are exactly the same, that is, qualitatively identical. By the same token, all the modes that are a certain value of rest mass are qualitatively identical. By modes, I always mean determinate properties and not determinable ones – that is, not properties such as, for instance, elementary charge or rest mass, but e.g. negative elementary charge and rest mass 0,51 MeV. All electrons in the world constitute a natural kind, because their characteristic properties – negative elementary charge and rest mass 0,51 MeV – are exactly the same (see Busse, 2008, for an elaborate position in this sense). Fundamental physical properties in the sense of modes hence are numerically distinct, but qualitatively identical. All and only those fundamental physical modes that are qualitatively identical make true the same description (concept, predicate) that expresses what these properties are – such as "negative elementary charge" or "rest mass 0,51 MeV". In sum, there are qualitative identities among funda-

mental physical modes that are sufficient to ground the similarities that there are among more complex properties, including the ones on which the special sciences focus.

3. Conservative identity of properties

Conceiving all properties (1) in a causal-functional way and (2) as particulars (modes) enables us to maintain that the properties in which the special sciences trade exist and are identical with physical properties (see also Whittle, 2008, as regards these two premises and their consequences; see furthermore the position that Gillet, 2007, describes as continuity functionalism). More precisely, they are identical with the manner in which certain microphysical objects are related with one another, producing certain specific effects as a whole under certain environmental conditions due to the manner in which they are arranged; by a configuration of physical properties, we mean such relations among microphysical objects. The reason for this identity is that both are causal-functional: the properties on which the special sciences focus consist in producing certain effects, and these effects are identical with the effects that certain physical configurations bring about as a whole in certain environments. It is therefore not possible to tell them apart on a causal basis. For instance, a certain gene produces certain proteins, and these proteins are identical with the effects that a certain DNA configuration brings about as a whole in a specific molecular context. More precisely, certain atoms are related in a certain way, forming a configuration of nucleic acids of a certain molecular type, and in virtue of the way in which they are related, they produce certain proteins and are therefore identical with a gene of a certain type.

There is no reason to distinguish between role and realizer properties when both are causal-functional and when properties are modes. The way a complex object is qua being a certain biological or mental, functional property token is the same as the way the complex object in question is qua being a certain physical configuration if and only if the effects that the object brings about qua being the former are identical with the effects that the object brings about qua being the latter. We hence get from a theory of causal-functional properties all the way down to a *conservative, ontological functional reduction*: all the prop-

erties that there are in the world are either themselves fundamentally physical, causal-functional properties or are identical with configurations of such properties. This position is a reductionism, since all the properties are physical, but only some are also biological, or mental. It is conservative, since there are biological and mental properties, etc. with all their specific causal features out there in the world (see Esfeld, 2007, for an argument showing that the causal theory of properties does justice to our experience of mental causation).

An argumentation in this vein is often supposed to run into the following problem: if one maintains that the properties of complex objects on which the special sciences focus are identical with physical properties, then there is the danger of this position ending *de facto* up in an eliminativism with respect to these properties of complex objects, because even given the premise of identity, it is not able to show how these properties can be causally efficacious. However, this problem cannot be simply a consequence of the claim of identity: identity is a logical relation that is symmetric. If the causally efficacious properties that complex objects have as a whole are identical with a configuration of fundamental physical properties, then some such configurations are causally efficacious properties of complex objects as a whole. In general, if all *A*s are identical with *B*s, then some *B*s are identical with *A*s. It does not make sense to ask whether a given complex object brings about certain effects qua its physical configuration or qua its properties as a whole, since both are the same. In general, if the property of being *A* is the same as the property of being *B*, then all the effects that an object brings about qua being *A* are the effects that it brings about qua being *B*, and *vice versa*.

One can thus not trace the eliminativism problem back to the identity claim as such. On the contrary, the mentioned problem is a consequence of the theory of properties against the background of which this identity claim is conceived. The claim of identity on its own does not say much. One has to show how the properties of complex objects on which the special sciences focus can be identical with physical properties. That is why the premises of the causal theory of properties and properties as modes or tropes are crucial for the argumentation of this paper. The properties of complex objects with which the special sciences deal are a paradigmatic example of causal properties. The only reason why we admit these properties is that complex objects have certain specific effects as a whole. The claim of these properties being identi-

cal with physical properties is therefore intelligible if and only if those physical properties are causal properties as well, instead of being pure qualities. Otherwise, one would be committed to the consequence that certain configurations of physical properties, which are not causal as such, make true certain causal descriptions of the special sciences, but that there are no causal properties of the objects to which the special sciences refer (see the discussion of Lewis' realizer functionalism above). An eliminativist consequence hence arises if and only if one presupposes a non-causal theory of properties as pure qualities.

The nominalist premise of properties being modes is as important as the premise of the causal theory of properties in order to make this position available. If properties were universals, then the properties of complex objects as a whole on which the special sciences focus could not be identical with physical properties. The reason why there are special sciences that focus on these properties is that the classifications that introduce these sciences in order to seize these properties capture significant similarities that are not expressed by the physical classification of these objects according to their composition. In other words, complex objects that differ in their physical composition can have significant causal properties as a whole in common. If these properties were universals, they could therefore not be identical with physical properties qua universals due to multiple realization.

If, by contrast, properties are not universals, but modes, it is no problem how causal properties that complex objects have as a whole can be identical with configurations of physical properties: the manner in which a complex object is insofar as it is the power to produce certain specific effects as a whole is the manner in which it is insofar as it is a configuration of physical properties. These are two different descriptions (concepts, predicates) of one and the same way (mode, trope) in which a complex object is.

The argumentation according to which one can vindicate the causal efficacy of the properties on which the special sciences focus if and only if one conceives these properties as being identical with physical properties is associated with the works of Jaegwon Kim in the contemporary discussion (see notably the books Kim, 1998; 2005). On the one hand, Kim tends to conceiving all properties in a causal manner (e.g. Kim, 2005, 159), and he rejects the Humean regularity theory of causation, at least as far as mental causation is concerned (Kim, 2007; 2009); on the

other hand, his position comes in the end close to the functionalism of David Lewis: Kim draws the consequence that there are in the last resort no properties that correspond to the descriptions of the special sciences; these descriptions refer to tokens of fundamental physical properties (see Kim, 1998, 111; 2005, 26, 58; as well as 2008, 108–112). Kim adopts Lewis' conception of a local, species-specific reduction (Kim, 1998, in particular 93–95; and, 2005, in particular 25). However, in that manner, Kim's position ends up in the eliminativism horn of the dilemma of the standard versions of functionalism. The position argued for in this paper can be received as a further development of Kim's position, having the aim in view to develop a version of functionalism that avoids the mentioned dilemma by being a conservative reductionism.

Like Kim, John Heil (2003) argues against the conception of there being different ontological levels of properties out there in the world. Heil also maintains that the properties which are supposed to exist on higher levels can only be causally efficacious by being identical with physical properties. Going beyond Kim, Heil argues in favour of a version of the causal theory of properties, and he conceives properties as modes. The position developed in this paper therefore is close to the one of Heil. However, my main criticism of Heil's ontology is that he does not develop the consequences that his premises would allow him to draw: his view finally amounts to the conclusion that the functional properties of the special sciences do not exist, there being only the fundamental physical properties (see Heil, 2003, 45, 153, and 2006, 18–21, for a clear and concise statement, as well as Esfeld, 2006, on this consequence of Heil's position). Again, the threat of the dilemma between eliminativism and epiphenomenalism is obvious.

The causal theory of properties is associated in particular with Sydney Shoemaker's paper "Causality and properties" (Shoemaker, 1980). In later work, Shoemaker argues for a view according to which the properties on which the special sciences focus are realized by physical properties without being identical with them: in brief, the powers that characterize a property of the special sciences are a subset of the powers that characterize the respective physical realizer properties (see Shoemaker, 2007, chapter 2, especially 11–14). Following Yablo (1992), Shoemaker draws a distinction between determinable and determinate properties and takes this distinction to illustrate that view: the powers that characterize a determinable property are a subset of the powers

that characterize the respective determinate realizer properties. Thus, the causal powers that characterize the property of being blue are a subset of the causal powers that characterize the property of being marine blue: marine blue has all the causal powers of blue and further powers, namely those ones that distinguish marine blue from, for instance, cobalt blue.

However, by making that ontological distinction between properties, Shoemaker runs into the same problem as the role functionalism of Putnam and Fodor. As Carl Gillet and Bradley Rives (2005, section 3) point out, the determinate or realizer properties include by definition all the powers of the respective determinables. Consequently, the determinate properties are sufficient to bring about all the effects that the determinables could cause. Unless one acknowledges token identity between realized and realizer properties, one thus faces again the epiphenomenalism objection that haunts role functionalism. That objection could only be avoided by either admitting systematic overdetermination or by accepting interactionism, thus rejecting the causal completeness of the physical domain (McLaughlin, 2007, interprets Shoemaker's position in that latter way; see also the emergentism that Gillet, 2006, contemplates).

It can with reason be maintained that the differentiation between determinables and determinates is not an ontological one between properties that there are in the world, but concerns only concepts and descriptions. The predicates that we use in order to describe the properties in the world can be determinable or determinate, such as the predicates "blue" and "marine blue". The properties that there are in the world are all determinate ones (see Gillet and Rives, 2005). Being blue or being marine blue are in no manner different properties that there are in the world. One and the same property in the sense of a way (mode) in which an object is can be described in a precise manner by using the predicate "marine blue" and in a less precise manner by using the predicate "blue".

Shoemaker's claim that a physical property that realizes a property in the domain of a special science brings about the effects that characterize the latter property only in virtue of a subset of its causal powers is disputable, if one bears in mind that what is causally efficacious are property tokens, not types. The properties of the special sciences cannot be realized by single physical properties, but only by configurations

of physical properties. Any property token of the special sciences can cause the effects that characterize the property type in question in the vocabulary of a special science only by bringing about the effects that a certain configuration of physical properties produces as a whole. For instance, any gene token can produce the specific protein that it brings about in a certain situation only by having all the molecular effects that a certain DNA configuration has as a whole in that situation, for it is through those effects that the protein comes into being. To take another example, any pain token can cause the specific pain behaviour that it brings about in a certain situation only by producing the neuronal effects that a certain configuration of neurons has as a whole in that situation because it is through those effects that the pain behaviour comes about. The properties on which the special sciences focus hence are in the same manner determinate properties as are the physical properties. It is only that their descriptions in the vocabulary of a special science are not as detailed as physical descriptions.

This conclusion could be avoided by admitting token multiple realization, that is, maintaining that, for instance, one and the same pain *token* can be realized by a neuronal configuration of another type in another possible world (Yablo, 1992, endorses token multiple realization). However, if one subscribes to this idea, one is committed to the position that each token of a functional property possesses a primitive thisness, since its being is independent of the physical configuration that realizes it in a given world. In this case, there are two property tokens a and b of the same property type in the domain of a special science, in the world w_1, a is realized by a physical configuration of type P_1, and b is realized by a physical configuration of type P_2. In the world w_2, by contrast, it is a that is realized by a physical configuration of type P_2, and it is b that is realized by a physical configuration of type P_1. The only difference between these two worlds is a swap of a and b. One thus is committed to haecceitism: worlds are recognized as being different whose only difference consists in a swap of individuals, without there being any qualitative difference between them. Haecceitism is widely considered to be an implausible position for this reason (see notably Lewis, 1986b, chapter 4.4, and 2009, section 4).

This short discussion shows that newer version of functionalism that are proposed in the contemporary literature also run into the dilemma of epiphenomenalism and eliminativism, which the classical versions of

functionalism of Putnam and Fodor on the one hand and Lewis on the other face. The way out of this dilemma consists in proposing a causal theory of properties together with a theory of properties as modes and to base on this view of properties the claim of the properties on which the special sciences focus being identical with configurations of physical properties.

4. Ontological and epistemological reductionism

Token identity in the sense of an identity of properties as tropes or modes as a proposal to resolve the problem of the causal efficacy of the properties on which the special sciences focus is not new. Notably David Robb (1997) has argued for token identity on the basis of a metaphysics of properties as tropes. What the present proposal adds to Robb is a causal-functional view of all properties, supported by independent arguments. Nonetheless, one may object that this move is not sufficient to solve the mentioned problem: the metaphysics of properties as particulars (modes or tropes) is employed in this context because there is no type identity between types of the special sciences – such as biological or psychological types – and physical types due to multiple realization. One can in this context object the following: insofar as there is only a reduction of tokens, but not of types, the problem remains whether the properties on which the special sciences focus cause anything qua biological, or mental properties, etc.

Let us briefly consider the background of that problem: Donald Davidson (1970) claims in his famous paper "Mental events" that mental events are identical with physical events. More precisely, all events admit a physical description, and some events admit also a mental description. It is not possible to reduce the mental to a physical description. This position is widely recognized to fail due to the following objection: it cannot show that events cause anything insofar as they are mental events (see the papers in Heil and Mele, 1993).

The position put forward in this paper differs from the one of Davidson in that identity applies to property tokens in the sense of modes or tropes. Nonetheless, Paul Noordhof (1998) objects to Robb (1997) that in the same way as it is reasonable to ask whether a Davidsonian mental event causes anything qua mental, it is reasonable to ask whether

a mental trope causes anything qua being a trope of a mental type (see, as regards this objection, also Kistler, 2009, chapter 5.2). Robb (2001) retorts that if identity is applied to those entities in virtue of which an object or event causes something, namely properties in the sense of tropes, it makes no sense to raise the qua-question for these entities, since they are already the most fine-grained ones (cf. also Whittle, 2007, section 4).

Even though that reply is correct, there remains a problem. If all that exists in the world are particulars (objects and their modes), then types are concepts that seize salient similarities among the ways objects are (natural kinds). As regards the mentioned ways in which complex physical objects exist, these modes make true descriptions in terms of physical concepts that focus on their composition as well as descriptions in terms of concepts of the special sciences that focus on the salient effects that they bring about as a whole in a given environment. Multiple realization is the epistemological fact that modes coming under one single concept of the special sciences often come under different physical concepts. The concepts of the special sciences and the corresponding physical concepts differ not only in meaning, but they are also not coextensive.

On the one hand, not only the physical concepts, but also the concepts of the special sciences possess a scientific quality, consisting in these concepts figuring in law-like generalizations that are projectible, support counterfactuals and provide causal explanations. On the other hand, not only Davidson in his "Mental events", but also most of the philosophers who favour token identity in the fine-grained sense of identity of properties as particulars (modes, tropes) maintain that the descriptions (laws, theories) in which concepts that are proper to the special sciences figure cannot be reduced to physical descriptions (laws, theories) (or remain at least neutral with respect to that latter point). In other words, they defend ontological reductionism combined with an epistemological anti-reductionism (or at least combined with neutrality as regards epistemological reductionism).

However, in that case, the problem that haunts Davidson and that Noordhof raises against Robb reappears: it has to be possible to relate the different descriptions in a systematic, reductive manner, if they are descriptions that are made true by one and the same way an object is and if each of them is to provide for law-like generalizations that are projectible, support counterfactuals and yield causal explanations. Oth-

erwise, it could not be vindicated that these descriptions are about the *same* entities in the fine-grained sense of modes instead of being about *different* properties that objects have, that is, different ways in which they are. Consequently, the position would end up either in property dualism with the threat that the ways objects are insofar as they make true descriptions of the special sciences are epiphenomenal, or in eliminativism as regards the scientific quality of the descriptions in terms of the special sciences.

The argumentation set out in this paper hence is not complete as yet. The ontological reductionism proposed in this paper stands or falls together with an epistemological reductionism. Multiple realization prevents the types in which the special sciences trade from being identical with physical types. Nonetheless, one has to achieve some sort of type reduction, that is, provide for a way that enables in principle the reduction of the descriptions (laws, theories) of the special sciences to physical descriptions (laws, theories) in order to retain the scientific quality of the former ones. Christian Sachse shows in his contribution to this volume how this can be done against the background of the causal theory of properties. Thus, if the argumentation in this paper and the following one is on the right track, the causal theory of properties opens up the way both for a conservative ontological reductionism and a conservative epistemological reductionism.

Note

1 The claims made in this paper are elaborated on in a more detailed manner in Esfeld and Sachse, 2010, chapter 2. I am grateful to an anonymous referee for criticism of the draft of this paper.

Bibliography

Armstrong, David M., 1999: The causal theory of properties: properties according to Shoemaker, Ellis, and others. In: *Philosophical Topics* 26, pp. 25–37.

Bennett, Karen, 2003: Why the exclusion problem seems intractable, and how, just maybe, to tract it. In: *Noûs* 37, pp. 471–497.

Bickle, John, 1998: *Psychoneural reduction: the new wave*. Cambridge, MA: MIT Press.

Bird, Alexander, 2007a: *Nature's metaphysics. Laws and properties*. Oxford: Oxford University Press.

Bird, Alexander, 2007b: The regress of pure powers? In: *Philosophical Quarterly* 57, pp. 513–534.

Black, Robert, 2000: Against quidditism. In: *Australasian Journal of Philosophy* 78, pp. 87–104.

Block, Ned, 1990: Can the mind change the world? In: Boolos, George (ed.): *Meaning and method. Essays in honor of Hilary Putnam*. Cambridge: Cambridge University Press, pp. 137–170.

Busse, Ralf, 2008: Fundamentale Eigenschaften und die Grundlagen des Ähnlichkeitsnominalismus. In: *Philosophia Naturalis* 45, pp. 167–210.

Davidson, Donald, 1970: Mental events. In: Foster, L.; Swanson, J. W. (eds.): *Experience and theory*. Amherst: University of Massachusetts Press, pp. 79–101. Reprinted in Davidson, Donald, 1980: *Essays on actions and events*. Oxford: Oxford University Press, pp. 207–225.

Esfeld, Michael, 2006: From being ontologically serious to serious ontology. In: Esfeld, Michael (ed.): *John Heil. Symposium on his ontological point of view*. Frankfurt (Main): Ontos, pp. 191–206.

Esfeld, Michael, 2007: Mental causation and the metaphysics of causation. In: *Erkenntnis* 67, pp. 207–220.

Esfeld, Michael, 2010: Causal overdetermination for Humeans? In: *Metaphysica* 10, DOI: 10.1007/s12133-010-0061-3.

Esfeld, Michael; Sachse, Christian, 2010: *Kausale Strukturen. Einheit und Vielfalt in der Natur und den Naturwissenschaften*. Berlin: Suhrkamp. English version *Conservative reductionism*. Forthcoming New York: Routledge spring 2011.

Fodor, Jerry A., 1974: Special sciences (or: The disunity of science as a working hypothesis). In: *Synthese* 28, pp. 97–115.

Gillet, Carl; Rives, Bradley, 2005: The non-existence of determinables: or, a world of absolute determinates as default hypothesis. In: *Noûs* 39, pp. 483–504.

Gillet, Carl, 2006: Samuel Alexander's emergentism: or, higher causation for physicalists. In: *Synthese* 153, pp. 261–296.

Gillet, Carl, 2007: A mechanist manifesto for the philosophy of mind:

a third way for functionalists. In: *Journal of Philosophical Research* 32, pp. 21–42.
Heil, John, 2003: *From an ontological point of view*. Oxford: Oxford University Press.
Heil, John, 2006: On being ontologically serious. In: Esfeld, Michael (ed.): *John Heil. Symposium on his ontological point of view*. Frankfurt (Main): Ontos, pp. 15–27.
Heil, John, 2009: Obituary. C. B. Martin. In: *Australasian Journal of Philosophy* 87, pp. 177–179.
Heil, John; Mele, Alfred (eds.), 1993: *Mental causation*. Oxford: Oxford University Press.
Kim, Jaegwon, 1998: *Mind in a physical world. An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
Kim, Jaegwon, 2005: *Physicalism, or something near enough*. Princeton: Princeton University Press.
Kim, Jaegwon, 2007: Causation and mental causation. In: McLaughlin, Brian P.; Cohen, Jonathan (eds.): *Contemporary debates in philosophy of mind*. Oxford: Blackwell, pp. 227–242.
Kim, Jaegwon, 2008: Reduction and reductive explanation: is one possible without the other? In: Hohwy, Jacob and Kallestrup, Jesper (eds.): *Being reduced*. Oxford: Oxford University Press, pp. 93–114.
Kim, Jaegwon, 2009: Mental causation. In: McLaughlin, Brian, Beckermann, Ansgar and Walter, Sven (eds.): *The Oxford handbook of philosophy of mind*. Oxford: Oxford University Press, pp. 29–52.
Kistler, Max, 2009: *La cognition entre réduction et émergence. Etude sur les niveaux de réalité*. Paris: Syllepse.
Langton, Rae; Lewis, David, 1998: Defining 'intrinsic'. In: *Philosophy and Phenomenological Research* 58, pp. 333–345. Reprinted in Lewis, David, 1999: *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press, pp. 116–132.
Lewis, David, 1980: Mad pain and Martian pain. In: Block, Ned (ed.): *Readings in the philosophy of psychology. Volume 1*. London: Methuen, pp. 216–222. Reprinted in Lewis, David (1983): *Philosophical papers. Volume 1*. Oxford: Oxford University Press, pp. 122–130.
Lewis, David, 1986a: *Philosophical papers. Volume 2*. Oxford: Oxford University Press.
Lewis, David, 1986b: *On the plurality of worlds*. Oxford: Blackwell.

Lewis, David, 1994: Lewis, David: Reduction of mind. In: Guttenplan, Samuel H. (ed.): *A companion to the philosophy of mind*. Oxford: Blackwell, pp. 412–431.

Lewis, David, 2009: Ramseyan humility. In: Braddon-Mitchell, David; Nola, Robert (eds.): *Conceptual analysis and philosophical naturalism*. Cambridge (Massachusetts): MIT Press, pp. 203–222.

Locke, Dustin, 2009: A partial defense of Ramseyan humility. In: Braddon-Mitchell, David; Nola, Robert (eds.): *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press, pp. 223–241.

Loewer, Barry, 2007: Mental causation, or something near enough. In: McLaughlin, Brian P.; Cohen, Jonathan (eds.): *Contemporary debates in philosophy of mind*. Oxford: Blackwell, pp. 243–264.

Martin, C. B., 1997: On the need for properties: the road to Pythagoreanism and back. In: *Synthese* 112, pp. 193–231.

McLaughlin, Brian P., 2007: Mental causation and Shoemaker-realization. In: *Erkenntnis* 67, pp. 149–172.

Mumford, Stephen, 1998: *Dispositions*. Oxford: Oxford University Press.

Noonan, Harold W., 2010: Bird against the Humeans. In: *Ratio* 23, pp. 73–86.

Noordhof, Paul, 1998: Do tropes resolve the problem of mental causation? In: *Philosophical Quarterly* 48, pp. 221–226.

Putnam, Hilary, 1967 / 1975: The nature of mental states. In: Putnam, Hilary, 1975: *Mind, language and reality. Philosophical papers. Volume 2*. Cambridge: Cambridge University Press, pp. 429–440. First published as Psychological predicates. In: Capitan, W. H.; Merrill, D. D. (eds.), 1967: *Art, mind and religion*. Pittsburgh: University of Pittsburgh Press.

Robb, David, 1997: The properties of mental causation. In: *Philosophical Quarterly* 47, pp. 178–194.

Robb, David, 2001: Reply to Noordhof on mental causation. In: *Philosophical Quarterly* 51, pp. 90–94.

Shoemaker, Sydney, 1980: Causality and properties. In: van Inwagen, Peter (ed.): *Time and cause*. Dordrecht: Reidel, pp. 109–135. Reprinted in Shoemaker, Sydney, 1984: *Identity, cause, and mind. Philosophical essays*. Cambridge: Cambridge University Press, pp. 206–233.

Shoemaker, Sydney, 2007: *Physical realization*. Oxford: Oxford University Press.

Strawson, Galen, 2008: The identity of the categorical and the dispositional. In: *Analysis* 68, pp. 271–282.
Whittle, Ann, 2007: The co-instantiation thesis. In: *Australasian Journal of Philosophy* 85, pp. 61–79.
Whittle, Ann, 2008: A functionalist theory of properties. In: *Philosophy and Phenomenological Research* 77, pp. 59–82.
Yablo, Stephen, 1992: Mental causation. In: *Philosophical Review* 101, pp. 245–280.

Christian Sachse

Conservative Reduction of Biology

Abstract*

The paper argues in favour of a reductionist strategy in the philosophy of biology in order to maintain the unity of science. After considering the problems in the current state of the philosophy of science that beset both the anti-reductionist and reductionist camps, I design a conservative, non-eliminativist, alternative reductionist strategy based on the theoretical possibility of constructing functionally defined sub-concepts in biology that are nomologically coextensive with physical descriptions. This *theoretical* link between biology and physics makes it possible to understand the original and operational biological concepts as abstractions from these biological sub-concepts. Thus, in a sense, we can 'serve two masters', preserving the fundamental role of physics while allowing biology its proper sphere of explanatory autonomy, and thus its scientific character. Since this abstraction step is an intra-theoretic one, the scientific quality of the original biological concepts can be vindicated because both epiphenomenalism and eliminativism are avoided, a result that is not available in standard reductionist or anti-reductionist approaches. Against this background, we can argue furthermore that biological kinds are natural ones and that biology adumbrates laws and explanations of different degrees of abstraction.

Zusammenfassung

Um die Einheit der verschiedenen Wissenschaften zu erreichen, wird eine neue reduktionistische Strategie im Bereich der Philosophie der Biologie vorgestellt. Dies geschieht im Anschluss an eine Untersuchung der Probleme, denen bisherige nicht-reduktionistische und reduktionistische Strategien gegenwärtig in der Wissenschaftsphilosophie gegenüberstehen. Aufgrund dieser Analyse wird eine sogenannte konservative, das heißt nicht-eliminativistische, alternative reduktionistische Strategie vorgestellt, welche sich im Wesentlichen auf die theoretische Möglichkeit der Konstruktion funktional definierter Subtypen im Bereich der Biologie stützt, welche nomologisch koextensional mit physikalischen Beschreibungen sind. Diese *theoretische* Verbindung zwischen Biologie und Physik ermöglicht es, die ursprünglichen operationalen biologischen Begriffe als Abstraktionen der konstruierten Subtypen zu verstehen. Hierdurch werden wir zwei Dingen gerecht – sowohl die

philosophia naturalis 47-48 / 2010-11 / 1-2

fundamentale Rolle der Physik beizubehalten, als auch einen eigenen Bereich von Erklärungsunabhängigkeit für die Biologie zu begründen und dadurch ihre wissenschaftliche Qualität herausstellen. Da der erwähnte Abstraktionsschritt rein theorieimmanent ist, kann die wissenschaftliche Qualität der ursprünglichen biologischen Begriffe begründet werden, ohne zu einem Epiphenomenalismus oder Eliminativismus zu führen – ein Resultat, das in herkömmlichen reduktionistischen oder anti-reduktionistischen Strategien nicht möglich ist. Vor diesem Hintergrund kann weiterhin argumentiert werden, dass biologische Arten natürliche Arten sind und dass die Biologie zur Formulierung von Gesetzen und Erklärungen unterschiedlicher Abstraktionsgrade fähig ist.

1. Introduction and the dilemmas for the scientific quality of biology

It is generally taken for granted that biology provides functional explanations. These are specific causal explanations often couched in terms of selected effects or given contributions to certain capacities or fitness functions of biological systems (the locus classicus is Wright, 1973; and Cummins, 1975). We shall say more about biological functions in section 5. Here, let us simply take for granted that functional explanations are somehow about causally efficacious properties. This is sufficient to ground the claim that each functional property token is identical with something physical (see among others Kim, 2005 ch. 2; Kitcher, 1984; and Rosenberg & Kaplan, 2005). Otherwise there would be serious problems with at least one of the following widely accepted working hypotheses: 1) biological properties supervene on complex configurations of physical properties (see among others Rosenberg, 1978; Weber 1996) and 2) physics is causally, nomologically and explanatorily complete with respect to biology (see among others Papineau, 2002, appendix).

Biology, for instance, seeks to explain causal relations between given genotypes, phenotypes and the environment or evolutionary changes of the genotype frequencies due to different selective forces in changing environments. Both such local causal relations and more extensive evolutionary processes, however, supervene on complex configurations of physical properties and their changes. These, in turn are *in theory* completely explicable in terms of physics. More precisely, under the condition that biological properties have functional effects, the production of

these effects supervenes on a corresponding causal interaction among physical properties, where no non-physical forces are involved. This fact suggests that any functional property token is identical with something physical (see especially Kim, 2005, ch. 2 for the general argument in detail). Biological property tokens therefore are not epiphenomena, but causally efficacious entities, and biological functional explanations can be understood as causal ones that focus on specific causal effects of these entities (according to the very notion of a biological function).

Against the background of this general framework, we come to the conclusion that biology refers in different terms to property tokens to which physics also refers. One may then ask how biological concepts, laws and explanations are related to the corresponding physical descriptions (that are composed concepts), laws and explanations. Of course, the answer depends on what we mean by 'laws' and there is an ongoing debate about whether there are any laws in biology and whether biological explanations presuppose underlying laws (see Rosenberg, 2001). The reductionist approach I propose constitutes an argument that biology does contain laws, and that biological explanations thus have a nomological underpinning (see also Sober, 2000, ch. 1.4). It furthermore enables us to take biological kinds as genuine natural ones. However, these issues presuppose at first a more detailed consideration of biological concepts and their relation to physics.

To launch this discussion, let's start with the main shortcoming of the anti-reductionist position (I'm thinking here of arguments going back to Fodor, 1974; and Putnam 1967/1975, like for instance in Kitcher, 1984). According to these arguments, biological concepts may and are often not bi-conditionally related to physical descriptions. Functional explanations that are couched in such terms may then constitute some kind of autonomous and unifying explanatory level. For instance, *one* genetic description and explanation may refer to entities that come in fact under *different* physical descriptions and explanations. Pace multiple realization, this fact should rather constitute an anti-*eliminativist* than an anti-*reductionist* argument, and a central focus of the paper is to spell this claim out. Why? Actually, in any case where a biological concept cannot be bi-conditionally related to a physical description, more precisely, if no nomological coextension between both descriptions can be established, then no *anti*-reductionist position has an argument to take these biological concepts as being about the same property tokens

as physics is. Within an anti-reductionist framework, it is moreover suggested that the biological concepts are actually not about the *same* entities in the fine-grained sense, but are instead about *different* properties (see Michael Esfeld's contribution to this volume; see also Esfeld & Sachse, 2010, ch. 5; and Sachse, 2007, ch. 3).

This consequence has direct relevance to the usual pattern of argument by which anti-reductionism maintains the autonomy and scientific quality of biology. If one sticks to the mentioned multiple realization as an anti-reductionist argument, claiming that bi-conditional relations cannot be sufficiently established to enable the theoretical reduction of biology, this then leads to property dualism, with the further implication that the biological properties are epiphenomenal. If we go to the logical end of the anti-reductionist argument, we encounter the paradox that anti-reductionism makes rather eliminativism plausible than the claimed autonomy of biology. In order to avoid epiphenomenalism, it thus *has to be theoretically possible* to construct biological concepts that are bi-conditionally related to physical descriptions, which then means to consider the relation between biology and physics from a reductionist perspective.

However, the fact that anti-reductionism suggests epiphenomenalism and therefore also eliminativism does not imply that *non-eliminativist* (conservative) reduction is a trivial task. Rather, reductionism faces two dilemmas. The first dilemma derives from the fact that a reductionist approach, which is classically based on bi-conditional relations between biological concepts and physical descriptions (like Nagel, 1961), seems to lead to another kind of eliminativism. To understand how this consequence comes about, just imagine for one moment that there were no multiple realization such that one could establish Nagelian bridge-laws in a bi-conditional manner, or, if you prefer, in terms of Kim style functional reduction (Kim, 1998, ch. 4; and 2005, ch. 4). Thus, any biological concept *B* would be nomologically coextensive with a physical description *P*. But even though we may then vindicate the claim that biological concepts are about causally efficacious properties (more on that in section 2), the causal, nomological and explanatory completeness of physics is such that ultimately, we can in principle eliminate biology in favour of physics, even though not in practice. In other words, if nomological coextension can be established between *all* biological concepts and physical descriptions, physics would theoretically replace biology.

The second dilemma can be spelled out as follows: by using the multiple realization thesis to refute classical reductionism (more on that in section 3), it is nonetheless possible to develop another kind of reductionism that also ends up implying eliminativism. It turns out that it is theoretically possible to construct physical theories in such a way that they model in a more or less isomorphic way the complete set of biological theories (see Bickle, 1998). This possibility arises in the following stages: a) due to token-identity and the completeness of physics, one may theoretically construct physical theories that model any causal relation considered by biological theories, even though such a modelling may not constitute bi-conditional relations because of multiple realization; b) due to their integration into physics, the constructed physical theories are the preferred ones, leaving no logical space for any sui generis biological theories (that seem to be disqualified anyway so far). This eliminativist result does not change if the reductionist approach is some kind of combination of the classical or functional model of reduction and "new wave reductionism", or if the "new wave" collapses into the classical model (see Endicott, 1998, for that collapse). To sum up, it seems that both the anti-reductionist and the reductionist approaches look somehow alike by suggesting the theoretical replacement of biology by physics or constructed physical theories.

De facto, one may currently not replace biology for simple instrumental reasons and also estimate that there is little likelihood to do so one day. However, instrumental reasons become stronger if one can argue how exactly biology can be *about causally efficacious property tokens* (I shall refer to this point by "*Cau*" in what follows) and at the same time show why its functional explanations are *theoretically not replaceable* by physical ones (I shall refer to this point by "¬*Rep*" in what follows). My discussion, then, follows the argument that can be made to conjoin "*Cau*" and "¬*Rep*". I shall come back to instrumental reasons at a later stage. At this point, to recap, it is obvious that "*Cau*" and "¬*Rep*" are two sides of the same coin called multiple realization in the mainstream approaches. We have shown that the antireductionist approaches cannot establish "*Cau*" and therefore hardly "¬*Rep*". If there were no multiple realization and nomological coextension thus could be established between all biological concepts and physical descriptions, then classical and functional models of reductionism may vindicate "*Cau*" but not "¬*Rep*". If there is multiple realization, then new wave reduc-

tionism enters the scene, however, without justifying the conjunction of "*Cau*" and "¬*Rep*" either.

The following section considers the possibility to establish nomological coextension in the framework of functional reduction under the condition that there is no multiple realization. In this context, it is possible to vindicate "*Cau*" for biological concepts. In section 3, I show how to establish nomological coextension starting with multiple realization. By means of a causal argument, I will establish nomological coextension between so-called functional sub-concepts and physical descriptions. The end result is once again the vindication of "*Cau*", here for the constructed sub-concepts. Section 4 will then focus on the relationship between these sub-concepts and the original (multiply realized) biological concepts. The aim of this section is to show how the original biological concepts may inherit "*Cau*" from their sub-concepts even though being non-replaceable by physics ("¬*Rep*") due to multiple realization. The last section will then apply these considerations to several other debates in the philosophy of biology – among others the notion of biological function, laws in biology and biological natural kinds.

2. Functional reduction without multiple realization

The structure of this section is threefold. First of all, I need to represent biology for the previously raised issues and aims of the paper. This will be done by some comments on biological functions and explanations. This, second, enables us to apply the model of functional reduction and reductive explanation, which in turn, third, shows how one can vindicate that biology is about causally efficacious property tokens ("*Cau*") if there is no multiple realization.

By no means, a single paper may represent the entire domain of biological inquiry. However, for a consideration of reductionism, the following two issues may sufficiently represent biology: 1) any biological concept can be functionally defined in order to provide special kinds of causal explanations. Thereby, biological concepts are about functions being defined in the context of evolutionary biology since, as Dobzhansky put it, "nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1973). While the so-called etiological approach refers to the evolutionary *past* in order to determine what biological functions

are, other approaches determine biological functions in terms of causal dispositions to contribute to the system's capacities or fitness under the *given* environmental conditions (see Wright, 1973; and Cummins 1975; see among others also Bigelow & Pargetter 1987, Millikan, 1989; Neander, 1991; Griffiths, 1993; Kitcher, 1993; Mitchell, 1993; Godfrey-Smith, 1993 and 1994, Amundson & Lauder 1994; Manning, 1997; Schwartz, 1999; Wouters, 2003; Arp, 2007; and Mossio et al., 2009). Here, nothing hangs on a particular concept of biological functions. However, section 3 will be an implicit critique of the etiological approach, and section 4 constitutes in turn an implicit strategy to spell out its compatibility with the causal-dispositional approach (that will be made more explicit in section 5).

2) Any functionally defined concept of biology refers to something physical (token-identity). Within biology, any entity is most completely described and explained in terms of molecular biology. This point pays heed to another all encompassing paradigm of biology – the molecular approach to any biological property is a mechanistic one (cf. Machamer, Darden and Craver, 2000). Once again, nothing special hangs on a particular interpretation of this paradigm or research strategy since both reductionists and anti-reductionists accept some pre-eminent importance of molecular biology. This paper shall only focus on how to establish a reductionist link between biology and physics. Since the arguments are constructed on a high level of abstraction, one may easily apply them to the debate about whether or not particular branches of biology are reducible to molecular biology as well.

Against this background, we now consider how to justify "*Cau*" – that biology is about causally efficacious property tokens. As explained in the introduction, one has to relate the physical and biological concepts, laws and explanations in a bi-conditional manner. Otherwise, we could not say that biological concepts, laws and explanations are about the *same* entities as physics instead of being about *different* ones, and keeping the completeness of the physical world in mind, *epiphenomenal* ones. If there is no multiple realization, nomological coextension between physical and biological concepts, laws and explanations can be established by the model of functional reduction. To see this, let us follow, with a few tweaks, the central steps of the functional reduction model that reconsiders previous issues in a more precise way (Kim, 1998, ch. 4 and 2005, ch. 4):

1. We functionally define any concept *B* of biology. Independently of the very notion of biological functions at this point, the general argument for this approach can be spelled out as follows: if biological concepts are not functionally formulated in causal terms, then no causal explanation could be based on them. Depending on the very notion of biological functions, the functional formulation spells out what the characteristic causes and effects are. In this sense, the difference between physical and biological definitions of biological properties is not only of terminological nature, but there is a difference in what is understood as characteristic. Keeping in mind the evolutionary context, the difference between the biological and physical ways to describe and explain biological properties becomes evident: the biological definition relates causes and effects in the most general way to the framework of natural selection, which is what endows it with *functionality.* These functional definitions are more abstract than those of physics, where the focus is on *any* causal power the property exhibits.

2. We look for the physical base, often called physical realization, of biological functions. For a lot of biological property tokens, their exact physical structure is of course an open empirical question. But this empirical fact concerning current research does not alter the metaphysical issue. Since we take ontological reductionism for granted, there is no question *whether* biological property tokens are identical with *something* physical. Additionally, the reductionist debate is not primarily concerned with instrumental applicability and thus does not imply any normative component like claiming that reduction *should be* effected in biological research projects. Reduction is not a pragmatic matter, but is instead generally concerned with creating a coherent and unified system of scientific theories and possible levels of explanations. Furthermore, in this paper, I shall argue in section 4 that there are good scientific reasons for retaining a biological vernacular in order to argue against its theoretical replacement by physics ("$\neg$*Rep*").

3. Given the first two steps, it is then in theory possible to explain reductively, which means in terms of physics, how biological property tokens are caused and cause the effects that characterize their functioning (see for the general idea of reductive explanations Chalmers, 1996, pp. 42–51). The main characteristic of any kind of *reductive* explanation is its relative nature – to explain something (e.g. biological) in different and more detailed (e.g. physical) terms. As taken for granted in

step 1, biology provides functional explanations. Since any of these causal explanations refers to something physical (step 2) and physics is causally, nomologically and explanatorily more complete than biology, physics can provide more detailed causal explanations of the biological causal relations (that are outlined in any functionally defined concept *B* and an according biological functional explanation). These reductive explanations are commonly seen as mechanistic explanations (see also Craver, 2001 and 2006). One may note that in fact such reductive mechanistic explanations in terms of physics are quite common in biology and, similarly, molecular biology may provide on its own reductive (mechanistic) explanations as concerns the properties and causal relations considered by any other biological branch. Let me shortly illustrate these three steps by means of some biological examples.

Step 1: *Escherichia coli*, a bacterium that is often used in genetic research, contains genes or regions in the genome that are responsible for its cell-wall biosynthesis. To simplify, let us focus on genes that code for membrane proteins. Since the synthesis of these proteins are required for the growth of the cell before cell division, it is accordingly possible to functionally define the genetic bases: the rate of protein synthesis means a contribution to the possible growth rate, which can be, under optimal growth conditions, equated with fitness (see also Waters, 1994 and 2007 for the discussion of the gene concept).

Step 2: Researchers have mapped the physical structure of the genetic base being responsible for the cell wall. For instance, they have sequenced the *mrdA* (Murein cluster d) that plays a crucial role for the cell-wall biosynthesis. This means in the ideal case that one has identified the physical structure that, given certain physical conditions in the cells, will bring about the expression of the proteins (or other effects) from the genes or genetic regions in question. To sum up and simplify the issue, the DNA sequences within the *E. coli* genome are physically identified as the coding sequence for these specific proteins, with these specific functions, produced under normal physical conditions in the cell.

Step 3: The description of any such gene and the explanation of the production of its characteristic effects may theoretically use only physical concepts. Because of the completeness of physics, it may employ only concepts of physics, explaining in a mechanistic manner the causal relation from the gene to its phenotypic effect. Simplified, it is a physi-

cal mechanistic explanation of how DNA sequences are transcribed into mRNA, which then is translated into chains of amino acids that, in turn, are folded into proteins that are incorporated in the cell wall, bringing finally about the phenotypical effects that define the gene in question.

Under our hypothesis that excludes multiple realization, these steps of the functional reduction model show how each biological description and explanation of a property can in principle be nomologically correlated with a corresponding physical description and explanation. To put it differently, if the biological concepts are functionally (and thus causally) defined, then the biological and the corresponding physical descriptions are nomologically coextensive because: a) our assumption that there is no multiple realization; b) ontological reductionism; and c) the completeness of physics. Once the nomological coextension between biological concepts and physical descriptions is established, it is possible to deduce biological explanations and laws from physics. There thus remains no threat of epiphenomenalism for biological properties ("*Cau*" is justified). Both biology and physics are, then, about the same properties even though they refer to these properties in different terms. In this way, the biological approach has a higher degree of abstraction since it does not spell out in detail the way in which the characteristic effects of genes are produced.

To sum up, if there is no multiple realization, functional reduction and reductive explanations establish an argument for the scientific quality of biological concepts, laws and explanations *in the sense that* they are about certain specific causal powers and causal relations ("*Cau*"). However, keep in mind that an eliminativist approach to biology can still be based on the preference of the physical reductive explanations because of their completeness. "*ffiRep*" of the biological explanations is not yet justified.

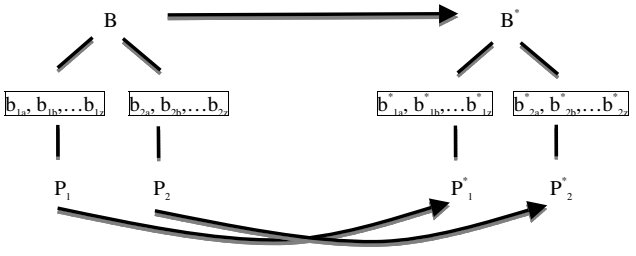
3. The challenge of multiple realization

The aim of this section is twofold. First of all, I shall make the representation of biology more realistic by incorporating the fact that its properties are often multiply realized or at least multiply realizable. To put it differently, biological concepts often refer or may refer to property

tokens in a homogeneous manner that are heterogeneously described in terms of physics. The question then is how to justify "*Cau*" for the functional similarities that are brought out by biological concepts – that they are about causally efficacious property tokens even though no correspondingly homogenous physical description exists. As we pointed out above, "*Cau*" depends on the theoretical possibility of establishing nomological coextension between biological concepts and physical descriptions. Since it follows from the definition of multiple realization that the original biological concepts cannot figure in such bi-conditional correlations, the second aim of this section is thus to outline an argument to theoretically construct so-called functionally defined biological sub-concepts that are no longer multiple realizable. On this basis, "*Cau*" can be justified as concerns these sub-concepts.

Taking ontological reductionism and the completeness of physics for granted, physics can account for any property token that biology describes and explains. Hence, there are biological property tokens that are described and explained in physical terms and in terms of biology. The characteristic difference between these two kinds of descriptions (and thus explanations) lies in the referring to the very same properties by allowing for the fact that biology focuses on the causal dispositions of the very properties that are or may become salient for selection. Consequently, the biological approach is an abstract one as compared to the physical approach that is more detailed and relatively complete. There thus is an asymmetry between the more abstract biological homogeneous descriptions and explanations and the more detailed heterogeneous physical descriptions and explanations. This fact is brought out by the argument of multiple realization that constitutes, because of this inherent asymmetry, an argument against the nomological coextension of physical and biological concepts, laws and explanations. As concerns biological properties in general, the central point of multiple realization can be illustrated as follows:

Biological concepts that constitute functional explanations refer, because of functional similarities, in a homogenous way to biological property tokens in the world.



Physical concepts that constitute causal reductive explanations refer, because of physical differences, in a heterogeneous way to biological property tokens.

Figure 1

Since the matter is quite crucial for what follows, let me be more precise on multiple realization and the interpretation of it. In order to avoid letting biological properties slide into epiphenomena, biological property types are taken to be biological concepts (see Esfeld's contribution to this volume, Esfeld & Sachse, 2007; and Sachse, 2007, ch. 2). The functionally defined concepts of biology then may refer *multiply* in the following sense: they refer *homogeneously* to biological property tokens that are identical with physical configurations, while the latter come under *different* physical descriptions due to their varying composition. Multiple realization is, so understood, an empirical fact; for our purposes, there is nothing astonishing about it, since the approach of biology is more abstract than the physical one. Thereby, the focus on natural selection constitutes the argument for multiple realization (see Papineau, 1993, 47, and also Rosenberg, 2001): depending on the given environmental conditions, only some of the causal powers of a given physical configuration are pertinent for selection. Here, if you like, one may take the abstractions from physical details that do not change the biological function as motivated by more than instrumental simplifications of complex physical structures and changes. The epistemological focus in biology, then, is on what matters in the context of evolution, and this constitutes the ground for any further biological simplification for instrumental reasons.

The question then is how this well-established fact of multiple realization fits with the token identity claim and the completeness of physics

so that we will be able to vindicate "*Cau*". According to the scheme of multiple realization, not everything that comes under *B* also comes under a single physical description P_1. Here, P_1 is a placeholder for a detailed homogeneous physical description that, because of multiple realization, only applies to a subset of entities that come under *B*. Therefore, functional reduction as it stands (section 2) does not provide for a nomological coextension between the descriptions of biology and physics. It shows a way from physical to functional concepts of biology but not the other way round.

The question is whether the so-called local or species-specific reduction of Lewis and Kim may be used to construct biological concepts that are nomologically coextensive with the corresponding physical description. To discuss this strategy, let us refer to the common example in the philosophy of mind that goes as follows: the concept of pain reduces in one species, say humans, to one physical concept – e.g. "firing of C-fibres" (P_1) –, it reduces in another species, say octopuses, to another physical concept (P_2), etc. (see Lewis, 1980; Kim, 1998, 93–95; 2005, 24–26). On that basis, one has so-called species-specific concepts such as "pain-in-humans" (B-P_Ω) and "pain-in-octopuses" (B-P_Δ). These concepts are not purely functional concepts of the special sciences but something like semi-functional-semi-physical concepts with P_Ω and P_Δ as their physical parts. The functional concept *B* (e.g. "pain") is in this way relativized to particular species (or even has to be relativized to local physical structures if there are physical differences within the species) such that no common property specification of the function *B* remains (see Kim, 1999, 17–18).

To probe the meaning of this loosening of the functional from the physical concept, one only has to focus on the problem how to argue that species-specific concepts like "pain-in-humans" (B-P_Ω) do not refer to *different* physical structures. To put it differently, is it possible to use *physical* criteria at the biological level in order to establish nomological coextension with physical descriptions? There are three possible relations between the semi-functional-semi-physical concepts B-P_Ω (B-P_Δ) and the physical description P_1 (P_2):

1) If the physical part in the semi-functional-semi-physical concept B-P_Ω has *nothing* to do with the physical concept P_1, there is obviously no argument for nomological coextension. Pain in humans is still multiply realizable. This can easily be seen in the context of the functional

model of reduction (section 2): the reductive explanations that are based on P_1 are not linked to P_Ω but only to B, to which the physical concept P_1 is *de facto* not nomologically coextensive; in other words, since there is no link between P_1 and the added physical criterion (P_Ω) in B-P_Ω, that physical criterion cannot establish a nomological coextension between B-P_Ω and P_1.

2) If the physical element in the semi-functional-semi-physical concept B-P_Ω *contains parts* of the physical concept P_1 (P_1 = conjunction of P_{1*} and P_Ω), then the link between both concepts is still too weak to constitute nomological coextension. The physical criterion P_Ω is of course nomologically coextensive with the corresponding part (P_Ω) of P_1, but it is not nomologically coextensive with P_1 since there can be configurations in the world that are described only by P_Ω but not by P_1. If there is any possibility that the *conjunction* of B and P_Ω would be nomologically coextensive with P_1, then B has to be about something that is included in P_1 but not in P_Ω. Similarly, in the conjunction of B and P_Δ, B has to be about something that is included in P_2 but not in P_Δ. What is that something the B expresses that is physically different in both cases? If B is not about something physically different, then there would be no multiple realization to begin with. So, B has to be about something physically different, say about P_{1*} in the first and about P_{2*} in the other case. Then B-P_Ω cannot be nomologically coextensive with P_1 since B-P_Ω may refer both to the conjunction of P_Ω and P_{1*} (= P_1) and other possible conjunctions like that of P_Ω and P_{2*}.

3) It thus seems that unless the semi-functional-semi-physical concept B-P_Ω contains entirely the respective physical concept P_1, any coextension between the B-P_Ω and the physical concept P_1 is a mere *contingent* fact. If, however, $P_\Omega = P1$, then we get B-$P1$ and thus we are back to the starting point where the link between B and $P1$ (B and $P2$) is unclear.

Since this problem appears in any application of the species-specific or structure-specific model in biology, it is clear why this approach cannot make "*Cau*" intelligible: unless a given semi-biological-semi-physical concept does not contain the respective physical concept (like P_1 or P_2), no nomological coextension with physical concepts can be established. If there is no such nomological coextension, it remains unclear what the biological part B in the semi-biological-semi-physical concept is about – what the common specification of B is. Somehow worse, if the semi-

biological-semi-physical concept in fact contains the physical concept P_1, then there is no link at all between *B* and physics. To sum up the essential point, adding *physical* criteria to biological concepts does not help to make intelligible how "*Cau*" can be justified for any biological concept *B* (and thus of biology in general).

Against this failure, one has to look for another theoretical strategy to establish nomological coextension between the descriptions of biology and physics. To do so, let me reconsider multiple realization once again in terms of the causal-functional theory of properties (see Michael Esfeld's contribution to this volume). If local physical structures coming under one concept *B* are described in terms of different physical concepts (like P_1 and P_2 in our schema), then there is a difference in composition among these structures. Each of these physical concepts picks out a minimal sufficient condition to bring about the effects that define *B*, given certain normal background conditions. In order to get from structures coming under P_1 to structures coming under P_2, one has to substitute at least one of the parts that are *necessary* to bring about the effects in question with a part of another type. If – and only if – one takes the causal-functional theory of properties for granted, any such replacement implies a systematic difference in the way in which these structures cause the effects that define *B*. It is then excluded that one can replace a local physical structure of type P_1 by a local physical structure of type P_2, thus obtaining a different physical realizer of *B without* making a causal difference (see also Kim, 1999 and 2005, 26).

If the effects that define *B* can be brought about by two or more different physical properties, we will find a difference in the production of side effects that are systematically linked with the main effects in question. Think of different causal interactions with the physical environment within the cell when a gene is transcribed and proteins are synthesized that make up its characteristic phenotypic effects. For any such difference in the causal sequence from the DNA transcription to the protein synthesis, there exists the possibility that the difference may become pertinent to the shift of selection pressures within the target environment (see Rosenberg, 1994, 32). Consequently, that difference can *in principle* also be considered in terms of the concepts that are proper to biology to which *B* belongs. Here, more precise functional definitions will help us to account for different reaction norms, and thus, physical differences. A reaction norm can be described by a mathematical func-

tion over the different probabilities of fitness contributions in different environments. Against this background, for the concept *B* (that is multiply realized by P_1 and P_2), it is possible to conceive two functional sub-concepts B_1 and B_2 taking different reaction norms into account (see also Bechtel & Mundale 1999 with regard to the more fine-grained functional concepts of the special sciences).

It follows from the outlined argument that the sub-concepts are thus no longer multiply *realizable* since *any* physical difference that is constitutive for multiple realization (that is a different way to bring about the effects that define *B*) leads to specific functional differences – that is, to a unique reaction norm. The functionally defined sub-concepts thus correspond by definition to one single type of physical configuration that brings about the effects that define *B* in one particular way. Having the sub-concepts so defined, they are thus nomologically coextensive with the physical concepts P_1 and P_2.

For instance, let us consider a gene of *E. coli* that has effects that are pertinent for its fitness and that is accordingly functionally defined in terms of biology. Think of our cell-wall biosynthesis example. The gene tokens coming under *B* are defined by their characteristic expression of membrane proteins that are crucial for the cell growth of the bacterium before cell division, etc. Independently of our chosen level of genetic simplification, the gene tokens coming under *B* are identical with certain physical configurations (DNA sequences) that are described differently in terms of physics (by P_1 and P_2) since there are differences in the physical composition of the DNA sequences in question. Nonetheless, because of the redundancy of the genetic code, all these physically different DNA sequences code for proteins of the same type (or any other effect that is considered in the functional definition *B*). The crucial point here is that according to the physical differences between P_1 and P_2, there are different physical paths to bring about the effect in *B*. These different ways to produce the effects (the proteins for instance) are, as current research confirms more and more, systematically linked with possible side effects or reaction norms (see below). Differences in side effects have an effect on the overall evolutionary trajectory, as for instance in that they express functional differences that lead to different selection pressures, such that so-called codon-bias. Codon-bias is a statistical skewing towards a specific DNA sequence (thus specific physical configurations of genes). This arises because the physical differences

of the DNA sequences have been and continue to be pertinent to natural selection *under certain environmental conditions*.

To have a better idea of such functionalizable side effects or reaction norms, think of differences in the speed or the accuracy of the protein production, of which we have quite illustrative and well-confirmed examples (see among many others Bulmer, 1991, Hartl et al., 1994 and Gerland & Hwa, 2009, for such functional side effects in certain genes of *E. coli*, see Mukhopadhyay et al., 2008, for functional side effects in plants, see Kimchi-Sarfaty et al., 2007; Yang & Nielsen, 2008 and Moses & Durbin, 2009, for functional side effects in certain genes of mammals, see Sotlzfus, 2006 and dos Reis & Wernisch, 2009, for general and comparative considerations). To sum up, depending on the environmental conditions, certain DNA sequences are more optimal than others (and thus not selectively neutral) and this can be taken into account in more precise functional definitions.

The issue is of course more complicated than it is sketched out here. One may thus object that the codon-bias that results out of the given selection pressure depends on many other factors than only on a faster production of proteins. Of course it does. The selection pressure (and thus the codon-bias) depends for instance on the selective importance of the produced proteins. If the protein is not that important for the organism, the selection pressure and thus the codon-bias for a particular DNA sequence is accordingly low. However, selection pressure depends on the adaptive landscape. Shifts in environmental conditions can activate hitherto latent selection pressures. At this point it becomes clear to what extent the sub-concepts (that take into account that very issue) are *theoretical* constructions. Let me note here that this quick example from the empirical data serves mainly to show at what point the construction of functionally defined sub-concepts may represent the successes and lacuna of current genetic research as it searches for functional differences that correspond to physical differences. This fact therefore suggests that biology has in principle the means to consider the reaction norms of P_1 and P_2 and to construct functionally defined sub-concepts. To put it differently, under certain environmental conditions, it seems *necessary* to take into account functional differences that result from even minor physical differences in order to coherently explain evolutionary pathways.

Another crucial issue is linked to mutations and the frequencies of

their appearance. One may object that mutation frequencies (and thus appearances of physical differences) are sometimes so high that in fact no selection pressure occurs in favour of or against specific DNA sequences. To put it simply, the number of generations with a specific DNA sequence for one gene is not large enough before it changes physically because of the given mutation rate for that very gene to allow us to speak of a specific selection pressure (codon-bias). This observation is at the heart of the so-called neutralism debate in genetics (see Nei, 2005). However, this possibility (or fact if you prefer) doesn't block the theoretical possibility of constructing functionally defined sub-concepts that do not necessarily tell us what fitness contribution the gene in question *in fact* provides. The fitness index is dependent on the given environment anyway. Moreover, the sub-concepts articulate the *dispositions* for fitness contribution. These dispositions are inherently sensitive to physical differences by definition. In saying this, we are not only making the case for the metaphysical underpinning of reductionism; we are, as well, reflecting the contemporary debate on the understanding of fitness in terms of propensities or dispositions (see classically Mills & Beatty, 1979; see also Weber, 1996; Ariew & Lewontin, 2004; Krimbas, 2004; Ariew & Ernst, 2009 for clarifications and critical comments that, however, do not affect the main line of reasoning of this section: physical differences have, under certain environmental conditions, an impact on the biological level).

Biological research currently suggests that the differences in side effects, or effects that have no seemingly adaptive purpose, are not insulated from selection, but given changes in a physical environment, become the target of selection. Depending on changes in the physical environment (including the overall genetic makeup of the species population), certain DNA sequences may obtain a selective advantage over other DNA sequences because of possible differences in, for instance, the speed and accuracy of the production of the same proteins. It is easy to imagine adaptive scenarios in which the accurate and fast production of the membrane proteins in question may become important for the survival of the bacterium. Since fitness differences can theoretically be measured, biology has the means to consider them. Consequently, for any concept *B* defining a certain type of gene of *E. coli*, it is possible to conceive functional sub-concepts B_1 and B_2 taking into account these side effects (like the speed and accuracy of the protein production) by

means of considering the resulting measurable fitness differences. Thus, once again simplifying in order to illustrate the idea, B_1 may be the conjunction of the gene tokens that express the protein in question (like all gene tokens coming under *B*) *and* the consideration of a certain time index of the protein production or the corresponding probabilities on fitness contributions, distinguishing it from gene tokens coming under B_2). This more precise rendering of sub-concept B_1 may be written as "*B* and production of the characteristic effect *X* in t_1" or "*B* and probability function *C*1 of fitness contribution", while the sub-concept *B*2 may be something like "*B* and production of the characteristic effect *X* in t_2" or "*B* and probability function *C*2 of fitness contribution" (see Sachse 2007, chapter 4 to 7, for a detailed case study of the reduction of classical to molecular genetics along these lines).

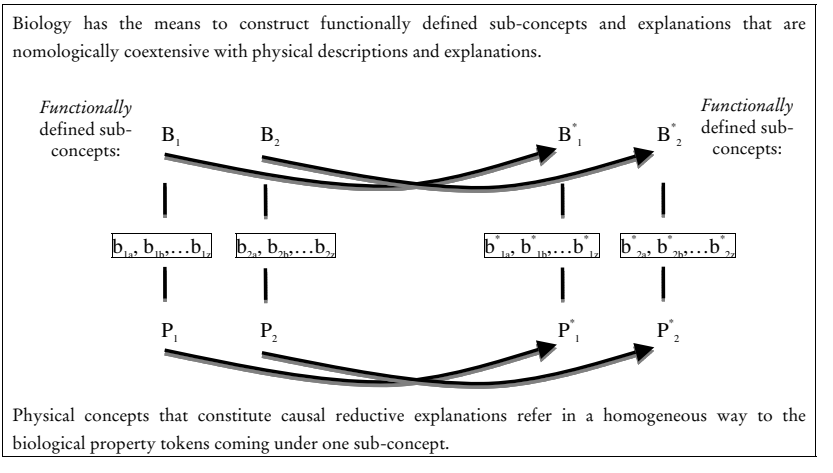


Figure 2

To sum up the crucial point of this section, biology has the means to construct, *in theory*, functionally defined sub-concepts that are, as we shall discuss in more detail in the following section, nomologically coextensive with physical descriptions. Therefore, it is possible to apply our reasoning of the previous section to these sub-concepts: they are about causally efficacious property tokens ("*Cau*") such that there is no danger of epiphenomenalism for biological tokens insofar as they are described by those sub-concepts.

4. The scientific quality of biology ("*Cau*" and "¬*Rep*")

By means of these sub-concepts we attain concepts of biology that are nomologically coextensive with physical concepts and thus make it possible to reduce biology to physical theories in a functional manner, with three steps: 1) within an encompassing fundamental physical theory P, we construct the concepts P_1, P_2, etc. to capture the differences in composition among the local physical structures that are all described by the same concept B; 2) B is more precisely articulated by constructing functional sub-concepts B_1, B_2, etc. of B, each of which captures the systematic side effects linked to the different ways of producing the effects that define B. To put it differently, the sub-concepts are constructed out of B in such a way that they are nomologically coextensive with the concepts P_1, P_2, etc. using the functional model of reduction shown in section 2; (3) B is reduced to P via B_1, B_2, etc. and P_1, P_2, etc. Reducing B (and thus biology) here means that starting from P, we can construct P_1, P_2, etc. and then deduce B_1, B_2, etc. from P_1, P_2, etc. given the nomological coextension. One derives B *by abstracting* from the conceptualization of the functional side effects contained in B_1, B_2, etc. for any environmental context where the functional side effects are not manifested or not pertinent to selection (see Esfeld & Sachse 2007).

The above-mentioned sub-concepts are not construed in a local or species-specific way that contains physical criteria, but in terms of purely functional differences only, say, different dispositions for fitness contributions. The functional sub-concepts B_1 and B_2 of B are distinct only by conceptualizing the different ways in which the effects that define B are or may be produced. Consequently, B always has the same substantial "specification of the function" in B_1, B_2: these sub-concepts clearly express for biologists what their referents *functionally* have in common (the disposition *to* produce certain effects) and what their *functional* differences are (*the way in which* these effects are produced). To simplify, B is both a) an abstraction from certain dispositions, what we called side-effects in the genetic example and that are only brought out by the sub-concepts, and it is b) a focus on dispositions, call them pertinent similarities under certain environmental conditions, that are contained in the sub-concepts as well. For example, *E. coli* gene tokens falling under any sub-concept B_1, B_2, etc. are biologically understood by taking into account fitness differences that are related to their expres-

sion of a certain protein, which means that our sub-concepts bring out salient causal similarities, the expression of the protein in question, and the laws relating to that effect. The concept *B* has the same substantial "specification of the function" in all these sub-concepts since the latter ones are constructed out of *B*. In any sub-concept, the disposition *to* produce the characteristic effect is contained as well. Therefore, this proposal does not put the scientific quality of the concept *B* and the laws in which it figures, couched in terms of *B*, in jeopardy, but on the contrary, justifies biology as a science by linking *B* and its laws in terms of *B* via its sub-concepts with physics.

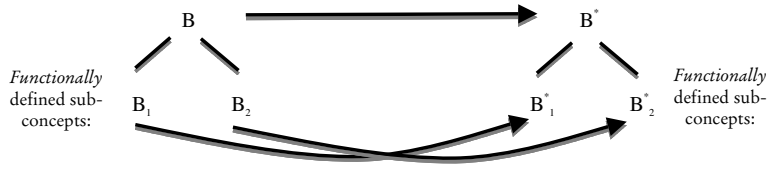
Let us see how we are now in the position to vindicate both "*Cau*" and "¬*Rep*" for *B*. On the basis of the fundamental physical laws, one can construct laws in terms of P_1, P_2, etc. that refer to the properties on which biology focuses. From those laws, one can deduce biological laws in terms of B_1, B_2, etc. given the nomological coextension of these concepts. These sub-concepts and any laws and explanations that are based on them are not about epiphenomena (thus vindicating "*Cau*"). Nonetheless, they were replaceable by physics because of nomological coextension (no vindication of "¬*Rep*"). However, one reaches the laws and explanations in terms of *B* by abstracting from the conceptualization of the functional side effects that are represented in *B*1, *B*2, etc. Since the "specification of the function" of *B* is contained in each of its sub-concepts, the abstract concept *B* cannot be eliminated. The abstract laws of biology couched in terms of *B* are non-physical or not replaceable by physics in the sense that there is no single physical law having the same extension as any of these laws, vindicating "¬*Rep*" for *B*. The fundamental physical laws are too general, applying to everything that there is in the world, and the law-like generalizations couched in terms of those physical concepts that focus on the composition of the complex objects in question (the concepts P_1, P_2, etc.) are too restricted. When talking about complex objects such as e.g. genes, cells, or whole organisms, the physical concepts focus on the composition of these objects. Due to selection there are salient causal similarities among effects that such complex objects produce as a whole, although they differ in composition. When we consider the concepts that capture these similarities, we don't consider them as physical concepts, but – since they are relative to selection – take them to be concepts of biology.

Since in our world many environments are such that there is no dif-

ferential selection for certain differences in composition among complex objects, being concerned only with the effects that these objects produce as a whole, the abstract concepts of biology possess a scientific quality, figuring in law-like generalizations that capture something that is objectively there in the world. Nonetheless, these concepts and law-like generalizations do not conflict with the completeness of physics and the supervenience of everything on the physical, since, as we have shown, there is a reductive method to express them on the basis of the fundamental physical concepts and laws. The outcome of our argument thus vindicates the scientific quality of these biological concepts, without epiphenomenalism or eliminativism (see also my reconsideration of the status of abstract concepts in the context of biological functions in the following section).

One may sum up the essential point of this strategy as follows: the asymmetry that is spelled out by multiple realization is only a problem if this asymmetry is *all* that one can say about the relation between two different sciences. In addition to the still given asymmetry between the functionally defined abstract concepts (B) of biology and physical descriptions (P_1 and P_2), the proposed strategy establishes *symmetric* links in the form of nomological correlations between functionally defined sub-concepts (B_1 and B_2) of B and the physical descriptions (P_1 and P_2). It is in this sense that one may either call B to be "reducible" (via its sub-concepts) to physics, or call B to be "abstractable" from its sub-concepts (that are reducible to physics), depending on one's preferences.

One derives the original biological concepts *by abstracting* from functional side effects contained in the sub-concepts for any environmental context where these functional side effects are not manifested or not pertinent to selection.



The sub-concepts are constructed out of the original concepts and thus share the same meaning. Therefore, the abstract concepts are also about causally efficacious property tokens even though not replaceable by physical descriptions.

Figure 3

Regardless of our terminological preference, my general argument establishes the scientific quality of the descriptions of biology only insofar as selection is concerned. Since selection is widely acknowledged to be the reason why there is multiple realization at all (Papineau, 1993, ch. 2), it follows that the approach can be applied to any other biological branch as well. The argument can be summed up as follows: a) classical reductionism is considered to be refuted by the thesis of multiple realization. Multiple realization is supposed to be based upon natural selection; b), our proposed conservative reductionist strategy derives from our acceptance of multiple realization, and of its basis. It should now be clear why neo-Darwinian evolution is the reason why there seems to be a problem for reductionism, as well as why we believed, given the construal of selection that we proposed to outline, we could nevertheless reduce biology with an alternative strategy.

Let me be clear on the second crucial ingredient of this alternative strategy. This proposal to establish a systematic link between biology and physics via functional sub-concepts presupposes the causal-functional theory of properties (see Michael Esfeld's contribution to this volume). If the fundamental properties were pure qualities, so that what they are is independent of the causal relations in which objects stand in virtue of having these properties, then any causal-functional description of biology could trivially be multiply realized by replacing pure qualities of one type with pure qualities of another type. In this schema, there would be no causal difference effected by that substitution (see also Jackson, 1998, 23–24). By the same token, this proposal presupposes the anti-Humean view of the laws of nature being metaphysically necessary that goes with the causal theory of properties: otherwise, the same biological laws could be reached from different sets of fundamental physical laws that are obtained by changes in the distribution of the fundamental physical, purely qualitative properties. Consequently, there would no longer be one single fundamental physical theory – or one coherent set of fundamental physical theories – to which biological theories could in principle be reduced. With my premises, then, in place, I showed, on the basis of all properties being causal-functional ones, that the proposed reductionist strategy is indeed able to show how both biological theories and their target entities are related to fundamental physics down to their specific characteristics.

5. Perspectives for other debates in the philosophy of biology

What is the utility of reductionism? The specific character of biology has posed a dilemma for the philosophy of science, in as much as science is considered to be unified. Conservative reduction redeems biology for a united science by showing how its discourse can be linked to physics and thus avoids primarily metaphysical conflicts. By adopting the proposed reductionist strategy, we block the danger of epiphenomenalism or eliminativism as the ultimate philosophical framework for abstract biological concepts, laws and explanations. In addition, conservative reductionism may constitute a plausible framework for approaches that normally have an anti-reductionist background. For instance, it possible to take those explanations of biology that have no coextensive counterpart in physics as *unifying* explanations in Philip Kitcher's sense (Kitcher, 1976 and 1981). By abstracting from physical differences, one and the same biological concept, explanation and law applies to physically different entities. It then seems that Kitcher's arguments in favour of the scientific quality of classical genetics (and of biology in general) can now be grounded within or made compatible with the proposed conservative reductionist framework (compare Kitcher, 1984, with the strategy outlined above). Following this reasoning, fitness, natural selection, genes, etc. can be understood as concepts with a high degree of abstraction, which lead to unifying biological explanations that have no coextensive equivalent in physics. Keep in mind that the given degree of abstraction is based on how the world actually is and evolves and whether certain physical difference may or may not imply functional differences that have to be taken into account.

The reductionist strategy constitutes hence an explicit argument to take the principle of natural selection and other biological generalizations as law-like (see also Rosenberg, 2001 2006, ch. 4 and Sober, 2000, ch 1.4). In order to show this point, the concepts that constitute the principle of natural selection (or any other biological generalization) only have to be theoretically connectable via sub-concepts with physical descriptions and laws. Biological generalizations that are couched in terms of sub-concepts get their law-like character from physics deductively, on account of nomological coextension. From this move, abstract biological generalizations inherit their law-likeness, since they only abstract from certain functional details, which is, depending on the

context, scientifically justified if coherent explanations can still be provided. Through this argument, the principle of natural selection appears to be the most abstract and unifying law-like generalization of biology that is, by means of its application to specific units of selection, connected to physics.

Following this conclusion, we can then specify the difference between so-called proximate and ultimate explanations in a particular way. Using Sober's question "Why do ivy plants grow toward the sunlight?", there are two possible answers that characterize proximate and ultimate explanations respectively (see Sober, 2000, ch. 1.2): an answer that is couched in terms of the physiological mechanisms that programme the plant to engage in phototropism is a proximate explanation, which refers to ontogenetic causes and provide mechanistic explanations. Alternatively, one may outline the phylogenetic causes and explain the capacity of phototropism in relation to adaptation and selective advantage. Postulating an evolutionary lineage that begins with the first organisms (so to speak) that possessed this trait, we can outline via genetic transmission from generation to generation (or across species boundaries due to lateral gene transfer or hybridisation) a clear causal relation between the ivy and its place in the timeline of natural selection on which an explanation of the adaptive purpose of phototropism can be based.

In the context of the reductionist approach, the difference between both explanations can be spelled out as difference in degrees of abstraction. The proximal explanation refers to local processes that occur in each generation again and again, and since a fitness contribution is thus implied, one may, by summing up and abstracting from several details, end up with ultimate explanations. Thereby, it depends on the given and changing environmental conditions what kind or degree of abstraction is justified. To put it differently, the reductionist approach constitutes a hierarchical system of concepts, laws and explanations and thereby shows how ultimate explanations are related to proximal explanations.

Against this background, one may also elaborate on a similar debate – about the different approaches to the very notion of biological functions. Let us consider for instance the two most current approaches – the etiological one that determines biological functions generally as selected effects, thus, by a reference to the evolutionary past (see also Millikan, 1989; Neander, 1991; Griffiths, 1993: Mitchell, 1993; Godfrey & Smith, 1993 and 1994; and Schwarz, 1999) and the systemic or causal-

dispositional approach that defines biological functions without such a reference to the evolutionary past (see also Bigelow & Pargetter, 1987; Amundson & Lauder, 1994; Manning, 1997; Weber, 2005 ch. 2.4; and Mossio et al., 2009; see furthermore Kitcher, 1993; and Arp, 2007 for some kind of compatibility of these approaches). Without going into the details of these approaches – since it is beyond the scope of this paper to add any new argument or counterarguments at this point – one may take the proposed approaches as differing mainly in their extension and thus explanatory force.

While the etiological approach refers to past occurrences and thus mainly to manifested and pertinent dispositions and therefore is somehow similar to ultimate explanations based on phylogeny, the systemic and causal-dispositional approach sticks more to local causal capacities of systems or dispositions and thus can be identified with ontogenetic mechanistic explanations. By adopting a reductionist perspective according to the outlined strategy – that the etiological approach is a more abstract approach or concept of biological functions than the other one but nonetheless reducible to it – one may explain away the main difficulties and make more explicit the advantages of both approaches. Thereby, once again, it depends on the given and changing environmental conditions whether the more abstract approach is justified. To put it differently, it is the environment that constitutes the normative aspect for any kind of functional ascription. If, for instance, minor physical differences have an impact on the function under the given environmental conditions, there is an argument to account for this fact and this is, in the most extreme case, done by the construction of functional sub-types in terms of causal dispositions. By contrast, many physical differences have no functional impact under certain environmental conditions such that more abstract approaches, even with a historical dimension, are admitted or even preferred to provide more unifying explanations.

What I have proposed with respect to the debate on biological functions and on biological concepts, laws and explanations of different possible degrees of abstraction applies as well to the debate on biological taxa being natural kinds (see for the debate among others Brigandt, 2003 and 2009; Dupré, 1981; Ereshefsky, 2007 and 2010; LaPorte, 2004; Mallet, 2010; O'Malley, 2010 and Richards, 2008). Conservative reductionism supports a realist attitude with respect to biological kinds in the following general way: since the sub-concepts are nomologically

coextensive with physical descriptions, it is possible to apply any argument in favour of (composed) physical kinds being natural ones to the biological sub-concepts as well. Thus, the more abstract biological concepts inherit their naturalness and counterfactual robustness from their sub-concepts, or, to put it differently, the reductionist framework makes explicit the hierarchical structure of a system of natural biological kinds that is *theoretically* achievable. Additionally, depending on the given and changing environmental conditions, the abstract biological concepts such as biological taxa can figure in biological laws and explanations. Thereby, neither inheritance nor the biological sphere's systematic hierarchical structure contains, in the ideal case, any conventionalist aspect.

Still, biological species are evolving while physical natural kinds are not and, from a biological perspective, evolution contains some kind of contingency. This suggests to understand biological species rather as individuals than genuine kinds with real essences. However, sub-concepts do not contain more contingency than (composed) physical kinds to which they are nomologically coextensive. Furthermore, there is no principal difference whether we consider multiple realization of a type at one specific time or for a period of time. For instance, imagine an abstract concept B_{t1} that applies to any member of a species at t_1 and this concept can be conservatively reduced via its sub-concepts to physics. Look at that species at a later stage in evolution (at t_2) and imagine once again that an abstract concept B_{t2} applies to any member of a species and this concept can be conservatively reduced via its sub-concepts to physics as well. If we now compare both abstract concepts B_{t1} and B_{t2}, it is likely that they differ somehow and it is even more likely that their sub-concepts differ somehow since evolution has taken place. However, there is no principal objections that both abstract concepts B_{t1} and B_{t2} may constitute themselves two sub-concepts for some more abstract concept that bring out salient characteristic similarities that figure in explanations. Call this a theoretical species concept that applies to B_{t1} and B_{t2}.

It is out of the scope of this paper to discuss this approach and seemingly problematic issues like speciation in detail. However, whether physical differences give rise to speciation depends on the given context. Within the framework of conservative reductionism it thus is suggested that differences in essence in combination with the given environmental

conditions constitute or not the starting point for speciation. In other terms, phylogenesis during evolution does not depend on us but on the world and the underlying physical structures and changes that can be, in theory, considered in terms of sub-concepts and more abstract concepts. On that theoretical basis, classifications that mostly focus on a historical dimension like common ancestry and that may exclude real essences *do not hinder* to construct biological kinds ahistorically with genuine essences. Here as well evolution gives us an impressive idea what kind of biological species were and actually are realized.

Note

* Special thanks to Michael Esfeld, Patrice Soom, the participants of the workshop on "Reduction, explanation and metaphors in the philosophy of mind" (Bremen, September 2009), those of the "European advanced seminar in philosophy of life science" (Geneva, September 2010) and the anonymous reviewer for constructive comments on this paper.

Bibliography

Amundson, Ron; Lauder, George V. (1994): Function without purpose. In: *Biology and Philosophy* 9, S. 443–469.

Ariew, André; Lewontin Richard, C., 2004: Confusions of fitness. In: *The British Journal for the Philosophy of Science* 55, pp. 347–363.

Ariew, André; Ernst, Zachary, 2009: What fitness can't be. In: *Erkenntnis* 71, pp. 289–301.

Arp, Robert, 2007: Evolution and two popular proposals for the definition of function. In: *Journal for General Philosophy of Science* 38, pp. 19–30.

Bickle, John, 1998: *Psychoneural reduction: the new wave*. Cambridge (Massachusetts): MIT Press.

Bigelow, John; Pargetter, Robert: 1987: Functions. In: *Journal of Philosophy* 84, pp. 181–196.

Brigandt, Ingo, 2003: Species pluralism does not imply species eliminativism. In: *Philosophy of Science* 70, pp. 1305–1316.

Brigandt, Ingo, 2009: Natural kinds in evolution and systematics: Metaphysical and epistemological considerations. In: *Acta Biotheoretica* 57, pp. 77–97.

Bulmer, Michael, 1991: The selection-mutation-drift theory of synonymous codon usage. In: *Genetics* 129, pp. 897–907.

Chalmers, David, 1996: *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Craver, Carl, 2001: Role functions, mechanisms, and hierarchy. In: *Philosophy of Science* 69, pp. 53–74.

Craver, Carl, 2006: When mechanistic models explains. In: *Synthese* 153, pp. 355–376.

Cummins, Robert, 1975: Functional analysis. In: *Journal of Philosophy* 72, pp. 741–764. Reprinted in Sober, E. (ed.), 1994: *Conceptual issues in evolutionary biology*. Cambridge (Massachusetts): MIT Press, pp. 49–69.

Dobzhansky, Theodosius, 1973: Nothing in biology makes sense except in the light of evolution. In: *American Biology Teacher* 35, pp. 125–129.

dos Reis, Mario; Wernisch, Lorenz, 2009: Estimating translational selection in eukaryotic genomes. In: *Molecular Biology and Evolution* 26, pp. 451–461.

Dupré, John, 1981: Natural kinds and biological taxa. In: *The Philosophical Review* 90, pp. 66–90.

Ereshefsky, Marc, 2007: Species, taxonomy, and systematics. In: Matthen, M.; Stephens, C. (eds.): *Handbook of the philosophy of science: Philosophy of biology*. Amsterdam: Elsevier, pp. 406–427.

Ereshefsky, Marc, 2010: Microbiology and the species problem. *Biology and Philosophy* 25, pp. 553–568.

Esfeld, Michael; Sachse, Christian, 2007: Theory reduction by means of functional sub-types. In: *International Studies in the Philosophy of Science* 21, pp. 1–17.

Esfeld, Michael; Sachse, Christian, 2010: *Kausale Strukturen. Einheit und Vielfalt in der Natur und den Naturwissenschaften*. Berlin: Suhrkamp. English version *Conservative reductionism*. Forthcoming New York: Routledge spring 2011.

Endicott, Ronald, 1998: Collapse of the new wave. In: *The Journal of Philosophy* 95, pp. 53–72.

Fodor, Jerry A., 1974: Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28, pp. 97–115.

Gerland, Ulrich; Hwa, Terence, 2009: Evolutionary selection between alternative modes of gene regulation. In: *Proceedings of the Nation-*

al Academy of Sciences of the United States of America 106, pp. 8841–8846.

Godfrey-Smith, Peter, 1993: Functions: consensus without unity. In: *Pacific Philosophical Quarterly* 74, pp. 196–208. Reprinted in: Hull, D.; Ruse M. (eds.),1998: *The philosophy of biology: Oxford readings in philosophy*. Oxford: Oxford University Press, pp. 280–292.

Godfrey-Smith, Peter, 1994: A modern history theory of functions. In: *Noûs* 28, pp. 344–362.

Griffiths, Paul E., 1993: Functional analysis and proper functions. In: *The British Journal for the Philosophy of Science* 44, pp. 409–422.

Hartl, Daniel L.; Moriyama, Etsuko; Sawyer, Stanley A., 1994: Selection intensity for codon bias. In: *Genetics* 138, pp. 227–234.

Heil, John, 2003: *From an ontological point of view*. Oxford: Oxford University Press.

Jackson, Frank, 1998: *From metaphysics to ethics. A defence of conceptual analysis*. Oxford: Oxford University Press.

Kim, Jaegwon, 1998: *Mind in a physical world. An essay on the mind-body problem and mental causation*. Cambridge (Massachusetts): MIT Press.

Kim, Jaegwon, 1999: Making sense of emergence. In: *Philosophical Studies* 95, pp. 3–36.

Kim, Jaegwon, 2005: *Physicalism, or something near enough*. Princeton: Princeton University Press.

Kimchi-Sarfaty, Chava; Mi Oh, Jung; Kim, In-Wha; Sauna, Zuben E.; Calcagno, Anna Maria; Ambudkar, Suresh V.; Gottesman, Michael M., 2007: A 'silent' polymorphism in the MDR1 gene changes substrate specificity. In: *Science* 315, pp. 525–528.

Kitcher, Philip, 1976: Explanation, conjunction, and unification. In: *Journal of Philosophy* 73, pp. 207–212.

Kitcher, Philip, 1981: Explanatory unification. In: *Philosophy of Science* 48, pp. 507–531.

Kitcher, Philip, 1984: 1953 and all that. A tale of two sciences. In: *Philosophical Review* 93, pp. 335–373. Reprinted in: Kitcher, P., 2003: *In Mendel's mirror. Philosophical reflections on biology*. Oxford: Oxford University Press, pp. 3–30.

Kitcher, Philip, 1993: Function and design. In: *Midwest Studies in Philosophy* 18, pp. 379–397. Reprinted in: Kitcher, P., 2003: *In Mendel's mirror. Philosophical reflections on biology*. Oxford: Oxford Uni-

versity Press, pp. 159–176 and also reprinted in Hull D.: Ruse, M. (eds.), 1998: *The philosophy of biology. Oxford readings in philosophy*. Oxford: Oxford University Press, pp. 258–279.

Krimbas, Costas B., 2004: On fitness. In: *Biology and Philosophy* 19, pp. 185–203.

LaPorte, Joseph, 2004: *Natural kinds and conceptual change*. Cambridge: Cambridge University Press.

Lewis, David, 1980: Mad pain and Martian pain. In: Block, N. (ed.): *Readings in the philosophy of psychology. Volume 1*. London: Methuen, pp. 216–222. Reprinted in Lewis D., 1983: *Philosophical papers. Volume 1*. Oxford: Oxford University Press, pp. 122–130.

Lewis, David, 1986: *On the plurality of worlds*. Oxford: Blackwell.

Lewis, David, 1994: Reduction of mind. In: Guttenplan, S. H. (ed.): *A companion to the philosophy of mind*, Oxford: Blackwell, pp. 412–431.

Machamer, Peter; Darden, Lindley; Craver, Carl F., 2000: Thinking about mechanisms. In: *Philosophy of Science* 67, pp. 1–25.

Mallet, James, 2010: Why was Darwin's view of species rejected by twentieth century biologists? In: *Biology and Philosophy* 25, pp. 497–527.

Manning, Richard, 1997: Biological function, selection, and reduction. In: *The British Journal for the Philosophy of Science* 48, pp. 69–82.

Millikan, Ruth Garrett, 1989: In defense of proper functions. In: *Philosophy of Science* 56, pp. 288–302.

Mills, Susan K.; Beatty, John H., 1979: The propensity interpretation of fitness. In: *Philosophy of Science* 46, pp. 263–286.

Mitchell, Sandra D. 1993: Dispositions of etiologies? A comment on Bigelow and Pargetter. In: *The Journal of Philosophy* 90, pp. 249–259.

Moses, Alan M.; Durbin, Richard, 2009: Inferring selection on amino acid preference in protein domains. In: *Molecular Biology and Evolution* 26, pp. 527–536.

Mossio, Matteo; Saborido, Cristian; Moreno, Alvaro, 2009: An organizational account of biological functions. In: *The British Journal for the Philosophy of Science* 60, pp. 813–841.

Mukhopadhyay, Pamela; Basak, Surajit; Ghosh, Tapash Chandra, 2008: Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and arabidopsis. In: *DNA Research* 15, pp. 347–356.

Nagel, Ernest, 1961: *The structure of science. Problems in the logic of scientific explanation*. London: Routledge.

Neander, Karen, 1991: Function as selected effects: the conceptual analyst's defense. In: *Philosophy of Science* 58, pp. 168–184.

Nei, Masatoshi, 2005: Selectionism and neutralism in molecular evolution. In: *Molecular Biology and Evolution* 22, pp. 2318–2342.

O'Malley, Maureen A., 2010: Ernst Mayr, the tree of life, and philosophy of biology. In: *Biology and Philosophy* 25, pp. 529–552.

Papineau, David, 1993: *Philosophical naturalism*. Oxford: Blackwell.

Papineau, David, 2002: *Thinking about Consciousness,* Oxford: Oxford University Press.

Putnam, Hilary, 1967/1975: The nature of mental states. In: Putnam, Hilary (ed.), 1975: *Mind, language and reality: Philosophical papers. Volume 2*. Cambridge: Cambridge University Press, pp. 429–440. First published as Psychological predicates. In: Capitan, W. H.; Merrill, D. D. (eds.), 1967: *Art, mind and religion*. Pittsburgh: University of Pittsburgh Press.

Richards, Richard A., 2008: Species and taxonomy. In: Ruse, M. (ed.), 2008: *The Oxford handbook of philosophy of biology*. Oxford: Oxford University Press: pp. 161–188.

Rosenberg, Alexander, 1978: Supervenience of biological concepts. In: *Philosophy of science* 45, pp. 368–386.

Rosenberg, Alexander, 1994: *Instrumental biology or the disunity of science*. Chicago: University of Chicago Press.

Rosenberg, Alexander, 2001: How is biological explanation possible?. In: *The British Journal for the Philosophy of Science* 52, pp. 735–760.

Rosenberg, Alexander; Kaplan, D. M., 2005: How to reconcile physicalism and antireductionism about biology? In: *Philosophy of Science* 72, pp. 43–68.

Rosenberg, Alexander, 2006: *Darwinian reductionism. Or, how to stop worrying and love molecular biology*. Chicago: The University of Chicago Press.

Sachse, Christian, 2007: *Reductionism in the philosophy of science*. Frankfurt (Main): Ontos-Verlag.

Schwartz, Peter H., 1999: Proper function and recent selection. In: *Philosophy of Science* 66, pp. S210–222.

Sober, Elliot, 2000: *Philosophy of biology. Second Edition*. Boulder: Westview Press.

Stoltzfus, Arlin, 2006: Mutation-biased adaptation in a protein NK model. In: *Molecular Biology and Evolution* 23, pp. 1852–1862.

Waters, C. Kenneth, 1994: Genes made molecular. *Philosophy of Science* 61, pp. 163–185.

Waters, C. Kenneth, 2007: Causes that make a difference. In: *Journal of Philosophy* 104, pp. 551–579.

Weber, Marcel, 1996: Fitness made physical: The supervenience of biological concepts revisited. *Philosophy of Science* 63, No. 3, pp. 411–431.

Weber, Marcel, 2005: *Philosophy of experimental biology*. Cambridge: Cambridge University Press.

Wright, Larry, 1973: Functions. In: *Philosophical Review* 82, pp. 139-168. Reprinted in: Sober, E. (ed.), 1994: *Conceptual issues in evolutionary biology*. Cambridge (Massachusetts): MIT Press, pp. 27–47.

Wouters, Arno G., 2003: Four notions of biological function. In: *Studies in History and Philosophy of Biological and Biomedical Sciences* 34, pp. 633–668.

Yang, Ziheng; Nielsen, Rasmus, 2008: Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. In: *Molecular Biology and Evolution* 25, pp. 568–579.

Douglas Kutach

Reductive Identities: An Empirical Fundamentalist Approach*

Abstract

I sketch a philosophical program called 'Empirical Fundamentalism', whose signature feature is the extensive use of a distinction between fundamental and derivative reality. Within the framework of Empirical Fundamentalism, derivative reality is treated as an abstraction from fundamental reality. I show how one can understand reduction and supervenience in terms of abstraction, and then I apply the introduced machinery to understand the relation between water and H_2O, mental states and brain states, and so on. The conclusion is that such relations can be understood either as metaphysical contingencies or as necessary type-identities.

Zusammenfassung

Ich charakterisiere das philosophische Programm des "Empirischen Fundamentalismus", das sich hauptsächlich durch seine Verwendung einer Unterscheidung in fundamentale Realität einerseits und derivative Realität andererseits auszeichnet. Innerhalb des Rahmens des Empirischen Fundamentalismus wird derivative Realität als Abstraktion von fundamentaler Realität verstanden werden können, und ich wende den dabei eingeführten Apparat an, um die Relation zwischen Wasser und H_2O, zwischen mentalen Zuständen und Gehirnzuständen und dergleichen zu erhellen. Die Konklusion besteht darin, dass diese Relationen entweder als metaphysisch kontingent oder als notwendige Typen-Identitäten verstanden werden können.

1. Empirical Fundamentalism

The following is a brief introduction to a philosophical program, Empirical Fundamentalism, and its application to the question of how to understand the identification of water with H_2O and similar claims. Empirical Fundamentalism constitutes a general philosophical system, and it can only be defended through an extensive examination of its

philosophia naturalis 47-48 / 2010-11 / 1-2

numerous implications for a broad array of philosophical issues because such frameworks are justified in terms of their utility, not their veracity. Providing an adequate argument that its approach to philosophical problems is superior to extant alternatives would require a multi-volume treatment. Consequently, for brevity, much of the defense of the underlying assumptions of Empirical Fundamentalism is provided elsewhere. All I can present here is a single test case, an illustration of how Empirical Fundamentalism allows us to make sense of reductive identities in a flexible way that avoids ontological profligacy. This justification fits within a broader argument for Empirical Fundamentalism: that it is able to solve many traditionally intractable philosophical problems by translating them into a debate about the character of fundamental reality. To the extent that we can answer or at least bracket the question of how fundamental reality is structured, the remaining philosophical debate must concern what is non-fundamental, and such debates (within the framework of Empirical Fundamentalism) are often definitional quibbles that can be solved on pragmatic grounds.

Empirical Fundamentalism is built on two philosophical pillars: empiricism and fundamentalism. The fundamentalism part, as I will soon elaborate, is a metaphysical framework that employs a certain conception of the difference between fundamental and derivative reality in order to resolve philosophical disputes. Empirical Fundamentalism declares that the distinction between fundamental and derivative should be the central focus of metaphysics. There is a long history of similar distinctions: the classic reality and appearance dichotomy, Boyle's and Locke's primary and secondary qualities, Sellars' scientific and manifest image, as well as the familiar distinction between objective and subjective. The Empirical Fundamentalist judges such distinctions to be suboptimal for understanding reality and instead seeks to enthrone the fundamental/derivative distinction as the new monarch of metaphysics.

The version of empiricism invoked by Empirical Fundamentalism adopts a thoroughly naturalistic approach to fundamental reality. It involves, first of all, a flexible conception of what should count as empirically accessible: taking for granted (at least initially) a common sense approach towards observability and then refining the scope of the empirical wherever needed. Empirical Fundamentalism is not committed to an extreme phenomenalist form of empiricism but is empirical

in the same way that practicing scientists see themselves as providing theories of empirical phenomena.

Second, Empirical Fundamentalism operationalizes its conception of fundamental reality through the hypothesis that our best guess about the components of fundamental reality comes by way of a global abduction. Fundamental reality, insofar as we can know it, is taken to match that model of fundamental reality which best accounts for all empirical phenomena. Underdetermination, of course, sets limits on the precision such an inferential technique can deliver, and Empirical Fundamentalism is compatible with the hypothesis that fundamental reality corresponds to a class of models that are empirically adequate and are equivalent with regard to their consequences for empirical phenomena so that the problematic inference from empirical adequacy to truth is somewhat mitigated.

Third, Empirical Fundamentalism is committed to empirical analysis as its method of conceptual analysis. Conceptual analysis is necessary for providing linkage between our conception of reality and reality itself. As discussed in Kutach (2010, 2011), empirical analysis differs from orthodox conceptual analysis primarily by rejecting the dogma that a conceptual analysis is deficient if it conflicts with common sense intuitions or a priori truths. Instead, empirical analysis adopts the well-entrenched scientific approach towards conceptual architecture by optimizing concepts in whatever ways enhance understanding rather than insisting that metaphysical pronouncements are inferior when they mismatch naive preconceptions.

Having noted the role of empiricism, I will concentrate hereafter on the fundamentalism part of Empirical Fundamentalism.

My application of Empirical Fundamentalism to reductive identities can be interpreted as a competitor to the more familiar approach advocated by Frank Jackson (1998). Although I share the goal of clarifying a target vision of reality where the aspects of the world described by fundamental physics constitute most (if not all) of a supervenience base for all reality, there are key respects in which our conception of the target differs. The points of disagreement, I suspect, stem primarily from a difference in perspective regarding the prominence that language and psychology should have in such an account. I think most contemporary approaches towards metaphysics have been hampered by the persistent influence of the century-old linguistic turn in philosophy, including

its various incarnations in logical positivism, ordinary language philosophy, and the Canberra Plan. My account, by contrast, attempts to engineer metaphysical concepts without hewing closely to linguistic or cognitive structures and without adhering to theories of reference or truth or intentionality that are adapted to the study of human language and thought. Instead, I will attempt to formulate the familiar reductive picture of reality using conceptual structures more closely resembling those used in physics.

I will initiate discussion of the details by clarifying the distinction between fundamental and derivative that plays the starring role in Empirical Fundamentalism. Then, using the example of kinetic energy as a derivative quantity in classical mechanics, I will clarify how derivative existents can bear a certain reductive relation to fundamental reality. Along the way, I will provide three brief suggestive arguments for Empirical Fundamentalism: that it provides a useful model of modality, that it helps to illuminate the disutility of many a priori arguments, and that it helps to explain away many metaphysical disputes. Finally, I will construct some additional theoretical machinery to help formulate a model of derivative properties. Two ways of defining derivative properties—what I will later describe as an unfocused way and a focus-fuzzed way—allow us to make sense of how the relation between water and H_2O can constitute a type identity.

2. Fundamental and Derivative Reality

Most people, I think, have some intuitive grasp of the difference between fundamental and derivative. In order to direct the reader's attention towards the particular form of the distinction employed in Empirical Fundamentalism, I will list a few guiding principles and then describe how we can think of kinetic energy as a derivative property that reduces (in a sense I will eventually clarify) to fundamental attributes like mass and relative speed. (An attribute is a property or relation, broadly construed.)

Perhaps the easiest way to get a grip on the fundamental and derivative is to start by thinking about reality in a rather naive way. Just consider everything that exists, including all objects, properties, relations, substances and whatever else you think needs to be included. The totality

of existents, including all their relations with each other is what we call 'reality'. Then, we can think of reality as subdivided into exactly two parts, fundamental and derivative. 'Fundamental' and 'derivative' are at this point placeholders for a distinction that one can elaborate by filling in with a description of the conceptual role that 'fundamental' plays. On a first pass, it is convenient to operate under the regimentation that every existent is either definitely fundamental or definitely derivative and that these are mutually exclusive categories. Afterwards, one can take up the project of characterizing how the boundary between the fundamental and derivative and between the existent and non-existent could be indeterminate.

The following principles capture several constitutive features of fundamentality.

(1) Fundamental reality is as determinate as reality ever gets.
(2) Fundamental reality is consistent.
(3) The way things are fundamentally is the way things *really* are.
(4) Fundamental reality is the only real basis for how things stand derivatively.

A fifth principle one could entertain is that relatively little of reality is fundamental. Certainly, many prominent speculations about fundamental reality assert that it consists of a relatively sparse structure. Perhaps fundamental reality is just some atoms bouncing around in the void. Perhaps it is merely a single conscious being with temporally ordered mental states. By and large, sparse theories of fundamental reality make for more interesting metaphysical hypotheses, but it is best to avoid incorporating a desire for a parsimonious model of fundamental reality as a constraint on what it is for something to be fundamental. Instead, it is better to think of this principle as a truth that makes it especially useful to employ the distinction between fundamental and derivative.

Some other prominent hypotheses about fundamental reality are also best excluded from the conception of fundamentality. For one, fundamentality is often associated with a so-called fundamental *level,* which suggests something like the Putnam-Oppenheim-Kemelny (1958) layer cake model of the unity of science where there are different theoretical levels – for example, ecological, biological, chemical – each of which (they hope) reduces to the level directly beneath it. On the Empiri-

cal Fundamentalist conception of reality, the ontological distinction between fundamental and derivative reality is inherently binary and rules out the possibility of multiple levels that bear to each other the same kind of relation that holds between fundamental and derivative. To avoid confusion, I advise not thinking of fundamental and derivative reality as levels.

Fundamentality is also sometimes associated with smallness and in particular the empirical hypothesis that as one focuses at ever smaller distance scales, one reaches some scale beyond which reality has no interesting further structure. This is another thesis best separated from fundamentality because ceteris paribus it is better to insulate the features that motivate a notion of the fundamental from the implementation details. The same goes for requirements that what is fundamental be metaphysically simple or be composed of only localizable property-instances; it is better to allow that fundamental entities can have complexity, can consist of parts, and can be non-local.

I will now attempt to specify the constitutive features of fundamentality in terms of the example of kinetic energy in classical mechanics, which will serve as a model of reduction.

3. The Kinetic Energy Example

The theory of classical mechanics is a scheme for modeling how material bodies move in accordance with force laws. I will focus on a specific interpretation of classical mechanics, *N*, whose purpose is to clarify ontological commitments. Other interpretations of classical mechanics exist, but it is not my aim here to settle disputes in the philosophy of physics or to represent classical mechanics as it was understood by its inventors.

The ingredients of *N* include (by stipulation) a classical spacetime inhabited by corpuscles bearing intrinsic properties like mass and charge. A corpuscle is a point particle; it has an identity through time and occupies a single point of space at any given moment, so that its history over any span of time is a smooth time-like path in spacetime. Corpuscles in classical mechanics bounce around according to exceptionless laws where each corpuscle's acceleration is a relatively simple mathematical function of fundamental attributes, for example the inverse-square

law of gravity and some sort of short-range repulsive interaction. In summary, *N* posits the following structures: a Galilean spacetime, corpuscles with charge and mass properties, a distance relation between any two corpuscles at any given time, a relative speed relation between any two corpuscles at any given time, and a dynamical law governing how these fundamental attributes evolve over time. A model of *N* consists of the laws as well as a full arrangement of the allowed entities and attributes throughout spacetime.

Though we know *N* is an incorrect theory, it is convenient to consider how we ought to think about reality under the pretense that the actual world perfectly matches one of *N*'s models. Having adopted *N* as a surrogate for a complete correct theory of fundamental reality, we can distinguish between fundamental and derivative. The corpuscles and spacetime are fundamental entities, their relative distances and speeds are fundamental relations, their masses and charges are fundamental properties, and the cited laws governing them are fundamental laws. Noises, patience, and asset forfeitures, by contrast, are not fundamental because they do not appear as elements in the model nor do the laws of the simple theory make any special use of them. Because noises, patience, and asset forfeitures exist and are non-fundamental, they are derivative existents.

In more generality, we can think of fundamental reality as a system of magnitudes and their structural relations, including laws that constrain the complete layout of magnitudes. Once we have adopted some particular specification of these magnitudes and structures as a complete specification of fundamental reality, we can construe derivative existents simply as existents that are not part of fundamental reality, quantities and attributes and entities that are unspecified. Unfortunately, ordinary language and much of extant philosophical terminology is too imprecise or contested to communicate clearly the kind of ontological distinction posited by Empirical Fundamentalism, so a bit of elaboration is required on the topic of derivative reality.

I advise adoption of the following sufficient condition: a quantity is derivative if its magnitude requires the specification of quantities that are not a part of fundamental reality. By definition, the kinetic energy of any given corpuscle is one-half its mass times its speed squared, $1/2\ mv^2$. But there is nothing in *N* that defines a given corpuscle's absolute speed; a corpuscle's speed is defined only relative to other corpuscles. How-

ever, if we choose some reference frame and stipulate that it counts as the standard for being at rest, we can say that a corpuscle's speed is its speed relative to this rest frame. Then, because we can associate a unique speed with each corpuscle, there will be a particular value for the corpuscle's kinetic energy. The kinetic energy of a corpuscle is an example of a derivative quantity because there is nothing in any model of N that corresponds to a unique correct value for the kinetic energy unless we augment the model with a parameter that doesn't correspond to anything in fundamental reality, namely this stipulation of what counts as at rest. Once we make a choice of rest, the fundamental magnitudes fix the kinetic energy of every corpuscle. The total kinetic energy is also thereby fixed because it is just the sum of the individual kinetic energies.

Whenever a parameter used for describing reality does not have a unique correct assignment given how fundamental reality is structured, let us say that it is *fundamentally arbitrary.* A choice of rest is an example of a parameter that is fundamentally arbitrary. More generally, reference frames and coordinate systems are fundamentally arbitrary.

There are several justifications for treating kinetic energy as derivative rather than fundamental. For one, we already have fundamental laws in classical mechanics governing the motions of particles, and if there were some brute (fundamental) fact about precisely how much kinetic energy existed, it would play no essential role in the temporal development of the physics. (It is possible to formulate classical mechanics so that energy plays a starring role in the temporal development, but N grants kinetic energy no special status.) Another reason to think of kinetic energy as derivative is that if there were a brute (fundamental) fact about the precise quantity of kinetic energy in the world, we would have no epistemic access to its value. A third reason is that there is no scientific account of anything that would be defective in any way if we treated kinetic energy as derivative, nor would any scientific account be improved by treating it as fundamental. These kinds of considerations are standard in scientific practice and provide a practical grip on why we construe some quantities as fundamental and others as derivative. If we try to allocate various attributes to the categories of fundamental and derivative using the methods of science, we have good reasons for keeping the fundamental ontology fairly restricted. Ceteris paribus, a sparser theory of fundamental reality can provide more reductive explanations,

posit fewer epistemically inaccessible facts, posit fewer quantities that fail to integrate well with the rest of the fundamental quantities, etc. Although these criteria are not sacred, it is reasonable to treat kinetic energy as metaphysically derivative, and the discussion from here on will do so.

4. Fundamentality

In this section, I will use the kinetic energy example to clarify the constitutive principles defining the notion of fundamentality that serves as the foundation of Empirical Fundamentalism.

(1) The principle that the way things are fundamentally is as determinate as reality ever gets is illustrated well by the kinetic energy example. The derivative is at least as indeterminate as the fundamental in the sense that we had to supplement the fundamental attributes of *N* with a fundamentally arbitrary parameter in order to get a definite value for the kinetic energy. Put simply, no specific amount of kinetic energy is fixed by fundamental reality even though all the fundamental attributes are absolutely precisely defined. (This principle does not rule out the possibility that fundamental reality includes some sort of ontic vagueness.)

(2) To say that fundamental reality is consistent is to say it obeys a metaphysical correlate of the law of non-contradiction.[1] Derivative reality, by contrast, is subject to a more permissive scheme of managed inconsistency where certain inconsistencies can be tolerated if there is a suitable scheme for blocking any troublesome logical implications. Because the details of how we should understand this second principle do not bear directly on the topic of reduction, I will forgo any further discussion of it here.

(3) The principle that the way things are fundamentally is the way things *really* are is intended to express the relationship between fundamental reality and ontology. I will attempt to describe the ontological difference between fundamental in several ways in order to mitigate some of the confusion that is generated by the variety of interpretations that could be given to the terms 'real' and 'exists'.

(a) Empirical Fundamentalism instructs us to think of the actual world as fundamental reality. The actual world does not consist of eve-

rything that is the case. It is not equivalent to the totality of propositions that are true of the actual world, nor does it consist of all states of affairs or all facts. Instead, the actual world is just the one fundamental reality and does not include any derivative existents as components or parts or constituents.

To explore this hypothesis in more detail, it helps to examine the Empirical Fundamentalist's conception of possible worlds:

A metaphysically possible world is a logically possible fundamental reality.

The function of the word 'logically' here is merely to signify that the operative notion of possibility is entirely unrestricted. Incoherent or inconsistent specifications of a fundamental reality will fail to refer to any possible worlds, but any coherent, consistent description of how fundamental reality could be will correspond to a metaphysically possible world. An important qualification to this principle is that if a description *W* of a possible fundamental reality is based on how the one actual fundamental reality is structured – for example, a possible world just like the actual world but twice as big – then there will exist a metaphysically possible world corresponding to *W* only if the actual world is suitable for such an alteration. It is arguably coherent to double the size of the universe when you disregard the relevant physics, but if the true structure of fundamental does not permit a sensible doubling, there will be no corresponding possible world. This feature suffices to block the general inference from conceivability to possibility. In particular, natural kind terms and words like 'zombie' incorporate an implicit reference to actuality that makes them untrustworthy predicates for describing a genuine possible world.

It is beneficial that in Empirical Fundamentalism, the set of metaphysically possible worlds is not a proper subset of the set of all possible worlds. If a possibility is cogent enough to count as a world at all, it is a metaphysically possible world. Having possible worlds that are not metaphysically possible would make it unclear how we could ever gain rational access to the boundary between metaphysical possibilities and metaphysical impossibilities. We have a workable though imperfect practical grip on the difference between nomological possibilities and nomological impossibilities by way of our standards for evaluating scientific theories. But if there were some dispute about whether a cer-

tain conceivable but nomologically impossible world is metaphysically possible, how would we be able to decide rationally? As discussed by Leeds (2001, p. 172–173), empirical evidence would be of dubious value because the world under consideration is nomologically impossible. Conceptual and logical resources would be of dubious value because by hypothesis the boundary we seek is a further division among possibilities that are already accepted as coherent and logically possible. I have no conclusive argument that such a model of metaphysical possibility is unworkable, but it is a mark in favor of the model of metaphysical possibility employed in Empirical Fundamentalism that it does not suffer from this liability.

For an illustration of how to individuate possible worlds and thus an illustration of what a world consists of, consider the following two possible worlds. Let w be a model of N that is superficially just like the actual world. As detailed previously, w consists of a Galilean spacetime with a bunch of infinitely long corpuscle world lines with mass properties, charge properties, distance and relative speed relations between every pair of corpuscles at every moment and a fundamental law that governs how the state at one time evolves over time. It contains nothing else. Let w^- be just like w except with all the relative speeds excluded. The mere fact that w differs from w^- solely in virtue of w's including the relative speeds suffices for w and w^- to count as distinct possible worlds. Notice that because w^- has the same spacetime structure and the same fundamental distance relations between every pair of corpuscles at any given time, the fundamental attributes of w^- entail the relative speeds at all times. That is, w^- has all the resources needed to specify w and no information that goes beyond what is specified in w. The only difference between them is that relative speeds are fundamental in w but are derivative in w^-. This example illustrates that there is no closure principle associated with being fundamental. The fact that all relative speeds in are logically implied by the structural relations of w^- is not sufficient to count these relative speeds as fundamental.

One of the central motivations for defining a metaphysically possible world as a logically possible fundamental reality is to ensure that the relation between fundamental and derivative is not part of actuality. Some competitors to Empirical Fundamentalism build the relation between fundamental and derivative into the structure of the actual world. In such models, what is fundamental and what is derivative are

both components of actuality and what distinguishes them is some metaphysical relation that is also a component of the actual world. For example, one might postulate that some parts of actuality are linked to one another by the grounded-by relation (Audi 2007). Or one might say that some parts bear a relation of ontological priority (Cameron 2008, Paseau 2009) to other parts. One could postulate in-virtue-of relations or realization relations as metaphysically robust elements of the actual world. Empirical Fundamentalism opposes all such devices for characterizing the relation between fundamental and derivative and instead holds the following: (1) The actual world and all of its parts are fundamental, and nothing else is fundamental. (2) Derivative existents and any relation they bear to fundamental existents are not part of the actual world. Truthful statements about derivative existents (and any linguistic or cognitive references to them) are vindicated not because there is something in actuality that precisely corresponds to them but because of the utility of certain ways of abstracting away from fundamental reality.

(b) Existence in Empirical Fundamentalism can be understood in terms of a tripartite distinction between fundamental existence, derivative existence and non-existence. Our ordinary talk of 'real' tracks the difference between existence (whether fundamental or derivative) versus non-existence. By contrast, debates about realism and anti-realism, according to Empirical Fundamentalism, ought to track the difference between fundamental existence versus derivative existence or non-existence. I will now briefly sketch how this tripartite distinction allows the Empirical Fundamentalist to dissolve a debate concerning the metaphysical status of colors. I hypothesize that similar dissolutions can be provided for many other philosophical squabbles. Further examples of this sort would bolster the case for Empirical Fundamentalism.

Consider whether colors exist. C. L. Hardin (1988, pp. 111–112) argues for a version of color eliminativism, a denial of the existence of colors. He does so on the grounds that no existent plays the constitutive role of color well enough to deserve the label. The platitudes characterizing what it is for color to exist include principles that are in tension with one another, for example that the surface colors of objects exist regardless of whether any creatures have visual abilities and that orange is more similar to red and yellow than it is to blue and green. Most other philosophers of color disagree by claiming that colors exist.

This disagreement can be adjudicated by first recognizing some common ground. Almost everyone in this debate agrees that there is good scientific reason to believe that color is not a fundamental attribute.[2] If that is correct, then the debate only concerns whether colors should count as derivative existents rather than non-existents.

According to Empirical Fundamentalism, to put it one way, a possible existent *X* is a derivative existent if and only if *X* is not fundamental and fundamental reality is such that reference to *X* is handy and not too misleading. To put it another way, *X* is derivative if and only if *X* is not fundamental and *X* is a useful abstraction from fundamental reality. Much more deserves to be said about derivative existence, but I have deliberately phrased these necessary and sufficient conditions in a non-technical (and arguably sloppy) manner because it is a tenet of Empirical Fundamentalism that the distinction between existence and non-existence has little metaphysical significance and that there is no need for a metaphysical scheme to clarify a precise boundary between derivative existence and non-existence. Although it may strike the traditional metaphysician as heresy to declare that there is no deep difference between that which exists and that which does not, the Empirical Fundamentalist embraces the role of iconoclast by claiming that the important ontological difference lies instead between that which exists fundamentally and that which does not.

Applying this to our mundane attributions of color, we are justified in setting aside philosophical niceties and using a very lax standard whereby color exists merely in virtue of fundamental reality being such that our talk of color is useful for getting along in the world. Because the platitudes that constitute our conception of color prove to be useful rules of thumb – objects are usually perceived as near enough the same color by most people in most relevant circumstances and so on – we should not be too picky about the coherence of such principles and just accept the utility of the color platitudes as reason enough to accept that color exists. When discussing color in the context of metaphysical debate, however, it is permissible to adopt stricter standards. The boundary between derivative existence and non-existence is meant to track any raising or lowering of our standards for how well the actual world vindicates the constitutive platitudes for color.

So the dispute over the existence of colors becomes, in the Empirical Fundamentalist framework, a merely pragmatic squabble about how

strictly the various platitudes concerning color should be interpreted in order for colors to exist derivatively. Under lax standards, color exists derivatively because 'color' is a handy and not too misleading term. Under very tight standards, Hardin is arguably correct that color does not exist because there is no simple non-trivial way to abstract away from fundamental reality to arrive at a single quantity that simultaneously satisfies all the platitudes constitutive of color.

(c) Derivative existence is the kind of existence that is adequate for securing the legitimacy of cognitive or linguistic reference whereas fundamental existence is the kind of existence needed for ontological significance. For example, it is irrational to believe that Eve owns a coat that does not exist, but it is perfectly reasonable to believe that Eve owns a non-fundamental coat.

This difference between existence and fundamental existence can play an important role in debunking many a priori arguments that attempt to establish what fundamental reality must be like (beyond what is analytic or true by stipulation). The Cartesian cogito, for example, can be re-imagined as an argument from the premise, "Any thought with the content 'thinking does not exist' is necessarily an existent whose content is false", to the conclusion, "Thought exists". Such an argument can be attacked from a number of directions, but its key deficiency from the perspective of Empirical Fundamentalism is that even if the argument were considered successful in establishing its conclusion, it would only demonstrate that thought exists, not that thought exists fundamentally. Thus, this argument does not motivate the hypothesis that thought is an attribute of an enduring soul, nor does it make any progress in attacking physicalism or the more narrow claim that all thought exists in virtue of the behavior of appropriate brains in appropriate physical environments.

It is easy to apply the same analysis to refute many other a priori arguments that attempt to establish conclusions about how fundamental reality is structured. Such demonstrations, I believe, count in favor of the tripartite distinction between fundamental, derivative, and nonexistent.

(4) The fourth constitutive principle of fundamentality is that fundamental reality is the only real basis for how things stand derivatively. It is intended to serve as something very close to a claim that derivative reality supervenes on fundamental reality. I hesitate to claim that it is a

bona fide supervenience claim because supervenience claims are often understood in terms of entailment, and it proves critical for the scheme I am proposing to avoid implying that fundamental reality *by itself* completely fixes the character of derivative reality. In order to clarify how derivative reality depends on fundamental reality, I will first introduce a new kind of reduction, and second show how it vindicates claims of supervenience, and third discuss how the supervenience-like relation between the derivative and fundamental can hold generally, even without an explicit reduction.

It is important to note although this fourth principle partly constitutes what it means to play the role of fundamental reality, it also presupposes the ontological distinction provided by principle (3).

Empirical Fundamentalism employs a proprietary notion of reduction called 'abstreduction'.[3] Abstreduction is a form of reduction that operates by explicitly representing derivative existents as abstractions from fundamental reality. In order to illustrate abstreduction, I will draw attention to a critical feature of the kinetic energy example. Remember that in order to derive any specific value for the amount of kinetic energy in a system, one needs the fundamentally arbitrary choice of rest. A complete specification of the fundamental attributes of classical mechanics does not by itself suffice for any particular value of kinetic energy. So, how things are situated fundamentally does not fix how much kinetic energy there is. Yet, given any choice of rest, every detail about the distribution of kinetic energy is fixed. So, there exists a complete conditional characterization of kinetic energy, a complete set of conditionals of the form, "If choice of rest R is made and the the state of fundamental reality at time t is S (t), the total kinetic energy is $K(R, S(t))$." By saying, "Fundamental reality is the only real basis for how things stand derivatively", I intend to communicate that the only thing besides fundamental reality that bears on how things stand derivatively are choices about how to abstract away from fundamental reality, choices that do not count as constituents of actuality.

Any parameter-dependent entailment from fundamental to derivative constitutes an *abstreduction.* In general, a derivative quantity q *abstreduces* to fundamental reality if and only if there exists a (possibly empty) set of fundamentally arbitrary parameters such that specifying those parameters is sufficient (in conjunction with a specification of fundamental reality) for q. Abstreduction having been defined so

broadly, any non-fundamental quantity can abstreduce to fundamental reality merely by contriving an ad hoc parameter, but non-trivial cases of abstreduction are those where the employed parameter is reasonably general, such as a choice of coordinate system, a choice of spacetime region, or a collection of possible fundamental property-instances, and where the resulting derivative quantity has some utility.

One can observe that abstreduction supports a form of supervenience by considering two possible arrangements of particles, a_1 and a_2. A reasonable supervenience claim is that a difference in the kinetic energy of a_1 and a_2 implies that a_1 and a_2 differ fundamentally, especially with regard to the number of corpuscles, their masses, or relative speeds. Because it does not make sense to compare the (derivative) kinetic energy of a_1 and a_2 without a choice of rest for each arrangement and because that choice is fundamentally arbitrary, a difference in the kinetic energy of a_1 and a_2 could result either from a_1 and a_2 being different fundamentally or from their having different standards of rest. In some cases, there are resources available such that a choice of rest for one arrangement will fix a choice of rest for the other, but in full generality, the only way to ensure that the fundamentally identical a_1 and a_2 are identical insofar as derivative quantities like kinetic energy are concerned is to impose a stipulation that (for the purposes of evaluating supervenience) when two arrangements are identical fundamentally, any conventions employed for abstracting away from one must be applied to the other. That suffices to ensure that whenever a derivative quantity abstreduces to fundamental reality, it supervenes on fundamental reality.

Explicit abstreductions may not be available for every derivative existent, and yet it may still be reasonable to believe that supervenience holds. For example, physicalists maintain that whether a certain government is communist supervenes on the complete physical history and laws of the actual world. Although no one is able to supply an explicit set of parameters that (together with the totality of fundamental physics) implies what is communist and what is not, we can still reasonably maintain that if two possible worlds that obey physicalism are physically the same, then whatever principles we use to evaluate the status of a given government as communist need to be applied to both worlds equally. And that is enough to ensure that both worlds agree on which existents are communist.

One of the benefits of construing supervenience in this way is that

supervenience by itself is inadequate for accurately characterizing the kind of relationship that arguably holds between kinetic energy and the masses and relative speeds of fundamental corpuscles. For one thing, A's supervening on *B* is in general compatible with a lack of any asymmetry in the relation between *A* and B. Abstreduction, however, presupposes an essential ontological asymmetry because abstreduction by definition only exists between a derivative quantity and fundamental reality (or some part of fundamental reality), and it is part of the Empirical Fundamentalist framework that fundamental existents are ontologically privileged over derivative existents. Thus, the sort of supervenience that holds in virtue of abstraction is inherently asymmetrical.

Another shortcoming of supervenience as a tool for representing how kinetic energy depends on the fundamental attributes is that supervenience does not help to represent that arbitrarily small changes in the fundamental arrangement of particles results in arbitrarily small changes in the kinetic energy. In the literature on the mind-body problem, this deficiency of supervenience was identified by Kim (1993) as the "lone ammonium molecule" problem. Supervenience alone, Kim noted, does not prevent the hypothetical addition of a single ammonium molecule to one of Saturn's rings from making a radical difference to Earthly mental states. Although we might have good scientific reasons to question this particular claim of counterfactual independence, the important lesson to draw from Kim's observation is that it is a mark of a good physicalist account of mentality that mental quantities vary in accordance with physical quantities in a way that is at least consistent with the background beliefs that make physicalism plausible. It is possible to concoct any number of crazy functions to represent how the severity of someone's pain depends on the arrangement of all the atoms in the universe. Some of these functions have a person's degree of pain varying greatly as distant atoms are shifted slightly in ways that make a negligibly small difference to the functional behavior of the person's brain. Even though such a pain-function would be compatible with the supervenience of the mental on the physical, it would undermine the reasonableness of our judgments about how much pain other people feel. The supervenience of the mental on the physical ought to fit neatly into a broader (if only dimly seen) account of how mental states vary as a function of physical states. Abstreduction helps to provide such a fit because the resulting supervenience is a consequence of a broader account of how derivative

magnitudes vary as a function of fundamental magnitudes (holding fixed all fundamentally arbitrary parameters).

5. Reductive Identities

At this point, enough of the central tenets of Empirical Fundamentalism have been sketched to permit the formulation of a scheme for how water relates to H_2O. In order to fill in the details, a fragment of a theory of reference is required so that the terms 'water' and 'H_2O' can be related to reality. My goal here is to construct a model rich enough to make sense of commonly held opinions about how 'water' is to be understood, especially Putnam's (1975) observation that if we were to discover a substance XYZ somewhere else in the universe that behaves like water but is chemically very much unlike H_2O, it would not count as water. The model I will be constructing says in effect that there are two ways someone could interpret 'water'. One kind of intension corresponds to what I call an *unfocused* derivative property, where anything that behaves superficially like water counts as water. The other kind of intension corresponds to what I call a *focus-fuzzed* derivative property, where only the stuff sufficiently similar to local instances of watery stuff counts as water. What Putnam in effect pointed out is that, contingently, our implicit concept of water more closely matches the second kind of intension. My conclusion is that in principle we can say anything we need to say about water in either the unfocused or the focus-fuzzed way. In the unfocused way, water is not the same as H_2O. In the focus-fuzzed way, they can be equated. Nature itself does not privilege one construal over the other; that we employ the focus-fuzzed construal is a result of its convenience and historical accident.

In order to present my model, I will mention concepts, intensions, and referents in order to relate the structures I define to familiar philosophical terms, but nothing in the model presupposes a prior notion of intentional content or requires that content be considered fundamental. Contents exist, of course, but I strongly suspect they are derivative like most everything else.

By formulating the indexical character of natural kind terms like 'water' in terms of fundamental reality and abstreduction, I will show how references to water can make sense in a world that is fundamen-

tally just a bunch of physics without any fundamental water. It will thus solve the so-called location problem for water. Unlike the Jackson (1998) methodology, which requires one to locate water by showing how truths about water are *entailed* a priori from truths about fundamental reality, Empirical Fundamentalism allows one to locate water by providing an account of how the concept of water is a *useful* device for abstracting away from fundamental reality. Jackson requires that in order for an entity to be allowed in reality as a bona fide existent, it must achieve "entry by entailment". Empirical Fundamentalism only requires entry by utility.

Before discussing what reductive identities amount to, several preliminaries are needed. Let us restrict discussion to *concrete* derivative entities and ignore non-concrete entities like numbers and algorithms. For a (concrete) derivative entity to exist is for fundamental reality to include an instance of that entity. The derivative entity itself can be thought of as merely some set of possible instances. For example, we can think of a giraffe as a derivative entity by associating it with a set, *G,* of metaphysically possible instances, the ones we intuitively think of as ways a giraffe could be instantiated. An instance is by stipulation always fundamental. If some part of fundamental reality – say a complete specification *g* of all the fields and corpuscles in some spacetime region – is a member of *G,* we say that *g instantiates* a giraffe (as precisified by G). Of course, what we have in mind when we think of giraffes and what we refer to when we refer to a giraffe do not perfectly match up with any particular precisification, G, but everything that needs to be said about the metaphysics of giraffes can arguably be cashed out in terms of its precisifications.

An instance can be defined formally in several ways. In order to cut to the chase, I will only present a model of instances rich enough to discuss reductive identities. So, for current purposes, let an instance be defined as an ordered pair consisting of a fundamental event and a set of fundamental laws. A fundamental event is a spacetime region together with a full specification of all the fundamental attributes throughout that region. The fundamental laws specify not only the rules for how fundamental attributes evolve over time but also specify any fundamental constants and what kinds of fundamental attributes are allowed. This model of instances can be easily extended to handle relational concepts,

but my discussion will focus on its application to rather simple derivative properties like being-water or being-a-table.

A derivative property is by stipulation a set of instances. For illustration, consider the derivative property being-a-giraffe. Some of its instances are a specification of quarks and electrons and electromagnetic fields arranged giraffe-wise somewhere in a 4m cube of space lasting a tenth of a second together with a set of dynamical laws that govern the evolution of these particles and fields. Other instances instantiate the four fundamental elements – fire, water, air, and earth – in appropriate combinations to make a giraffe-ish material body that behaves like a giraffe.

There are no restrictions whatsoever on which sets of instances count as a derivative property. For example, the predicate "frog or carburetor" can be precisifed as a derivative property by specifying an appropriate set of instances, the set that includes instances of what we intuitively take to be frogs as well as instances of carburetors. I emphasize this example because philosophers commonly restrict use of the word 'property' in order to block the inference from the existence of some meaningful predicate to the existence of a corresponding property. Derivative properties, however, are ontologically innocuous, and for the sake of simplicity it is better not to impose any restrictions. Derivative properties that are gerrymandered or consist of unduly heterogeneous instances typically have little utility, so we can set them aside merely on pragmatic grounds.

In order to keep the discussion manageable, I will focus on possible worlds that have a four dimensional spacetime as their sole container for all fundamental fields and corpuscles and that have fundamental laws governing the temporal development of these fundamental attributes such that the complete state of the world at any one time fixes objective probabilities for all later states of the world. Thus, I will be ignoring how water and tables and giraffes can exist in worlds with two dimensional time, or where physical objects are fundamentally a perceptual state of God, or anything else too outlandish. Furthermore, from here on, I will assume that fundamental reality resembles paradigm models of fundamental physics, at least so that the fundamental ontology does not include properties like being-a-giraffe.

For the simple cases under discussion, talk of derivative properties and concept intensions are interchangeable. Any precisification of the

intension of the concept of a giraffe is a set of those possible instances that count as a giraffe, and that in turn just *is* (a precisification of) the derivative property being-a-giraffe.

5.1. Unfocused Properties

One important kind of concept is the functional concept, which can be associated with derivative properties through what I call a "test". For reasons that will soon become clear, the derivative properties associated with purely functional concepts will be a special case of what I call "unfocused properties". Our ordinary concept of a table can be construed in terms of its function in the sense that a table is anything that behaves in a table-like way. (Remember that it is not important in Empirical Fundamentalism whether thinking of tables in a purely functional way perfectly matches our intuitive concept of a table. It suffices that our concept of a table comes reasonably close to being merely functional.) In order to cash out what it means for something to behave in a table-like way, one can use the following procedure: Start with any spatial region r in a single time slice of spacetime as shown in Fig. 1, and specify some fundamental laws, L. Let e be any r-shaped fundamental event compatible with L. Then consider various possible background conditions, $\overline{B}_i$, occupying the entire space outside r. Each $\overline{B}_i$ is a set of fundamental events with a probability measure over the set to allow it to represent in a fuzzy way what could happen outside r. Now consider what we get when we graft e on to $\overline{B}_i$ by letting $\overline{C}_i$ be just like $\overline{B}_i$ except that each of $\overline{C}_i$'s members has e planted into the r-shaped hole. Thus, each $\overline{C}_i$ represents the precise event e embedded in some fuzzily characterized background field.

The laws, L, we have assumed, provide deterministic or chancy rules for how to evolve a completely specified time slice of physics towards the future. Thus, L is sufficient to propagate each $\overline{C}_i$ throughout the future, thereby establishing probabilities for any future event one chooses to consider.

In order to get an adequate characterization of table-like, one could choose a $\overline{B}_1$ that includes a human who is just about to poke the table with a finger. If some e is such that the resulting $\overline{C}_1$ fixes a very high probability for the finger being blocked at the edge of r, then e has answered one "test question" correctly for being table-like (or more accurately, being disposed to behave like a table). One could choose a

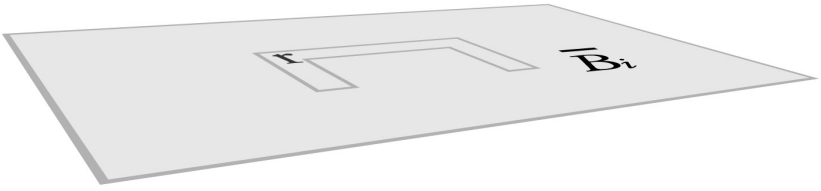


Figure 1. (e, L) instantiates a table at r iff e fixes suitable probabilities under L for certain future effects when e is embedded in various background conditions $\overline{B}_i$.

$\overline{B}_2$ that includes a human attempting to drag the contents of r across a floor. If e is such that the resulting $\overline{C}_2$ fixes a high probability for the material in r retaining its shape as it moves along with the human, then e has gotten another test question right for being table-like. One can formulate arbitrarily many such $\overline{B}_i$, each of which represents a way to probe whether the stuff in region r is likely to behave like a table in some respect. Each such test question can be formalized as an ordered triplet, $(\overline{B}_i, E, p)$, where the E is a coarse-grained description of some possible event located in a spacetime region after $\overline{B}_i$, and p is a probability range. An instance (e, L) is said to answer a test question correctly iff the C_i formed by grafting e onto $\overline{B}_i$ fixes a probability for E in the range p using the laws L. A test consists of a set of test questions together with some function that specifies whether a given instance (e, L) passes the test as a function of which test questions it answers correctly. One might stipulate that passing the test for being table-like requires an instance to answer absolutely all of the test questions correctly, or nine-tenths of them, or perhaps some weighted measure of them. If an instance (e, L) passes the test, that means that e counts as table-like (given the fundamental laws L and the chosen precisification of the functional test for being disposed to act like a table). Finally, the set of all (e, L) that count as table-like is a precisification of the property of being disposed to act like a table. Because the concept of table was interpreted as a purely functional concept, any such derivative property also counts as a precisification of being-a-table.

5.2. Focus-Fuzzed Properties

Concepts that are purely functional can be modeled effectively using tests that result in unfocused derivative properties, but concepts like

'water' require a more sophisticated treatment, what I call "focus-fuzzed derivative properties". The idea behind focus-fuzzing is quite simple. We start with an unfocused derivative property, $water_{uf}$, which corresponds to some precisification of the predicate "being disposed to act like water". Then we formulate a second derivative property, $water_{f}$, a so-called focused property, which contains only those members of $water_{uf}$ that are instantiated in our local environment. The third step is to form the focus-fuzzed derivative property, $water_{ff}$, that includes all the members of $water_{f}$ plus any other members that are suitably similar to the ones in $water_{f}$. I will now construct these three in more detail.

Stage 1: Any unfocused (derivative) property, $water_{uf}$, corresponding to "that which is disposed to behave like water" can be defined functionally just like the table, in terms of a test for watery behavior. Our ordinary concept of water seems to be modeled best when the stipulated test is only for the superficial behavior of water including its liquidity at room temperature, its density of roughly 1g per cc, and its translucency, but excluding its esoteric behavior such as its disposition to produce an explosive combination of gasses when an electric current is passed through it. Nothing in my account of functional properties places any restrictions on the kind of tests that can be used to define $water_{uf}$, so there is nothing wrong with using such a test if one wishes.

Stage 2: In the second stage, we focus $water_{uf}$ into a new set, $water_{f}$, by tossing out any of its members that are not instantiated in a certain prescribed region of the actual world, say the region around Earth. This implies a restriction to the actual fundamental laws, so any members of $water_{uf}$ whose fundamental laws differ from the fundamental laws of the actual world are automatically discarded. One could also impose further rules to discard more instances so that $water_{f}$ represents the *predominant* local substance that behaves watery or perhaps the watery substances that have interacted appropriately with our ancestors. I will set aside these refinements for the sake of brevity, but they are not difficult to incorporate.

For further detail, we can consider two hypothetical chemicals that are instantiated as part of $water_{uf}$. Let us suppose there exists a chemical, XYZ, that is possible according to the actual fundamental laws and passes the chosen test for being disposed to act like water. Let us also postulate a chemical, PDQ, that does not exist in the actual world but passes the test in virtue of alien fundamental laws. So, $water_{uf}$ includes

instances (e_{XYZ}, L_a), where e_{XYZ} is an event that instantiates XYZ and L_a represents the actual fundamental laws, and it also includes members ($e_{PD}Q$, L_n), where $e_{PD}Q$ is an event that instantiates PDQ and L_n represents non-actual laws that help PDQ to behave like water. The point is, $water_f$ automatically excludes all members of $water_{uf}$ that instantiate PDQ because they involve alien fundamental laws. Furthermore, if XYZ does not occur anywhere in the local region of spacetime chosen as the focus region, then all members of $water_{uf}$ that instantiate XYZ are excluded. So, the presence of XYZ on Twin Earth makes no difference as to what counts as an instance of $water_f$.

Stage 3: In the third stage, we fuzz $water_f$, by defining $water_{ff}$ to include all the members of $water_f$ plus any other instances that we choose to count as similar enough to the members of $water_f$. The intuitive reason for adding this fuzzing stage is that there is no reason to think that $water_f$ includes absolutely every microscopic configuration of water. $Water_{ff}$ includes all the instances one gets by taking an instance from $water_f$ and shifting its electrons and quarks a bit and altering it in other microscopic ways until $water_{ff}$ forms a set whose members we group together as being suitably similar.

There is no uniquely correct way to fuzz a set. Just like kinetic energy, where we are free to pick any standard of rest we like for purposes of convenience and scientific utility, we can fuzz a derivative property as much or as little as we like. A few general rules of thumb, though, guide the implicit conventions for fuzzing we tend to employ so that focus-fuzzing has substantial utility.

– The fuzzings that are most convenient as conceptual devices for understanding the behavior of ordinary macroscopic objects tend to be those that preserve the macroscopic character of an object but permit whatever microscopic variations are consistent with the macroscopic character. $Water_{ff}$ should include all instances formed by taking a member of $water_f$ and shifting some hydrogen nuclei to other locations near their oxygen nuclei, but $water_{ff}$ should not include instances where each oxygen nucleus has two protons shifted out to the neighborhood of the existing hydrogen nuclei because that would in effect convert the H_2O into methane, CH_4.
– $Water_{ff}$ should include instances that differ merely in size and shape. Even though $water_f$ contains no instances of water that occupy a cubic light-year, such cubes of water should count as $water_{ff}$.

– When a linguistic community agrees with Putnam that XYZ is not a form of water, that reveals that the implicit standards for fuzzing in that community are such that fuzzing H_2O samples enough to include XYZ molecules counts as fuzzing too much. Any population that does count XYZ as water is not committing a metaphysical error; they just have more liberal conventions for fuzzing.
– After we have selected a test that identifies some precisification of an unfocused property, it often proves convenient to re-use that test to establish boundaries so that fuzzing cannot include any events outside that boundary. For example, if we decide to fix a borderline for $water_{uf}$ so that samples of very muddy water with a density greater than 1.02 g/cm^3 do not count as watery enough, then when we fuzz the resulting $water_f$, we should exclude samples of H_2O that contain enough dirt contaminants to cross that borderline.
– In other cases, it proves useful to ignore $water_{uf}$'s boundaries. For example, a cubic meter of empty space with a couple of stray H_2O molecules intuitively counts as water even though it does not behave like water superficially. Similarly for snow, ice, steam, clouds, and so on.
– It is reasonable to allow some fuzzing in the laws as well, so that the H_2O in worlds where gravity is slightly stronger still counts as water. Yet, worlds should be excluded if they have substantially different laws or substantially different kinds of fundamental attributes.

6. Water = H_2O

All the resources have now been assembled to make sense of how water can be associated with H_2O. The key observation is that we can make sense of all the empirical phenomena associated with water either by thinking of water in an unfocused way as $water_{uf}$ or in a focus-fuzzed way as $water_{ff}$. Even though ordinary language treats 'water' in a focus-fuzzed way, in an empirical analysis of water, there is no interesting fact of the matter as to which is water.

Consider water as $water_{uf}$. All actual nearby instances of $water_{uf}$ are instances of H_2O, but water cannot be equated with H_2O because XYZ and PDQ are also forms of $water_{uf}$. So, there is no identity holding between composed-of-H_2O and $water_{uf}$.

Consider water as water_{ff}. All members of water_{ff} instantiate H_2O, and the conventions for fuzzing could reasonably be chosen so that all and only instances that count as composed-of-H_2O are included in water_{ff}. If that choice is made, there is a identity between water_{ff} and composed-of-H_2O. This identity furthermore counts as a kind of type identity because derivative properties play the role of types; they are sets of possible instances.

It should be noted that the interesting metaphysical structure is not especially well illuminated by talk of identities. The derivative property water_{ff} can indeed be equated with the derivative property composed-of-H2O, but this identity is largely the result of just choosing the convention for fuzzing water_{f} that results in an identity with composed-of-H_2O. A less permissive convention would have water_{ff} as a proper subset of composed-of-H_2O, and other conventions would have them overlapping but with neither one included in the other. What is not a matter of convention is that all local watery instances, water_{f}, are instances of composed-of-H_2O. The interesting metaphysical structure underlying the natural kind, water, is partly that the fundamental laws and fundamental kinds are such that the only stable configurations of particles that instantiate the functional kind, water_{uf}, are H_2O configurations and partly that if there are any other chemicals that behave superficially like water, they are not around here in any sufficient quantity. With an alien fundamental physics, there could be a continuum of substances or properties with no clear practical boundary between the watery stuff and the golden stuff and the feathery stuff, etc., in which case there would not be a well demarcated water kind. That is the pragmatic contingency that makes it handy to associate 'water' with water_{ff} rather than water_{uf}.

It is often claimed that the identity between water and H_2O holds necessarily. Kripke (1971) famously argued that such relations hold necessarily in virtue of the logical character of identity. However, there is another way to look at it that does not invoke any special features of identity. Some propositions are necessary because they result from semantic devices that convert contingent truths into necessary truths. For example, one can stipulate that 'plake' designates the actual number of tangerines eaten by the reigning Dutch monarch during the year 2033 in the actual world. Suppose that by happenstance, there is exactly one reigning Dutch monarch, and she eats nine tangerines during 2033.

It follows that the proposition *P*, expressed by the statement, "Plake exceeds three," is true in the actual world, w_a. When we evaluate the truth value of *P* from the perspective of some non-actual world, *w*, the definition of 'plake' requires us to find its magnitude by looking at what happens in w_a, not at what happens in *w*, whence plake's magnitude exceeds three. Thus, *P* is true in every possible world. Thus, *P* is a necessary truth despite its incorporating a description whose numerical value is contingent. Notice that there is nothing metaphysically interesting about the necessity of *P*, and it certainly does not demand that we postulate an essence of plakitude in our metaphysical scheme. The necessity arises entirely from a cheap semantic trick that has no bearing on issues of ontology or fundamental reality. Because it was built into the definition of 'plake' that contingencies in the actual world fix its intension, the extension of plake in *w* is independent of *w*'s tangerines and royalty.

The semantic trick that makes "Plake exceeds three" a necessary truth is the very same trick that makes the statement "All instances of $water_{ff}$ are instances of H_2O" a necessary truth. $Water_f$ – because it focuses on the *actual* world – fixes a precisification of water that excludes the way watery stuff is instantiated in alien worlds. $Water_{ff}$ will presumably expand this set to include some instances that have slightly different fundamental laws, but all such instances are still included only because they are suitably similar to the actual instances of $water_f$. How people in alien worlds use the word 'water' and how they drink and what substances play a watery role in such worlds is entirely irrelevant to what instances are included as members of $water_{ff}$. Once the set $water_{ff}$ has been constructed by focus-fuzzing, it remains constant across all possible worlds. Thus, all instances of $water_{ff}$ are *necessarily* instances of composed-of-H_2O. The inference from "In the actual world, anything that is $water_{ff}$ contains H_2O" to "In all possible worlds, anything that is $water_{ff}$ contains H_2O" is a semantic triviality and says nothing about the structure of fundamental reality.

7. The Mind-Brain Identity Theory

The scheme I have presented for relating water to H_2O is quite general and can be applied without alteration to the relation between the mental and the physical. It is true that the derivative property, $water_{uf}$, was

defined functionally in terms of the probabilities that an instance would need to fix for certain effects characteristic of water when conjoined with certain background conditions. However, nothing in the scheme I presented for focus-fuzzing requires a functional characterization of the unfocused derivative property. If there are some aspects of mentality, phenomenality perhaps, that resist a purely functional characterization, so long as there is some available precisification of the mental state, for example a stipulation of what instances count as being-thirsty, one can use that as the unfocussed derivative property that serves as a starting point for focus-fuzzing.

To explore the mind-body problem in more detail, let us introduce some neologisms by saying that a component of fundamental reality is *fphysical* iff it resembles (near enough) something that is uncontroversially an entity, attribute, or law of fundamental physics. For example, all of the following are fphysical: spacetime, electromagnetic fields, corpuscles, super-strings and the eleven-dimensional arenas they inhabit, quantum mechanical configuration states, and the classical inverse-square law of gravitation. If any of the following are components of fundamental reality, they are paradigmatically non-fphysical: Cartesian souls, phenomenal states, volitions, economic stratification, angelic influence, and any fundamental laws that impart a special swerve to particles that compose the brain of a decision-making creature.

Fphysicalism is the thesis that fundamental reality is entirely fphysical. Worlds where fphysicalism holds are worlds composed of nothing non-fphysical. Fphysicalism is a version of physicalism because it implies that the only way something non-physical can exist is derivatively, by being merely an abstraction from a fphysical fundamental reality. Dualism, by contrast, is the hypothesis that fundamental reality is partly physical and partly mental.

Now consider what we should say about thirst under the assumption that fphysicalism is true. Presumably, $\text{thirst}_{\text{uf}}$ contains some instances of brains together with the actual fphysical laws. It should also contain instances of a thirsty Cartesian soul together with fundamental laws that blend the interactions of fphysical properties and fundamentally mental properties. Thus, there is no special relation between $\text{thirst}_{\text{uf}}$ and fphysicality.

But when we focus $\text{thirst}_{\text{uf}}$ to form thirst_{f}, only fphysical instances

will be members of thirst_f because we are operating under the assumption that the actual laws and fundamental properties are all fphysical. Then, when we fuzz thirst_f to form thirst_{ff}, it is arguably reasonable to have a convention for fuzzing where we stick to at least roughly the same kind of fundamental laws and fundamental materials as the actual world, so that we are left with a thirst_{ff}, all of whose instances are purely fphysical. This result allows us to make sense of how thirst (and mentality generally) can be understood as a form of type identity physicalism. If fundamental reality is just a bunch of fundamental physics without any fundamental mentality, then the property we get by focus-fuzzing all metaphysically possible instances of thirst is a property all of whose instances are entirely fphysical. Again, it is trivial that this type identity, if true, counts as a necessary truth.

Of course, we originally come to the mind-body problem without knowing the true nature of fundamental reality. If our interest is in determining whether *thirst* is a physical property, then Empirical Fundamentalism tells us first to work on establishing – as best as we can – whether the better overall model of fundamental reality is one that is entirely fphysical, or one that incorporates a mixture of physical and mental components, or one that includes only mental components, or some other option. Assessing which model of fundamental reality is best involves considering which model provides the superior account of all empirical phenomena. Among other things, one would need to investigate whether the quarks and electrons in people's brains exhibit some unusual motion that is best explained by a fundamental libertarian volition. Also, one would need to consider the delicate issue of whether phenomenal happenings should count as empirical phenomena, and if so, whether phenomenalism or epiphenomenalism or mind-body parallelism would provide a better overall account of the totality of everything that is empirically accessible. These issues are all too contentious to address here, of course. What my account of focus-fuzzing does is to say that if you settle on the hypothesis that fundamental reality is entirely fphysical, then you ought to believe that (1) in the unfocused sense, mental properties are not physical properties but they are contingently everywhere physically instantiated, and (2) in the focus-fuzzed sense, mental properties are necessarily physical properties (given reasonable conventions for fuzzing).

Much more deserves to be said about the relationship between men-

tality and physicality, but owing to space limitations, I can only make two brief additional observations. Because the degree of fuzzing is a fundamentally arbitrary parameter, there is no deep fact of the matter as to whether some mental property – say, the property of understanding Chinese – should be focus-fuzzed into a derivative property that only includes human brains, or instead into a derivative property that is inclusive of computers, or of a guy shuffling papers around based on rules written in English, or of a galaxy-sized wooden contraption with the gross functional behavior of an ordinary Chinese speaker. According to Empirical Fundamentalism, such decisions are to be made on the basis of convenience and utility; there is no fact of the matter to be discovered as to which ones *really* understand Chinese.

This consideration suffices to insulate my account from the multiple realizability argument. The multiple realizability argument tries to attack reductive theories of mind on the ground that the target mental existent can be realized by distinct physical kinds. Multiple realizability, it must be said, is a nearly trivial claim. On any remotely plausible version of physicalism, mental existents will be multiply realizable in the sense that any mental state can be instantiated by microscopically distinct instances. It is uncontroversial that mental states are not sensitive to absolutely every last physical detail. Every distinct instance of a mental state (under physicalism) must be an instance of a distinct physical kind because every instance belongs to the kind that includes itself and nothing else, a simpleton kind.

Advocates of the multiple realizability argument have in mind a less discriminate conception of physical kinds, but within the framework provided by Empirical Fundamentalism, there is no reason to ascribe any metaphysical privilege (any fundamentality) to such kinds. Physical kinds are merely groupings (sets) of physical instances, and they can be as convoluted or gerrymandered or ad hoc as you like while still being physical kinds. A precisification of thirst may not be identifiable with a physical kind that is easy to express in English, but there is no barrier to its being identified with some convoluted set of physical instances. If there is any fact of the matter at all as to what physical instances count as an instance of (some precisification of) thirst, then that set of instances *is* the physical kind. Of course, given my account above, it would be misleading to say thirst is identical with some specific physical kind because there are numerous ways to precisify thirst as $\text{thirst}_{\text{uf}}$, and

it is merely a matter of convention and pragmatics how much fuzzing to use when constructing a focus-fuzzed property, $\text{thirst}_{\text{ff}}$.

8. A Brief Justification

The preceding discussion, I hope, explains how one can make sense of reductive identities, but it is fair to ask why it should count as a good way, or at least as a better way than existing alternatives. As I noted in the introduction, any attempt to evaluate the relative merits of the Empirical Fundamentalist program as a whole is too vast a project to take up here, but within the general guidelines established by Empirical Fundamentalism, there are some advantages to the approach I have taken that can be briefly noted beyond the three suggestive arguments given in section 4.

The primary merit of my account of how water relates to H_2O has been that it treats water and H_2O and their relation as derivative. Rather than having to complicate a model of fundamental reality that already includes a bunch of fundamental physics by having additional components to stand as a referent for 'water' and a referent for 'H_2O' and further relations to connect them appropriately to each other and to the fundamental physics, I have left all these out of the actual world. This more parsimonious model of the structure of the actual world counts as a benefit according to the standard scientific practice of treating an ontologically sparse model as ceteris paribus preferable.

The advantage of a parsimonious account of reductive identities would be worth little if other philosophical issues required tables, giraffes, and water to be treated as ontologically on a par with electrons and spacetime. So, an important factor in the value of the model is whether it relies on any concepts or constructs that require a more heavily populated fundamental reality, for example treating ordinary objects and their causal relations as fundamental. To that end, I intentionally formulated the individuation of various properties without the concept of causation. Instead, I employed the concept of a fundamental law, which is similar to causation in that fundamental laws govern how events at one time are related to events at another, but it is not encumbered with all the baggage that comes along with causation as it is normally understood. This means that my account is in a position to avoid some of

the potential pitfalls that challenge other accounts. Models of realization along the lines of Melnyk (2001) or Shoemaker (2007), for example, reckon ordinary attributes – being-water, being-a-table, and so on – in terms of their causal roles or powers or profiles. Given the notoriously contentious issue of figuring out what causation amounts to, there is at least the potential for such causal-role accounts to be saddled with a bloated ontology.

Another benefit is that the extensive conventionality (or fundamental arbitrariness) built into my account of terms like 'water' helps to explain several familiar aspects of natural kinds. If we bracket issues concerning fundamental kinds and focus solely on derivative kinds, then it is relatively easy for the focus-fuzzing conception of natural kinds to make sense of why the category 'natural kind' exists on a continuum. On the one end, there are clear-cut derivative natural kinds like water and gold, and on the other end there are clear cut derivative artificial kinds like games and tables. The difference can be drawn as follows. A (derivative) natural kind is a category such that it is reasonable to conduct a scientific investigation of its hidden nature. For example, it is prima facie reasonable to investigate the properties of diamonds to see if they share a common hidden nature with coal. By contrast, it is prima facie silly to conduct a scientific investigation of chess to see if it shares a common hidden nature with football. The focus-fuzzing model of natural kinds can make sense of this distinction in terms of our implicit knowledge of our default conventions for fuzzing. Without much reflection, we can recognize that focus-fuzzing the property being-chess is going to result in a set of instances that is far more dependent on our conventions for fuzzing than on the details of how chess matches are locally instantiated. There may be many surprising commonalities among chess and football, but these will inevitably turn out to be uninteresting historical contingencies, such as the hard-to-predict commonality that they are both disliked by Laura Monroe of West Bromwich. Without much reflection, though, we can also recognize that focus-fuzzing that which is disposed to behave like water may well result in some surprising properties that play a prominent role in science.

The focus-fuzzing model of natural kinds also helps to explain derivative kinds that lie between the natural and artificial extremes. For example, biological kinds are located somewhere near the natural end, because before we engage in much zoology, it is plausible that focus-

fuzzing the set of instances that behave like a koala, panda, or grizzly will result in a set whose members share some hidden factors. But once we have a more thorough understanding of all the underlying genetics, their commonalities and differences will appear more like historical contingencies. That is, knowing the vastness of the space of nomologically possible biological diversity, it is plausible that koalas, pandas, and grizzlies exist on a phenotypic continuum and that it is largely just historical happenstance (grounded in the constraints imposed by evolutionary factors) that accounts for the limited range of phenotypic mixtures of koalas, pandas, and grizzlies. Furthermore, current technology makes it difficult to create a continuum of in-between animals. By contrast, it is not largely historical happenstance or technological limitations that make it difficult to discover or manufacture an element that lies halfway between carbon and nitrogen. My remarks here accord with Russell's (1956) and Quine's (1969) observation that as sciences mature, they tend to obviate their natural kinds. A more extensive discussion would point out that mixture kinds like air and clay also lie between the natural and artificial extremes for obvious reasons, and that some artificial kinds like the dollar bill kind have a focus-fuzzing character, but one whose only hidden structure is its various anti-counterfeiting attributes and its historical origins. The benefit of thinking of derivative kinds in terms of focus-fuzzing is that it allows us to make sense of why it is reasonable to think of the world as having natural kinds, but without requiring any deep fact of the matter as to whether planets or eskimos or hurricanes are genuine natural kinds. It thereby avoids some unnecessary ontological clutter.

Notes

* This research was produced with the support of the National Endowment for the Humanities through a Summer Seminar stipend. I received very helpful comments from John Heil, Robert Rupert, Kevin Morris, Jaegwon Kim, and two anonymous referees from *Philosophia Naturalis.*

1 Tahko (2009) defends such a version of the law of non-contradiction.

2 I count Cornman (1975) and Campbell (1993) as dissenters on this point.

3 See Kutach (2011) for a case study applying abstreduction to the concept of causation.

Bibliography

Audi, Paul, 2007: *Beyond Causal Theories of Mind.* PhD. Dissertation. Princeton.

Cameron, Ross, 2008: Turtles all the Way Down: Regress, Priority, and Fundamentality. In: *The Philosophical Quarterly* 58 (230), pp. 1–14.

Campbell, John, 1993: A Simple View of Colour. In: Haldane, John; Wright, Crispin (ed.): *Reality, Representation and Projection.* Oxford: Oxford University Press, pp. 257–268.

Cornman, James W., 1975. *Perception, Common Sense, and Science.* New Haven, CT: Yale University Press.

Hardin, Clyde L., 1988: *Colour for Philosophers.* Indianapolis: Hackett.

Jackson, Frank, 1998: *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford: Oxford University Press.

Kim, Jaegwon, 1993: *Mind and Supervenience.* Cambridge: Cambridge University Press.

Kripke, Saul, 1971: Identity and Necessity. In: Munitz, Milton K. (ed.): *Identity and Individuation.* New York: New York University Press, pp. 135–164.

Kutach, Douglas, 2010: Empirical Analyses of Causation. In: Hazlett, Allan (ed.): *New Waves in Metaphysics.* Hampshire, UK: Palgrave Macmillan, pp. 136–155.

Kutach, Douglas, 2011: *Causation and Its Basis in Fundamental Physics.*

Leeds, Stephen, 2001: Possibility: Physical and Metaphysical. In: Gillett, Carl; Loewer, Barry (ed.): *Physicalism and Its Discontents.* Cambridge: Cambridge University Press.

Melnyk, Andrew, 2001: *A Physicalist Manifesto.* Cambridge: Cambridge University Press.

Oppenheim, Paul; Putnam, Hilary, 1958: Unity of Science as a Working Hypothesis. In: Feigl, Herbert; Scriven, Michael; Maxwell, Grover (ed.): *Concepts, Theories, and the Mind-Body Problem. Minnesota Studies in the Philosophy of Science, II* Minneapolis: University of Minnesota Press, pp. 3–36.

Paseau, Alexander, 2009: Defining Ultimate Ontological Basis and the Fundamental Layer. In: *The Philosophical Quarterly,* pp. 1–7.

Putnam, Hilary, 1975: The Meaning of 'Meaning'. In: Gunderson, Keith

(ed.): *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science, VII.* Minneapolis: University of Minnesota Press, pp. 131–193.

Quine, Willard Van Orman, 1969: Natural Kinds. In: Rescher, Nicholas (ed.): *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday.* Dordrecht: D. Reidel, pp. 5–23.

Russell, Bertrand, 1956: *Human Knowledge: Its Scope and Limits.* Allen, Shoemaker, Sydney, 2007: *Physical Realization.* Oxford: Oxford University Press.

Tahko, Tuomas, 2009: The Law of Non-Contradiction as a Metaphysical Principle. In: *Australian Journal of Logic* 7, pp. 32–47.

Robert Van Gulick

Non-Reductive Physicalism and the Teleo-Pragmatic Theory of Mind

Abstract

I begin with a basic account of teleo-pragmatic functionalism and its main features. I then discuss what that view implies about the nature of cognition, theories and understanding and thus about the limits on our ability to explain the mental and its relation to the non-mental. I show that teleo-pragmatic functionalism leads naturally to a version of non-reductive physicalism that combines theoretical pluralism with a strongly contextualist and pragmatic view of theories and models. Though non-reductionist at the theoretical and conceptual level, the view is nonetheless thoroughly and robustly physicalist in its ontology.

Zusammenfassung

Ich beginne mit einem basalen Ansatz des teleo-pragmatischen Funktionalismus und charakterisiere seine Hauptzüge. Im nächsten Schritt diskutiere ich die Implikationen dieses Ansatzes in Bezug auf die Natur der Kognition, auf Theorien und Verstehen, sowie die Implikationen dieses Ansatzes in Bezug auf die Grenzen unserer Fähigkeit das Mentale und seine Relation zum Nicht-Mentalen zu verstehen. Ich zeige, dass die Annahme eines teleo-pragmatischen Funktionalismus zur Annahme eines nicht-reduktiven Physikalismus führt, der Theorien-Pluralismus mit einer streng kontextualistischen und pragmatischen Interpretation von Theorien und Modellen vereinigt. Während dieser Ansatz auf der Ebene von Theorien und Begriffen zwar nicht-reduktionistisch ist, so ist er in ontologischer Hinsicht dennoch eindeutig physikalistisch.

Introduction

Theorizing about the nature of mind, at least of the sort often done by philosophers, involves a degree of reflexivity and reciprocal intertheoretical dependence. Questions about the nature of mind intersect

philosophia naturalis 47-48 / 2010-11 / 1-2

in many ways with questions about the nature of theories, explanations and understanding, both in general and specifically with respect to our theorizing about mind itself. The implications run in both directions. Our view of mind should inform our view of theories and vice versa. Theories and explanations are cognitive constructs and thus mental products, and it is minds or minded creatures that understand. Thus our views about the nature of mind should affect our philosophy of science, and it in turn will determine the sorts of explanations and understanding of mind that we might expect or hope to achieve.

Within the context of that general interdependence, I will discuss the specific implications that follow from what I take to be the most plausible view of mind, what I have elsewhere called "teleo-pragmatic functionalism" or TPF (Van Gulick, 2004). I will begin with a basic account of teleo-pragmatic functionalism and its main features. I will then discuss what that view implies about the nature of theories and understanding and thus about the limits on the explanations we should expect to be able to give of the mental and its relation to the non-mental. I will show that teleo-pragmatic functionalism leads naturally to a version of non-reductive physicalism that combines theoretical pluralism with a strongly contextualist and pragmatic view of theories and models. Though non-reductionist at the theoretical and conceptual level, the view is nonetheless thoroughly and robustly physicalist ontologically.

Teleo-pragmatic functionalism

As its name implies, teleo-pragmatic functionalism includes three main elements. It is a version of functionalism (Putnam 1967, 1975; Fodor 1968; Block and Fodor 1972), but one that interprets the relevant notion of function teleologically and with a strong pragmatic emphasis on mind/world engagement. In its generic form, functionalism as a theory of mind consists of four basic claims.

1. Minds differ from non-minds in their systemic organization, not in their underlying substrate or matter per se.
2. Complex systems (including minds) exhibit many levels of organization in structure and activity.
3. New patterns, powers and regularities often arise at different levels of organization.

4. Items and features at a level are often or normally type-individuated in terms of the roles they play at the relevant level of organization.

In particular, what makes a given mental state (or process) a mental state of the particular type it is – e.g., a pain in one's toe, a memory of lunch, a belief that I am late, or a desire for a cup of coffee – is the function or role that it plays within the relevant system, i.e. within the internal organization of which it is a part and which regulates that system's or organism's interaction with its world. Hence the name "functionalism". What matters is the function that the relevant state or process plays within its containing system or organism (Levin, 2009).

Different versions of functionalism interpret the notion of "function" in diverse ways. Some view it as "machine function" defined in terms of the inputs, outputs and state transitions of a finite state computational system, as in Turing machine functionalism. Others interpret it in terms of causal roles specified in terms of such basic relations as causing, jointly causing, inhibiting, or blocking the production of various effects. Causal role functionalism can itself take many forms, depending on the specific types of effects and transitions specified in the relevant networks, e.g. the relevant external effects might be limited to simple movements or they might include intentional actions (Van Gulick, 2009).

Teleological functionalism includes an element of purpose or goal-directedness in its view of function. Like organic functions (the function of the kidneys is to filter the blood) and artifact functions (the function of an automotive cam shaft is to open and close the cylinder valves), the functions specified by the teleo-functional theory of mind focus on (causal) roles described in terms of how they contribute to the overall goals or ends of the relevant system (Lycan, 1987; Van Gulick, 1980; Sober, 1990) If the overall goal of a mind is to successfully guide the interaction of an organism with its world or environment, then the functional roles associated with particular mental states or processes should be specified at least partly in terms of how they contribute to that goal.

The pragmatic aspect of the teleo-pragmatic functionalism has a similar basis. Minds are systems that guide the practical causal interaction of an organism or system with its world. In so far the defining functional roles associated with mental states are anchored in the purposive engagement between the organism and its world, those profiles will be

sensitive to pragmatic factors affecting how the given state might contribute to the success of the larger systemic organization in mediating that engagement.

These aspects of TPF derive in large part from the fact that it views minds from a biological perspective. All the natural minds of which we know are biological. Non-biological minds may be possible, and indeed artificial minds may well exist in the not too distant future. But as yet we have no actual examples of non-biological minds. Given that as a matter of natural history, actual minds are biological in nature and origin, it is apt to think about them from a biological perspective, without excluding the possible of other useful ways of looking at them.

Biological minds developed primarily to enhance the adaptive engagement of organisms with their environments through informationally-sensitive goal-directed behavior (Dennett 1969, 1997; Van Gulick, 1980). Indeed the process of evolution by natural selection can itself be viewed as an information process, a means by which a lineage of organisms acquires and stores information about the world and its causal structure (Lorenz, 1965).

The word "information" is a frozen metaphor, which originally derives from an Aristotelian notion of one item "taking in (or taking on) the form" of another, i.e., being "in formed by it" or taking on its form, shape or nature. A related metaphor underlies the notion of "adaptation" which implies coming to better "fit" the environment and its causal structure, much like the fit of hand-in-glove (Lorenz, 1965). Through the process of natural selection, the structure of the relevant life forms and their interactive control systems are modified in ways that reflect the nature and causal structure of the environments they inhabit. From the TPF perspective, that process in itself constitutes the acquisition of information, a means by which the organism or lineage of organisms comes to be informed by the environment.

The notion of minds as information processing systems has been widespread in psychology since the so-called "cognitive revolution" in the 1960's and 70's (Gardner, 1985). Minds are viewed as systems for acquiring, storing, integrating and applying information, as happens in perception, memory, reasoning and the control of action. A key distinction concerns the difference between the mere passive possession of information and the active possession of information (Dennett, 1970; Van Gulick, 1980). Both involve the sort of covariance that grounds

information theory. Features of one system X carry information about another system Y in that basic sense just if there is a systematic and lawful covariance between the features of X and Y which makes it possible to determine the facts about X from those of Y. It is in this sense that tree rings carry information about past climate and the structure of the light array impinging on the retina carries information about the shape, size and location of the objects in the surrounding environment from which that light is reflected (Dretske, 1981).

Some information in the mind or nervous system is of this sort, but of more importance is active information possession, which in addition to covariance requires that the possessing system be able to use or apply that information to the achievement of its goals or ends. As the information that falls on the retina is processed through the visual system, it produces cortical states that do more than merely detect the presence and nature of environmental features. They also have the potential or capacity to modify the organism's response to those features in ways that enhance the probability of achieving its goals. If for example, they detect the presence of food at a specific spatial location, they can guide the organism in moving toward it and ingesting it.

The details of how information is processed and applied in specific cases varies greatly and is a matter for scientific research, but for present purposes it is only the general account that matters. Two inter-related general parameters are especially important in determining the content of the information that the organism comes to possess by being in the relevant state. Both admit of many matters of degree. First is the range and diversity of ways in which the organism might apply the relevant information. And second is the number and diversity of content-sensitive ways in which the given state can interact with other informational and contentful states.

At one extreme would be states that guide or trigger fixed tropisms like the neural states that trigger tongue motions in frogs responding to small irregularly moving objects (bug-like stimuli) in their nearby visual field. At the other extreme, would be full propositional beliefs as when I believe that my laptop is sitting in front of me on my desk. The range of the appropriate ways in which such a belief might shape my behavior is open-ended and nearly infinite, and the flexibility of its impact derives in part from its ability to appropriately interact with any number of other beliefs, desires, or intentions that I might have. From

the TPF perspective, it is crucial that the movement from crude informational states to sophisticated contentful states such as propositional states is not a move from practical states to purely intellectual states, but rather a move from states that have a narrow limited practical dimension to states that have a far wider and richer role in guiding practical engagement. At every point along the spectrum, the capacity or potential to affect practical engagement is central to the active possession of information. The move from informational states with crude content to those with sophisticated content is not a matter of *eliminating* the practical component but of *opening it up*, of *expanding* and *enriching* it.

Although the cognitive science literature focuses on the notion of information, from the TPF perspective the notion of "understanding" may be even more important. And in thinking about cognition, the notion of understanding accords better with the TPF view than the more traditional appeal to knowledge.

It does so for at least four reasons. First, understanding admits of degrees. One can understand something fully, partly or to only a limited extent. Second, the notion of understanding makes the practical dimension of cognition more evident. We are accustomed in everyday use to the idea that understanding often involves practical abilities and the capacity to successfully engage with what one understands. Though we can similarly talk of "know-how", it is all too easy to intellectualize the notion of knowledge and lose sight of its practical dimension. Third, understanding is often a reciprocal relation, especially in social contexts where understanding can exist between or among a group of individuals. This reciprocal nature of understanding coheres with the TPF view that cognition typically involves a dynamic engagement between the organism and its environment, and that cognition is often embodied not in a static representation but in a capacity for successful interaction with the relevant aspect of the world.

The fourth reason is less obvious and involves another hidden metaphor, this one in the word "understanding" itself. It concerns what we can call "constitutive understanding", i.e. cases in which the existence of the object or entity that is being understood depends (at lest partly) upon that understanding for its very existence. If one attends to etymology, one sees that the English word "understanding" and Latin derived word "substance" appeal to the same metaphor – that which "stands under and supports" something. The Latin roots are "sub" meaning

"under" plus the participle "stans" of the verb "sto/stare" meaning "to stand" as in the English word "stance". A substance is that which "stands under" and supports properties or qualities. One might dismiss the parallel between the literal meanings of "understand" and "substance" as a trivial coincidence or a mere pun. But I believe there is reason to take it more seriously, namely the existence of "constitutive understanding" in which the process of mutual and reciprocal understanding actually brings into existence the entities that are understood and thus in a real sense "under stands" them in the substantial sense of supporting their existence.

The point is relatively abstract. So let me give an example to make it more concrete. We are the people that we are in large part because we are disposed to act and interact as we do, and because we play the various social roles that characterize our individual lives. With people whom we have interacted closely with for a long time – our parents, our children or our partner in a long marriage – our reciprocal relationships are often central to our identity. After fifty years of marriage, a husband and wife typically understand each other in intimate and defining ways. They engage each other and interact with each other in myriad ways that lie at the core of their respective identities, but that emerge only in the context of their life together and their reciprocal interpersonal understanding. Each enables the other to be the person he or she is by providing the complementary side of their dynamic mutual relationship. That is what makes it a case of constitutive understanding. The way in which each successfully engages the other, both reflects the others nature and helps to create that nature, or at least those aspects of it which exist within the context of their relationship. In that sense, each supports – and thus "under stands" in the constitutive sense – the existence of the other as the person he or she is. Through their mutual reciprocal understanding, each under stands the other.

TPF's emphasis on understanding, like its biological perspective, serves to strengthen its pragmatic and teleological aspects. Three consequences merit note.

First on the TPF view, understanding and cognition can be *embodied* in procedures, and some contents may be only implicitly rather than explicitly represented. For example, the organism's understanding of the causal structure of some part of it world may be embodied solely in the control systems that successfully guide its interactions with the

relevant real world items. And even when explicit representations are used, their contents ultimately depend upon the roles they play – however direct or indirect – in guiding the organism's successful engagement with its world.

Second, understanding and cognition are of a large degree *interest-relative*. What counts as understanding some fact or feature F will depend in part on one's goals and interests, and thus on the ways in which one is concerned to engage or interact with F. A given system of organization and representation might suffice for understanding F in the ways relevant to one set of interests, but not for those to another.

Third, cognition and understanding are contextual in multiple dimensions. All the familiar forms of contextualism that have been discussed in the contemporary philosophical literature on context apply quite naturally to the TPF. For example, TPF easily accommodates the many varieties of externalism (Van Gulick, 2004) according to which intentional or representational content shifts depending upon various sorts of contexts in which the cognitive agent might be embedded. For example, content may depend upon facts concerning the person or organism's actual or historical environment (Putnam), the environment's causal or informational structure, or the agent's embedding with a social community (Burge). In each case, TPF is well positioned to accommodate and explain the relevant contextual shifts, within its overall account of how embedded cognitive agents successfully engage and understand their worlds. Border-shifting comes naturally to TPF. How wide or narrow to define the content-determining system and which parameters to include is just the sort of meta-level pragmatic choice that the basic TPF would lead us to expect. The best way to individuate – how widely or narrowly for example – will depend our particular interests, the sorts of questions we aim to address and the sorts of understanding we hope to achieve.

Having provided a brief survey of the teleo-pragmatic functionalist view of mind, let us turn to ask what implications it has for the nature of theories and explanations, both in general and more specifically with respect to theories and explanations of mental phenomena. Some of the implications are immediate and obvious. Given its emphasis on pragmatic engagement, context, interest-relativity, and embedded cognition, TPF leads naturally to a pluralist view of theories. Theories and models are cognitive constructs produced by situated agents to provide suc-

cessful engagement with their worlds. That engagement can take a wide diversity of forms, reflecting that many ways in which agents with varying interests may aim to deal with a complex many layered world that exhibits pattern and regularities of many kinds at many levels. Thus it is plausible to suppose that we will likely require a wide variety of different theories and models to guide those diverse cognitive interactions. There is not likely to be any single all-purpose form of representing or conceptualizing that would meet our diverse pragmatic needs and interests in dealing with that complex many layered reality. Thus a rich theoretical and representational pluralism is just what one would expect given a TPF view of mind.

Moreover, that pluralist outlook recurs when we consider the nature of inter-level and inter-theoretic relations. Given the multiplicity and diversity of levels of order in the world that our various theories aim to model and describe, the *relations between levels* are themselves likely to be of many sorts, as are the *relations between the representations* we construct to understand them.

Thus it is not surprising that TPF functionalism is typically combined with the sort of non-reductive physicalism that has been the plurality view (perhaps even the majority view) among analytic philosophers for the past several decades (Fodor, 1974, 1997; Boyd, 1980). Non-reductive physicalism combines a denial of representational reduction (REP-Reduction) with an acceptance of ontological physicalism (ONT-reduction) (Van Gulick, 2001). The basic idea is to embrace pluralism and a degree of autonomy at the *representational* level in terms of the cognitive resources we require to understand our world, but at the same time to claim that at the *ontological* level everything real is in a non-trivial sense basically physical in the sense that it depends essentially and constitutionally on the physical. The view at first may seem contradictory, and there is admittedly a certain tension between the two main elements of the view. But as I will show below, there is no actual conflict and the two can be consistently and coherently conjoined.

The Non-reductive physicalist (NRP) faces two complementary challenges. First he must show that his denial of representational reduction (Non REP-reduction) is strong enough to have real import and merits it "non-reductive" label. Secondly, he must show that his version of physicalism is robust enough to legitimate his physicalist credentials. In particular he must unpack the dependence relation (the relevant sense

in which everything real depends upon the (strictly) physical that is strong enough to support his physicalist claims without undermining his denial of Rep-reduction. The advocate of NRP thus faces a potential dilemma that might posed by his critics. Stated with specific regard to the mental and the physical, the dilemma D might be put as follows:

D1. If the NRP proponent cannot spell out in detail how the mental depends on the physical, then his physicalism is suspect or hollow.
D2. However, if he can spell out in detail how the mental depends on the physical, then he should be able to construct derivations of the sort required for REP-reduction.

He appears thus to be "damned if he does, and damned if he doesn't". If he can spell out the dependence, then he should give up his non-reduction, and if he cannot then he should give up his physicalism. To see how the NRP supporter can respond and avoid the dilemma, we need first to get clear about both components of his view.

Representational reduction (REP-reduction) is a claim about the relations between representational and cognitive items: theories, models, concepts, or frameworks. And it typically asserts a very strong and tight connection between them as on classic reductive unity of science of view according to which our theories at each level should be strictly derivable from those at a lower level of underlying structure (Oppenheim and Putnam, 1958; Putnam, 1970). Chemistry should be derivable from physics, physiology and biology from chemistry, and psychology from neuroscience. The concepts at each higher-level should be definable in terms of those at the lower level and the laws and truths of upper level theories should be derivable from those of lower level theories together with appropriate bridge laws. Since the 1970's that sort of reductive picture has been largely out of favor, and other less reductive views of inter-level relation such as those asserting the autonomy of the special sciences have been widely embraced (Fodor, 1974, 1997; Boyd, 1980). Weaker forms of representational reduction may not require a inter-theoretical connection quite as tight as strict derivability, but they are require that the reducing theory be able to capture most (if not all) of the legitimate cognitive content of the reduced theory (Van Gulick, 1992).

For that reason, claims of representational or theoretical reduction typically include an assertion that the reduced theory could at least *in principle* be eliminated or replaced by the reducing theory without

any loss expressive power or content, even if it involved some costs as a purely practical matter. The representational reductionist may concede that as a practical matter the reduced theory might be easier to use, but he would maintain that strictly speaking it adds nothing over and above what is contained in the reducing theory.

However, that distinction itself – between what is *in principle* derivable and what can be done *in practice* – is problematic from a non-reductionist perspective, i.e. from the perspective of those who deny that representational reductions are typically not possible and who champion the necessity and autonomy of the special sciences. This is especially so if the non-reductionist is also a teleo-pragmatic functionalist for whom theories are cognitive tools constructed to enable situated agents to effectively engage the relevant parts of their world. On such a view, the cognitive content of a theory cannot be divorced from the role it plays (or could play) in mediating that interaction. And what roles it does or could play will depend not only on the causal structure of the world, but also on the causal structure of the cognitive agent and on the possibilities for engagement provided by the nature of the situation in which the agent and the object of the theory are embedded. The cognitive equivalence of two theories (or their equivalence in content) would in general require that they provide the relevant agents causal-interactive equivalence in dealing with the relevant target parts of the world. The interface between agent and world that the theory or representational system provides is central to its cognitive content and to the understanding that it affords the theory-user. It cannot be dismissed as a merely practical matter.

From the TPF perspective, the pragmatic and contextual nature of cognition, representation and content make it highly unlikely that many inter-theoretical and inter-representational relations will be of the sort required by REP-reduction. Theories (and other systems of representation) apt for understanding different levels of order and organization will likely differ along many parameters in ways that will make it very difficult (or impossible) to bring them into the sort of tight correspondence (derivation or expressive equivalence) required by REP-reduction (Van Gulick, 1992, 2002). Theories applied at different levels will typically differ in many content-determining aspects including the following four among others:

1. The nature of the *causal engagement* that the theory affords the theory-user.
2. The nature of the *interests or goals* toward which the theory might be adaptively applied.
3. The nature and role of *indexical and demonstrative* elements in fixing the reference and content.
4. *The contextual* principles for individuating kinds, e.g. how *widely or narrowly individuated and along what parameters.*

Given the disparity in such content-determining aspects between theories (or other systems of representation) applied to different levels of organization, it will in many (but not all cases) be difficult (and practically impossible) to put the representations of the two theories into the sort of tight correspondence required by REP-reduction.

The denial of REP-reduction has a number of consequences, some of which we have already noted. It supports the need for theoretical and representational pluralism, as well as the autonomy of the special sciences. It leads us to expect the inter-level links between theories to be more diverse in kind and often looser than on the REP-reduction modes. Importantly, it gives us good grounds to reject the strictly hierarchical view associated with the classic logical empiricist unity of science model.

On that classic unity of science view, the theories at each level could be derived from and defined in terms of the theory just below them, which in turn would bear the same relation to the theory at the next level down, with as many recursions as needed to reach theories at the micro-physical base level. Psychological theories could be REP-reduced to neurological theories, and they in turn be REP-reduced to physiological and bio-chemical theories which would be reduced to physics. The NRP proponent who denies REP-reduction as the norm for inter-theoretical relations has a couple of reasons to reject the hierarchical view.

First the hierarchical view involves a linear non-looping sequence of theories, each fully definable in terms of the theory at the level below it. However, in reality concepts and kinds at a given level will often depend for their individuation on kinds at levels above them as well as below. For example on the classic unity of science view, social theories together with their associated social kinds are to be reduced to psychological

theories describing the minds of the individuals who compose the relevant society. In reality, the concepts and individuating kinds at the two levels are to some degree mutually inter-dependent and inter-penetrate each other. While it may be true that social facts supervene on facts about the minds of the individuals who compose the relevant society, it is also true that the factors relevant to defining and individuating mental kinds in many cases will be sensitive to social factors (Burge, 1979). There are definitional loops within the collection of theories at different levels, e.g. the nature of the social depends in part upon he psychological, but the nature and kinds of the psychological also partly depend on the social.

Second, features of a given type may occur at more than one level in diverse forms, and as a result they may inter-penetrate each other in a series of recurrent loops. For example, semantic and contentful kinds may occur at multiple levels, with the "objects" semantically represented at a given level n serving as syntactic entities for use at some higher level n+m. Is the semantic defined and determined by the syntax or does the dependence flow in the other direction? Again the answer is likely to be, "Both" with many iterations.

If we begin at lower mental-levels of the sort associated with only crude and unsophisticated content, it may be possible to implicitly represent the nature of simple objects purely in terms of procedures that govern the organism's interactions with those items, without any need for explicit symbols or symbolic structures to encode that information. It is in virtue of those procedures, which allow the organism to *target and engage* the external object x, that x comes to exist in a very basis sense as an *intentional object for* the organism. However, once such objects exist as notional or *represented objects*, they can be recruited and used as *representing items* and "put to work" as syntactic structures underlying perhaps more sophisticated forms of content at a higher level. That is, once the system has succeeded in procedurally defining and understanding a domain of objects, it can use those objects as syntactic items to construct more sophisticated representations of other objects.

Moreover, such patterns of inter-dependence and "boot strapping" may iterate and occur more than once, with semantic relations defining or inducing the existence of a set of "objects: that then go on to serve as syntactic structures to enable the expression of more sophisticated semantics content at a higher level. On such a picture, the syntactic is

not *per se* more basic than the semantic or intentional, nor is the reverse true. Rather both the syntactic and the semantic occur at multiple levels that allow each in some case to underlie the other. The answer to the question, "Which is primary: syntax or semantics?" may be that in complex mental systems they develop in recursive loops with many successive layers of syntax and semantics building upon each other.

Having addressed the issue of how TPF relates to the denial of REP-reduction, let us turn our attention to the physicalist aspect of Non-reductive physicalism. Given the non-reductivist's inability to establish tight connections of derivation or content-equivalence between even his best theories of the mental and the physical, can he nonetheless give an account of how the mental depends upon the physical that is sufficiently robust to justify his claimed status as a physicalist? Does his rejection of representational and theoretical reduction also force him into being an ontological non-reductionist and denying that everything real is essentially and constitutively physical?

Since TPF is by definition a version of functionalism, it typically explains the relation between the mental and the physical as a matter of *realization.* A given level of organization in a mind (or other complex system) must be concretely implemented in some underlying structures that are able to play the requisite functional roles. As was noted very early in the history of functionalism, the distinction between *structure* and *function* is relative and not absolute (Kalke, 1969; Lycan, 1987). Characterizing something as a *neuron*, a *transistor*, or a *circuit breaker* might serve either as a structural description or as a functional one, depending on one's current interests and project. A neuron could be treated as an underlying *structure* that realizes an abstract role by carrying out some specified computational integrations or as a f*unctional* entity realized by underlying subcellular structures and mechanisms. There is no simple answer in the abstract to the question whether "neuron" designates a structural kind or a functional kind. It can serve as either depending on the explanatory context.

Nonetheless, there is a general principle that functional organizations at any level n must be realized or implemented by underlying structures at a lower level n-1. And the downward progression of realizations must eventually terminate at some level composed of basic mechanisms and structures TPF like most versions of functionalism regards the ultimate base level as consisting of entities that appear in the ontology of our

most fundamental physics, whatever they might be – subatomic particles, fields, strings, branes, or other entities as yet unknown.

In order to state the TPF proponent's commitment to physicalism, we first need an appropriate definition or demarcation of what is to count as physical. PD provides a recursive definition of the physical in terms of realization and consists of a base clause i. and recursion clause ii.

PD – Definition of the Physical:

i. If x is an entity or any sort – object, force, property, relation or otherwise – that appears in the fundamental ontology of our best physical theory, then x is physical.
ii. If x is an entity of any sort that is realized entirely by physical entities, then x is physical.

Since there is no limit on how many times the recursion clause ii can be applied, we are able to progressively build up the ontology of the physical through cycles of realization beginning at the fundamental base and gradually including atoms, molecules, proteins, enzymes, organelles, cells, organs, organisms, social groups, economies, financial systems, and even bubble and bust financial markets. They will all count as physical as long as each is produced by a series of realizations that ultimately terminate at the fundamental physical base. We can now state the TPF proponent's version of physicalism relying on the definition of the physical provided by PD. It is quite simple.

P. Physicalism: Everything real is physical.

Given the definition PD, P amounts to asserting that everything real is either fundamentally physical or ultimately realized by the physical. The possible existence of abstract entities such as numbers or sets might seem to fall outside of the domain of the physical, and thus if they are real to refute P. The TPF theorist can set concerns about the status and reality of abstract entities to one side, by adding a restriction on P limiting its claim to what exists in space and time, or even just in time. Thus P* might provide a more nuanced statement of physicalism.

P* Physicalism: Everything real that exists in space and/or time is physical.

P* is silent about the status of abstract entities, but it remains a robust ontological claim. The fact that numbers and other abstract entities may

fall outside the scope of the physical does not really diminish its significance, and it clearly implies that everything that is *real and mental* is also physical. Moreover, it is physical in a strongly *constitutive* sense. The dependence of the mental is not merely a causal dependence but a constitutive one. Minds – as well as mental states, events and processes – are realized (i.e. made real) and fully constituted by underlying physical structures, properties and organizations. There are no super-added or radically emergent qualities or entities that enter the picture.

But what of the dilemma posed earlier to the TPF theorist? If everything real, including everything that is real and mental, is realized by the underlying physical and thus fully constituted by it, why can we not spell out those dependencies and realization relations in ways that would enable us to carry out a REP-reduction of the theories that we use to describe and understand the mental to the theories that we use to describe the underlying physical substrates that realize the mental? And the same applies to all the other special sciences theories that we use to describe and understand the many higher layers of our complexly organized but ultimately physical world. Put crudely, why does the ontological reduction (ONT-reduction) of the mental to the physical via realization not imply the representational reduction (REP-reduction) of our mental theories to purely physical theories?

The answer is provided by the pragmatic and situated nature of cognition and understanding on the TPF view. As we saw above in discussing the denial of REP-reduction, those contextual and pragmatic factors make it unlikely in most cases that the sorts of tight inter-theoretical links required for REP-reduction will be available, at least not in any remotely practical sense. From the TPF perspective, theories and models are cognitive tools constructed to help guide and mediate our interaction with the relevant parts of our world. Their success and capacity to function in that way depends in part on the relation between their formal structure and the causal structure of the items to which they are applied. But it also depends in part on the causal structure of the *cognitive agent* who uses those constructs and on the causal structure of the *interface* between the agent and the relevant portion of the world, which itself depends in many ways on various aspects of the larger context in which the agent is embedded.

Hidden metaphors again play a role in our thinking about this. In particular, things are likely to look different if we follow TPF in thinking

about theories and models as cognitive *tools* rather than as *pictures* of reality. If one implicitly adopts the picture metaphor for thinking about theories, then it will likely seem that in a case of realization it should always be possible to reduce our representation of the realized feature to a theoretical representation of its realization base. If the elements of the underlying structure were like pixels, then once we had a complete representation of them, we would also have all the means we need to represent all the larger macro features and patterns realized by those pixels. Fix the pixel elements and you have also represented the larger picture and all its features.

However, if theories and models are instead thought of as *cognitive tools*, then the inference from realization to REP-reduction seems problematic and certainly far less obvious. For example, our understanding of some higher level properties or regularities may turn on facts about our particular causal organization that allows us to use the relevant higher-order theory to engage those higher level items in a resonant or dynamically reciprocal way. It might be quite impossible for us to use the resources adequate for dealing with the lower level features to fashion a representation sufficient for guiding our interaction with the realized higher level feature.

The basic point is well illustrated by the familiar economics example which goes back at least as far as Jerry Fodor's early articles on behalf of the autonomy of the special sciences (Fodor, 1974, 1975). Every real world economic or financial transaction requires some physical realization. I can pay you fifty Euros in many different ways – with euro bills, with coins, with a personal check, a plastic credit card or a purely electronic funds transfer. Every economic or financial transaction requires some physical implementation or realization base, no matter how heterogeneous they may be from the perspective of the underlying physics. But it would be utter madness to suppose that the conceptual and representational resources used in micro-physics could in any practical sense be used to give us economic understanding or provide us with appropriate means for guiding our adaptive engagement with the economic aspects of our world.

Self-awareness and introspection provide another and perhaps even more compelling example. Let us suppose that every mental state or process of which I can have direct first-person awareness must be physically (or neurally) realized. Even so, it seems unlikely in the extreme

that I could fashion third-person neuroscience representations that would provide me with causal interactive access to those states that was cognitively equivalent to that provided by the representations that I use from within my subjective point of view to monitor and control my own experience and mental activity. The causal interfaces are so different and they depend in such different ways on the causal profile of the cognitive agent, that there seems little or no possibility that the two systems of representation could be put in any tight REP-reductive correspondence, despite the fact that the one provides the complete realization base for the other.

The causal profile of higher-level features must be such that it enables the relevant situated and embodied agent to use those higher-level representations to successfully engage those features. It is on that basis that those representations can be said to apply to those features or describe them. And on a realization view of the sort embraced by TPF, the causal profile of any organized system depends solely on the causal profiles of its constituents plus their mode of combination. Thus the applicability or aptness of the higher-level representations to the system ultimately depends on the causal properties of the system's parts and their mode of combination. But the cognitive agent's ability to understand and engage that higher-level causal profile will typically depend on the agent's own causal profile, the context in which he is embedded, and on various contextual aspects of that embedding.

Thus the fact that the lower level realization determines the causal profile of the realized complex does not entail that the theories we use to understand the complex could be REP-reduced to (or replaced by) the theories we use to understand the features of the realization base. Just because we have good tools to successfully engage the elements of the underlying (micro-)base, it does not follow that we *as the cognitive agents that we are* can use those same tools to successfully engage every real pattern or complex organization realized by those underlying entities. How we can use those tools and what sorts of understanding they can provide us will depend in all sorts of ways on the nature of our causal interface with the relevant parts of the world and how those tools can mediate that engagement.

Thus TPF provides a means to reject the dilemma earlier posed to the Non-reductive physicalist. Realization does not entail REP-reduction, and we should not typically expect to be able to put our theories of the

realized features in tight correspondence – derivation or strict content equivalence – with our theories or representations of the realizing items. We can legitimately expect some looser and more approximate linkages. We should in general be able to explain how the relevant realizers are the sorts of things that could underlie the features that they realize, e.g. by showing how their causal profiles are of the sort that if suitably organized could give rise to the general sorts of causal profiles that occur at the realized level. But that sort of general possibility account would fall far short of the sorts of linkages required by strict REP-reduction.

Let me finally return to address two challenges that were earlier raised against the sort of non-reductive physicalism I have been endorsing, which aims to combine a rejection of representational reduction with the acceptance of the ontological reduction of the mental to the physical. First does the proposed view provide an account of the dependence relation that is sufficiently robust to justify its physicalist credentials. And second does it qualify as a genuinely and non-trivially non-reductionist view? I believe the answer to both questions is clearly "yes".

As we saw above, realization is a constitutive relation. And if everything real is either fundamentally physical or physically realized, then everything real is physically constituted. That is a strong ontological claim that surely merits the "physicalist" label. With respect to its non-reductive status, some anti-physicalist might argue that the term "non-reductive" should be reserved for views that reject ontological reduction rather merely denying representational reduction. For example those who believe in immaterial substances (Foster, 1996; Swinburne, 1997), property dualism (Chalmers, 1996) or radically emergent causal powers (Hasker, 1999) might regard the denial of representational reduction as "non-reduction light", a minor and uninteresting claim undeserving to be classed as non-reductive.

However, there are good reasons to the contrary. First as a matter of actual recent history in the discipline, many (likely the majority) of those who have called themselves non-reductivists have done so because they reject theoretical or representational reduction and championed the autonomy of the special sciences. And they did so in response to specific influential claims and movements that held otherwise, such as the logical empiricist project on the unity of science, which advocated across the board theoretical reduction as the working aim of science. Though few current philosophers would explicitly endorse that pro-

gram, its influence lingers and affects a lot of what gets said and thought about these matters. For example, I believe it underlies the numerous and mistaken attempts to derive metaphysical and ontological conclusions about the distinctness of the mental and the physical from epistemological premises about what our concepts allow us to can imagine or what our theories do not allow us to explain in the strict deductive sense (e.g., arguments involving thought experiments about zombies (Chalmers, 1996) or Mary the super color scientist (Jackson, 1982; Ludlow et al, 2004). Once one recognizes what the TPF view has helped us see, namely that theoretical and representational non-reduction is fully compatible with ontological reduction, then these anti-physicalist arguments lose most or all of their force. And that I claim is a good thing.

Bibliography

Block, Ned and Fodor, Jerry, 1972: What psychological states are not. In: *Philosophical Review* 81, pp. 159–181.

Burge, Tyler, 1979: Individualism and the mental. In: *Midwest Studies in Philosophy* 4, pp. 73–121.

Chalmers, David, 1996: *The Conscious Mind*. New York: Oxford University Press.

Dennett, Daniel, 1969: *Content and Consciousness*. London: Routledge

Dennett, Daniel, 1971: Intentional systems. In: *Journal of Philosophy* 68, pp. 87–106.

Dennett, Daniel, 1997: *Kinds of Minds*. New York: Basic Books.

Dretske, Fred, 1981: *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Fodor, Jerry, 1968: *Psychological Explanation.* New York: Random House.

Fodor, Jerry, 1974: Special sciences: or the disunity of science as a working hypothesis. In: *Synthese* 28, pp. 97–115.

Fodor, Jerry, 1975: *The Language of Thought*. New York: Thomas Crowell.

Fodor, Jerry, 1997: Special Sciences: Still Autonomous After All These Years. In: Tomberlin, J.(ed.): *Philosophical Perspectives 11: Mind, Causation, and World*. Boston: Blackwell, pp. 149–64.

Foster, John, 1996: *The Immaterial Self*. London: Routledge.

Gardner, Howard, 1985: *The Minds New Science.* New York: Basic Books.

Hasker, William, 1999: *The Emergent Self* Ithaca: Cornell University Press.

Jackson, Frank, 1982: Epiphenomenal qualia. In: *American Philosophical Quarterly*, 32, pp. 127–36.

Kalke, William, 1969: What is wrong with Fodor and Putnam's functionalism. In: *Nous* 3, pp. 83–94.

Levin, Janet, 2009: "Functionalism". *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/functionalism/

Lorenz, Konrad, 1965: *Evolution and the Modification of Behavior.* Chicago: University of Chicago Press.

Ludlow, Peter, Nagasawa, Yujin, and Stoljar, Daniel (eds.), 2004: *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument.* Cambridge. MA: MIT Press.

Lycan, William, 1987: *Consciousness.* Cambridge, MA: MIT Press.

Oppenheim, Paul and Putnam, Hilary, 1958: The unity of science as a working hypothesis. In: Feigl, H. et al., (eds): *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: Minnesota University Press, pp. 3–36.

Putnam, Hilary, 1967: Psychological predicates. In: Capitan, William; Merrill, D. (eds.) *Art, Mind, and Religion.* Pittsburgh, PA: University of Pittsburgh Press, pp. 37–48.

Putnam, Hilary, 1970: On Properties. In: Rescher, Nicholas et al (eds): *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*, Dordrecht: D. Reidel, pp. 235–254.

Putnam, Hilary, 1975: Philosophy and our mental life. In: Putnam, Hilary (ed.): *Mind, Language and Reality.* Cambridge: Cambridge University Press, pp. 291–303.

Swinburne, Richard, 1997: *The Evolution of the Soul.* Oxford: Oxford University Press.

Van Gulick, Robert, 1980: Functionalism, information and content. In: *Nature and System* 2:139 (Reprinted in Bermudez, José; McPherson, F. (eds.) 2006: *Philosophical Psychology: Contemporary Readings.* London: Routledge, pp. 64–89.

Van Gulick, Robert, 1982: Functionalism. *The Philosophy Research Archives* 8, pp. 185–204.

Van Gulick, Robert, 1992: Nonreductive materialism and intertheoretical constraint. In: Beckermann, Ansgar; Flohr, Hans; Kim, Jaegwon (eds.): *Emergence and Reduction*. Berlin and New York: De Gruyter, pp. 157–179.

Van Gulick, Robert, 2001: Reduction, emergence and other recent options on the mind-body problem. In: *Journal of Consciousness Studies* 8–9, pp. 1–34.

Van Gulick, Robert, 2002: "Nichtreduktiver Materialismus – noch immer das beste Angebot auf dem Leib-Seele Basar" (Nonreductive physicalism: still the best buy at the mind/body bazaar). In: Pauen, Michael; Stephan, Achim (eds.): *Phänomenales Bewusstsein. Rückkehr zur Identitätstheorie?* Berlin: Mentis Verlag, pp. 297–327.

Van Gulick, Robert, 2004: Outing the mind. In; Schantz, Richard (ed.): *The Externalist Challenge: New Studies on Cognition and Intentionality.* Berlin and New York: De Gruyter, pp. 255–84.

Van Gulick, Robert, 2009: Functionalism. In: McLaughlin, Brian; Beckerman, Ansgar (eds.): *Oxford Handbook of Philosophy of Mind.* Oxford: Oxford University Press, pp. 128–51.

Part II Reduction, Phenomenality, and the Explanatory Link

Markus I. Eronen

Replacing Functional Reduction with Mechanistic Explanation

Abstract

Recently the functional model of reduction has become something like the standard model of reduction in philosophy of mind. In this paper, I argue that the functional model fails as an account of reduction due to problems related to three key concepts: functionalization, realization and causation. I further argue that if we try to revise the model in order to make it more coherent and scientifically plausible, the result is merely a simplified version of what in philosophy of science is known as mechanistic explanation. Hence, instead of analyzing reduction in philosophy of mind in terms of functional reduction, it should be analyzed in terms of mechanistic explanation.

Zusammenfassung

In letzter Zeit ist das Modell funktionaler Reduktion zu so etwas wie dem Standardmodell von Reduktion in der Philosophie des Geistes geworden. Im vorliegenden Artikel argumentiere ich, dass das Modell funktionaler Reduktion als Reduktionsmodell versagt, und zwar aufgrund von Problemen dreier zentraler Begriffe: Funktionalisierung, Realisierung und Kausale Verursachung. Darüber hinaus argumentiere ich, dass eine im Lichte dieser Probleme revidierte Fassung des Modells funktionaler Reduktion, die sowohl kohärenter als auch wissenschaftlich plausibler ist, sich nicht vom Modell mechanistischer Erklärung unterscheidet. Aus diesem Grund sollte Reduktion in der Philosophie des Geistes unter Rekurs auf ein Modell mechanistischer Erklärung und nicht mit Hilfe eines Modell funktionaler Reduktion analysiert werden.

philosophia naturalis 47-48 / 2010-11 / 1-2

1. Introduction

In recent years, the functional model of reduction has become something like a standard model of reduction in philosophy of mind. However, the model is by no means new: its main ideas are already present in the filler-functionalism of David Lewis (1972). Lewis' idea was roughly that a given mental state *M* is defined functionally in terms of its causal role, but in the end *M* is nothing more than the physical states that occupy this role. For instance, "pain" is a functional concept specified by its causal role, but in the end pain *just is* the physical (neural) state that fills that causal role. This physical state can be one thing in humans, another in octopuses, and still something else in Martians. These different states are all picked out by the functional concept "pain", which (non-rigidly) designates different physical fillers in different species.

More recently philosophers like Joseph Levine (1993), David Chalmers (1996), Frank Jackson (Chalmers and Jackson, 2001), and Jaegwon Kim (1998; 2005) have presented somewhat varying models of functional reduction based on this general approach. All of these authors have then applied the supposedly general model of reduction to the problem of phenomenal consciousness, arguing that phenomenal properties are fundamentally irreducible, or that there is an "explanatory gap" between phenomenal properties and the physical domain.

I will focus here on Kim's version of functional reduction, since it is exceptional in its clarity, and has also been extremely influential. I will argue that the functional model fails to capture the nature of reduction in psychology and neuroscience. Furthermore, I will show that if we try to revise the functional model in order to make it more scientifically credible, it turns out that the revised model is not significantly different from mechanistic explanation. Hence, instead of analyzing reduction in philosophy of mind in terms of functional reduction, it should be analyzed in terms of mechanistic explanation.

2. The functional model

Kim's main motivation for invoking the model of functional reduction is to show that mental properties (with the exception of phenomenal properties) can be saved from the *causal exclusion argument*, which I

will briefly sketch here. Several different versions of the argument exist; the formulation here reflects Kim's most recent accounts (Kim, 2002; 2005). The argument is based on certain principles that together create a problem for mental causation (Kim, 2002, 278):

The Problem of Mental Causation: Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (1) physical causal closure, (2) exclusion, (3) mind-body supervenience, and (4) mental/physical property dualism (i.e., irreducibility of mental properties).

The principle of physical causal closure states that every physical occurrence has a sufficient physical cause. The principle of exclusion states that no effect has more than one sufficient cause, except in cases of genuine overdetermination, such as two bullets hitting the heart of a victim at exactly the same time, both causing death.

It is easy to see how the four principles taken together lead to trouble. Let us start by assuming that (the instantiation of) a mental property *M* causes (the instantiation of) another mental property *M**. Due to mind-body supervenience, *M* supervenes on some physical property *P*, and *M** supervenes on some physical property *P**. Since *M** supervenes on *P**, *M** must be necessarily instantiated whenever *P** is instantiated, no matter what happened before: the instantiation of *P** alone necessitates the occurrence of *M**. Thus, according to Kim, the only way that *M* can cause *M** is by causing *P**.

This is where the principle of causal closure kicks in: *P** must also have a sufficient physical cause. This means that *P** has a sufficient physical cause *P* and a mental cause *M*, and the exclusion principle states that one of these must go – if we would accept cases like this as genuine overdetermination, we would get massive overdetermination of physical effects by mental causes, which is highly implausible. Obviously *M* is the one that has to go, since if *M* was the only cause of *P**, this would violate the principle of physical causal closure. Therefore, *M* cannot be the cause of *M** or of any other mental or physical property. This holds for all mental properties, and we have the striking conclusion that, under mind-body supervenience, mental properties are causally impotent.

According to Kim, physical causal closure and mind-body supervenience are among the inescapable commitments of all physicalists. The exclusion principle is taken to be a general metaphysical constraint that can hardly be challenged. This leaves only mental/physical property

dualism (i.e., the irreducibility of mental properties) as the principle that has to go. Therefore, Kim's conclusion is what he calls "conditional reductionism": "If mentality is to have a causal influence in the physical domain – in fact, if it is to have any causal efficacy at all – it must be physically reducible" (Kim, 2005, 161).

What does reduction then amount to? Kim's answer is the functional model:

> To reduce a property, say being a gene, on this model, we must first "functionalize" it; that is, we must define, or redefine, it in terms of the causal task the property is to perform. Thus, being a gene may be defined as being a mechanism that encodes and transmits genetic information. That is the first step. Next, we must find the "realizers" of the functionally defined property – that is, properties in the reduction base domain that perform the specified causal task. It turns out that DNA molecules are the mechanisms that perform the task of coding and transmitting genetic information – at least, in terrestrial organisms. Third, we must have an explanatory theory that explains just how the realizers of the property being reduced manage to perform the causal task. In the case of the gene and the DNA molecules, presumably molecular biology is in charge of providing the desired explanations. (Kim, 2005, 101)

Kim presents the functional model as a better and more scientifically credible alternative to Ernest Nagel's (1961) classic but problematic model: "Nagel reduction of pain requires an all-or-nothing, one-shot reduction of pain across all organisms, species, and systems. It is clear that functional reduction gives us a more realistic picture of reduction in the sciences" (Kim, 2005, 102). In Nagel's model, reduction of a theory T_2 consists in deducing it from a more fundamental theory T_1, with the help of "bridge laws" that connect the terms of the two theories. What Kim sees as the main problem with Nagel's model is that it gives us reductions that do not explain (Kim, 1998, 90–97; 2005, 98–101). This is because, according to Kim, the reductive work in Nagel's model is done by the biconditional bridge laws that connect properties of the reduced theory to properties of the reducing theory, and these bridge laws are just "unexplained auxiliary premises" that are themselves in need of explanation.

Ausonio Marras (2002) has pointed out that bridge principles do not in fact play a key role in Nagelian reductions, and therefore Kim's critique is largely misplaced. However, in the present context, Nagelian reduction faces other, more fundamental, problems. The main problem of Nagelian models of reduction in the context of psychology and neu-

roscience is that they require the theories involved in reductions to be formalized, either according to the syntactic (e.g., Nagel, 1961) or the structuralist/semantic (e.g., Bickle, 1998) view of theories.[1] The problem is that while formal theories that are suitable as starting points of logical derivations may be available in theoretical physics, most special sciences simply do not have any well-structured theories that could be handled formally. Rather than trying to formulate such theories, psychologists and neuroscientists typically look for descriptions of mechanisms that can serve as explanations for patterns, effects, capacities, phenomena, etc., and this explanatory enterprise at best involves fragments of formal theories (Craver, 2007; Machamer et al., 2000; McCauley, 2007; Walter and Eronen, forthcoming). Furthermore, some generally accepted cases of scientific reduction – for instance the reduction of genetics to molecular biology – do not seem to involve formal theories (Sarkar, 1992). In this light, the model of functional reduction is prima facie promising, since it is a model of property reduction, not theory reduction, and does not require formal theories.[2]

Let us then take a closer look at Kim's model of functional reduction (Kim, 1998, 97–103; 1999, 10–13). The reduction of property *M* consists of three steps:

Step 1: *M* must be functionalized – that is, *M* must be construed, or reconstrued, as a property defined by its causal/nomic relations to other properties. As Kim puts it:

> [W]e must first "prime" *M* for reduction by construing, or reconstruing, it *relationally* or *extrinsically*. This turns *M* into a relational/extrinsic property. For functional reduction we construe *M* as a second-order property defined by its causal role – that is, by a causal specification *H* describing its (typical) causes and effects. So *M* is now the property of having a property with such-and-such causal potential[.] (Kim, 1998, 98)

Thus, property *M* is defined as a "second-order" property: it is a property that some first-order properties have.

Step 2 consists of finding the realizers of *M*. These are the first-order properties in the reduction base domain that have the right causal/nomic relations, i.e., the properties that fit the causal specification *H*. The realizers can be different in different systems, allowing for multiple realizability. Step 2 is a matter of scientific research, or as Kim puts it, "a scientifically significant part of the reductive procedure" (Kim, 1999, 11).

Step 3 is to find a theory that explains how the realizers actually perform the causal role specified in Step 1. Like Step 2, Step 3 is also a matter of scientific research, and these steps are intertwined, since figuring out what the realizers of *M* are certainly involves theories about the causal/nomic relations in the reduction base.

One of the central points of Kim's account is that functionally reduced properties are nothing "over and above" the reducing properties: "Central to the concept of reduction evidently is the idea that what has been reduced need not be countenanced as an independent existent beyond the entities in the reduction base – that if *X* has been reduced to *Y*, *X* is not something 'over and above' *Y*" (Kim, 1999, 15). According to Kim, this means that reduction has to lead either to identities (conservative reduction) or eliminations (replacement/eliminative reduction). Is functional reduction then conservative or eliminative?

First of all, Kim argues that when *M* has been functionally reduced to *P*, *instances* of *M* can be identified with the instances of *P* (Kim, 1999, 15–16). He invokes the "causal inheritance principle", which states that "[i]f a functional property [*M*] is instantiated on a given occasion in virtue of one of its realizers, [*P*], being instantiated, then the causal powers of this instance of [*M*] are identical with the causal powers of this instance of [*P*]." If we accept this principle, it follows that the instances of *M* and *P* have exactly the same causal powers, and it is hard *not* to identify the instances, since if they were not identical, the difference could not even be detected. However, what is at issue in the exclusion argument is not token causation (one instance or event causing another instance or event), but type causation. The problem is whether mental *properties* can have causal powers – in other words, whether some event can cause a mental or physical event in virtue of being an instantiation of a mental property. Therefore, for avoiding the exclusion argument it is not enough that *instances* of *M* are identical to instances of physical properties, also the property *M* itself has to be identical to a physical property *P*.

The situation is made even more complicated if (as is generally assumed) *M* can have multiple realizers. Therefore, Kim sees only two options: we can (1) identify *M* with the disjunction of its realizers, or (2) give up *M* as a real property and only recognize it as a property designator that picks out many different properties (the realizers of *M*).

Identifying *M* with the disjunction of its realizers is problematic. The

realizers must have different causal roles, since otherwise they wouldn't be different realizers (Kim supports a causal theory of properties). If *M* is identical to a set of causally and nomologically heterogeneous properties, Kim reasons, then *M* itself must be causally and nomologically heterogeneous, and is unfit to figure in laws, and is thereby not a scientific property (see Kim (1992) for more details of this argument).[3]

Therefore, Kim is inclined to accept the second option:

> One could argue that by forming "second-order" functional expressions by existentially quantifying over "first-order" properties, we cannot be generating new properties (possibly with new causal powers), but only new ways of indifferently picking out, or grouping, first-order properties, in terms of causal specifications that are of interest to us. (Kim, 1999, 17)

This makes functional reduction eliminative: we have to accept that mental properties are not genuine properties in their own right. Kim accepts this only because the other alternatives (disjunctive identities or property dualism) are wrought with major philosophical problems (Kim, 2008, 112). I will return the problems of this option in Section 3.2 below.

3. What Is Wrong with the Functional Model

The functional model has been recently criticized from different angles. Ausonio Marras (2002; 2005) has argued that when we analyze the model carefully and accept certain plausible background assumptions, it in fact leads back to Nagel reduction, which it was supposed to replace. In the same vein but with different arguments, Max Kistler (2005) has argued that functional reduction requires local bridge laws that are left just as unexplained as in a Nagel reduction. John Bickle (2008, personal communication) does not criticize the model itself, but points out that it is based almost entirely on logical and metaphysical considerations, and that the examples given to support it reflect an elementary school understanding of science. In this sense, the functional model is a step *backward* from Nagelian models, which were at least based on science (though not psychology and neuroscience).

I will develop the last line of argument in more detail, and show that from the point of view of philosophy of science and scientific practice,

the functional reduction approach is unacceptable. I will focus on three salient problems of the model. 1) Where do the functional definitions of properties to be reduced come from? (2) What is the "realization" relation between the property to be reduced and the reducing properties? (3) What notion of causation does the model require? These are by no means the only problems or points that need clarification, but they suffice to show why the model fails as a general account of reduction.

3.1. Functionalization

As we have seen, Step 1 in the functional model consists in defining or redefining the property to be reduced in terms of its causal role. However, it is not clear how we get the causal definition of the property to be reduced. Kim seems sympathetic to the view of Chalmers and Jackson (2001) and Levine (1993), according to which reductive explanation requires analytic definitions grounded in (a priori) conceptual analysis (see Kim, 2005, Ch. 4). The first step of functional reduction would thus consist in finding the analytic definition for the property to be reduced through conceptual analysis.

However, if the functional definition of the mental properties is to be based on conceptual analysis that is (at least relatively) a priori, this leads to a fundamental problem: our a priori ideas about psychological states or processes are often simply wrong. Consider for example memory. An armchair conceptual analysis would indicate that memory is some kind of a simple storage, where our past experiences are waiting for retrieval – Plato compared memory to an aviary of birds, from which we take the correct bird when memory retrieval is successful, and the wrong bird when it is not. However, scientific research has revealed that memories are not just retrieved, but actively constructed, and subjectively compelling memories sometimes turn out to be radically inaccurate. Furthermore, memory comprises several subsystems (short term memory, long term memory, episodic memory, visual memory, etc.), which neither individually nor taken together correspond to the simple storage envisioned by *a priori* analysis (see Bechtel (2008, Ch. 2) for a detailed philosophical analysis of memory research). Similar considerations apply to pain (Hardcastle, 2001), which has for decades been a standard example in philosophy of mind.

It is thus clear that mere conceptual analysis is not sufficient for working the properties "into shape" for reduction. One has to either allow

for scientific revision of common sense definitions of mental properties, or simply focus on properties as defined by empirical psychology.[4]

Furthermore, in both cases we have to allow for the revision and adjustment of the definitions as science proceeds. Such revision and interplay across levels is commonplace in science. One of the first philosophers to emphasize the importance of this co-evolution of concepts and theories was William Wimsatt, drawing from scientific practice in biology:

> A lower-level model is advanced to explain an upper-level phenomenon which it doesn't fit exactly. This leads to a closer look at the phenomenon, and perhaps results in some change in the way in or detail with which it is described. This will also lead to changes in the lower level model and may suggest new phenomena to look for. (Wimsatt, 1976, 231)

Also Bechtel and Richardson (1993) have described in detail the complexities involved in characterizing the phenomena to be explained in biology, based on detailed analyses of cases from history of biology, and one of their points is that scientists often have to constantly redefine the phenomena they are trying to explain. More broadly speaking, in the mechanistic explanation paradigm (Bechtel and Richardson, 1993, Machamer et al., 2000; Craver, 2007; Bechtel, 2008), a crucial point is that there is constant interplay between different levels of explanation, and both top-down and bottom-up influences.

There is also a further problem related to functionalization, even if take empirical psychological properties to be the targets of reduction and allow for constant revision of their functional definitions. It is quite possible that in the end we are unable to find any neuroscientific properties playing the causal role of some psychological properties, and thus we cannot functionally reduce them. The easiest solution in these cases would be to revise the functional definitions of the psychological properties, but this is not always justifiable. We might want to retain some psychological properties more or less as they are, since they are useful in scientific explanations. For example, Khalidi (2005) takes up the psychological property of fear, and shows (based on empirical results in cognitive neuroscience) that distinctions made at the neurophysiological level cross-cut the distinctions made at the psychological level. That is, from the vantage of neurophysiology, there is nothing playing the functional role associated with the psychological state of fear. Impor-

tantly, this is not a case of multiple realizability, which is a one-to-many relationship. In this case, there is simply just mismatch: a "one-to-none" relationship. However, we would not want to eliminate or revise the psychological concept of fear, since it still plays an important role in research and scientific explanations.

In this case, it seems that there are no neurophysiological states playing the causal role of fear, and the option of redefining fear does not seem very fruitful. Hence, Step 2 in functional reduction of fear fails. But should we conclude from this that fear is fundamentally irreducible and threatened by the exclusion argument? Or should we eliminate the property of fear from our ontology? Both options seem implausible. The framework of functional reduction seems unsuitable for dealing with situations like this.

Certainly the basic idea that the properties to be reduced have to specified causally is correct and in accordance with scientific practice. However, functionalization is not just a matter of conceptual analysis, it is not even remotely an a priori matter, and functional definitions can change as research proceeds. Furthermore, in some cases we might not be able to find neural realizers that play the functional role definitive of a mental property. This does not mean that Kim's functional model is fundamentally wrong, but it surely is too crude and simplified.

3.2. Realization

The second step in Kim's account of functional reduction is finding the "realizers" of the functionally defined property to be reduced. But what makes some property a realizer of another property? How should we understand this realization relation? And what sorts of things are the realizers of mental properties?

The roots of talk of "realization" in philosophy of mind go back to multiple realizability. Hilary Putnam (1967) famously argued that it is extremely plausible that a given mental state (like "being in pain") can be realized by different physical-chemical states in different organisms. In the debate that followed, very little attention was paid to the notion of realization itself. However, as several philosophers have recently shown (e.g., Polger, 2000; 2004; Shapiro, 2004), the realization relation is much more problematic than has been generally assumed. For example, a computer realizing an abstract algorithm or computation can hardly involve the same realization relation as a brain realizing a mental state,

since mental states are thought to be individuated causally, but abstract algorithms or computations are not individuated causally (Polger, 2004; 2007). This means that there might be no general realization relation that applies to all the different cases that are presented as typical cases of realization. However, I will not pursue this line of argumentation here, since others (i.e., Polger and Shapiro) have already elaborated it in detail. It might also be that Kim's account does not need any general notion of realization, and that a more "local" notion would suffice. In this section I will show that even if we limit the discussion to psychological properties and their realizers, and accept that there is no general notion of realization, Kim's notion of realization leads to problems.

Let us consider the case of mental properties and their neural realizers. The mental properties are to be functionally defined in terms of their causal relations to other mental properties. What is it then for a neural property to realize a mental property? According to Kim, the realizers have to perform the causal task specified in Step 1, that is, they have to "occupy" or "fill" or "play" the causal role definitive of the mental property.

But what does this mean? If we take the realizers to be properties, it seems that the only way to make sense of this is that the realizing neural property has to be embedded in a causal structure that is isomorphic to the causal structure in which the mental property is embedded. That is, the causal "context" of the neural property has to be isomorphic to the causal "context" of the mental property. What else could it mean for the neural property to occupy the causal role definitive of the mental property?

However, this leads to problems, since Kim's aim is to reduce all (non-phenomenal) psychological properties, not just one of them. This implies that, in order to accomplish a psychoneural reduction, we would have to figure out the causal structure of all the mental properties we want to reduce, and then find an isomorphic causal structure among the neural properties. If we also assume that laws underlie causal relations, and that theories are sets of laws (both assumptions are controversial, but commonly accepted in philosophy of mind), the implication is that Kim's model comes very close to theory reduction: in order to reduce a psychological theory, we need to find in (or derive from) the neuroscientic theory a structure that is isomorphic to the psychological theory. This is not so different from the "New Wave" model of psychoneural

reduction (e.g., Hooker, 1981; Bickle, 1998), where a psychological theory is reduced by deriving from neuroscience an "analogue" or "equipotent image" that is isomorphic to the psychological theory.

Marras (2002; 2005) makes a similar point with a somewhat different reasoning: in a closer analysis, Kim's model turns out to be a model of intertheoretic reduction. If this is the case, the functional model only appears to be an advance over the intertheoretic models, and faces exactly the same problems (see section 2 above).

Another fundamental problem with Kim's notion of realization was already briefly mentioned at the end of section 2: if we accept multiple realizability, the realized properties have to be either identical to the disjunction of the realizers, or just concepts (or predicates or designators). Kim rejects the first option for philosophical reasons and accepts the second one. However, in the context of realization, the problem with second option is that it seems to leave no room for the idea that neural properties realize mental properties. According to the second option, the mental concepts simply (non-rigidly) designate different neural properties in different species, just like in Lewis' (1972) filler-functionalism. If this is true, there is no realization relation here. Mental properties cannot be realized, since *there are no* mental properties, just mental concepts (or property designators) that group physical properties in interesting ways.[5] And mental concepts cannot be realized, since concepts in general are not the sorts of things that are realized. But if this is the case, the whole talk of realization has been misleading, and the claim that the functional model can accommodate multiple realizability turns out false.[6]

Perhaps, however, there are yet other ways of understanding realization. As Polger and Shapiro (2008) have pointed out, one problematic assumption that underlies many of these issues is the assumption that the realizers have to be *properties.* Particularly in more recent writings, Kim himself has been less strict and allows the realizers to be *mechanisms*: "Find the properties (*or mechanisms*) in the reduction base that perform the causal task *C*" (Kim, 2005, 102, my emphasis).

If we (unlike Kim) take this idea of mechanistic realization seriously, it leads to a more complicated picture of mental realization than the one the functional model presents. The idea is that a functionally (causally) defined psychological state, property, or capacity is realized by a neural mechanism that plays that functional role.[7] A crucial aspect of this kind

of mechanistic realization is the *multilevel* nature of the mechanisms: on any reasonable understanding of neural mechanisms, they have to be hierarchically organized into levels. Therefore, instead of a simple two-level model with the mental property and its neural realizers, we have a more complicated picture where the realizer is also organized into levels.

An often-cited example of a psychological property or capacity that is realized by a multilevel neural mechanism is memory consolidation (Craver, 2007). Memory consolidation can be functionally defined in psychological terms as the transformation of short-term memories into long-term memories. A key component in the neural mechanism realizing it is Long Term Potentiation (LTP), a well-studied cellular and molecular phenomenon that exhibits features that are closely connected to memory consolidation. Craver (2007, 165–170) defines the following four levels in the case of spatial memory and LTP: the level of spatial memory, the level of spatial map formation, the cellular-electrophysiological level, and the molecular level. They are levels of composition, where the relata are behaving mechanisms at higher levels and their components at lower levels. Levels of mechanisms in general are local and case-specific, and not intended as universal divisions of nature or science.

In fact, instead of making a distinction between realized properties and realizer properties, it is more appropriate to consider psychological properties simply as higher-level properties of neural mechanisms. For example, it is quite natural to consider psychological properties of memory consolidation as properties at the highest level of the memory consolidation mechanism. In this sense, they are neither identical to the realizing mechanism nor "just concepts" – they are real higher-level properties. The general idea is (roughly) that psychology defines and discovers functional properties that are then integrated into multilevel mechanistic explanations (and undergo "co-evolution" as science proceeds in different disciplines).

In this account, there is no special metaphysical realization relation. Indeed, no such relation is needed for understanding reductive explanation (see section 4). Whether there is multiple realizability in the sense of one-to-many mappings from higher to lower levels is an issue that has little to do with the possibility of reductive explanation.[8] Talk of realization can be preserved, as long as it is understood in a weak or metaphoric sense: a functionally defined psychological property is

"realized" by the underlying neural mechanism in the sense that the activity of the mechanism constitutes the function that is definitive of the psychological property.

The key requirement for "realization" in the functional model is that it would somehow save mental causation. In Kim's account, the only way this could work is that the "realized" properties turn out to be just concepts. Now if we adopt the mechanistic approach to realization, and see the realizers as multilevel mechanisms, what happens to mental causation? Aren't the multilevel mechanisms problematic regarding causation? In the next section, I will argue that the answer is no.

3.3. Causation

As we have seen, the properties to be reduced are defined by their *causal* roles; they are reduced by finding the first-order properties that have that *causal* role; the aim of functional reduction is to save mental properties from the *causal* exclusion argument; reduced properties have no *causal* powers of their own, and so on. Causal notions seem to play a key role in Kim's account. Indeed, the whole motivation for developing the functional model comes from the causal exclusion argument and from worries regarding the causal efficacy of mental properties. But what is causation? What does it mean to say that *X* causes *Y*?

The kind of notion causation Kim has in mind is very strong and robust:

> We care about mental causation, it seems to me, chiefly because we care about human agency, and evidently agency involves a productive/generative notion of causation. An agent is someone who brings about a state of affairs for reasons. If there indeed are no productive causal relations in the world, that would effectively take away agency – and our worries about mental causation along with it. (Kim, 2009, 44)

As the quote indicates, Kim thinks of causation as a relation where the cause generates, produces, or brings about the effect. According to Kim, a weaker account of causation in terms of, for example, counterfactual relations would not be satisfactory, since we would still need the metaphysical account of what makes the counterfactuals we want for mental causation true (Kim, 1998, 71).

In the next section, I will first briefly present one such weaker account of causation, and then argue that, *contra* Kim, this is all we need for understanding mental causation. In the section after that, I will argue

that on this account the causal exclusion argument does not threaten mental causes, and thus a large part of the motivation for the functional reduction of mental properties fades away.

3.3.1. Interventionist Causation

In recent years, several philosophers have presented accounts of causation in terms of interventions and manipulability (Pearl, 2000; Woodward, 2003; 2008; Woodward and Hitchcock, 2003; also Spirtes, Glymour, and Scheines, 1993). I will focus here on James Woodward's (2003) version, which is exceptional in its scope and clarity. The guiding insight of the account is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. To put it very roughly, in this model a necessary and sufficient condition for X to cause Y or to figure in a causal explanation of Y is that the value of Y would change under some intervention on X (in some background circumstances).

An intervention can be thought of as an (ideal or hypothetical) experimental manipulation carried out on some variable X (the independent variable) for the purpose of ascertaining whether changes in X are causally related to changes in some other variable Y (the dependent variable). Of course, several restrictions on interventions must be added – see Woodward (2003) for details. Interventions are not only human activities, there are also "natural" interventions, and the definition of an intervention makes no essential reference to human agency. This sets the interventionist account clearly apart from previous manipulability theories of causation (e.g., Menzies and Price, 1993).

According to Woodward, causal relationships are relationships that are *invariant* under interventions. Suppose that there is a relationship between two variables that is represented by a functional relationship $Y = f(X)$. If the same functional relationship f holds under a range of interventions on X, then the relationship is invariant within that range. For example, the ideal gas law "$PV = nRT$" continues to hold under various interventions that change the values of the variables (P, V, and T), and is thus invariant within this range of interventions. One consequence of this model is that relata of causation must be represented as variables, but states or properties can easily be represented as binary variables, such that, for example, 1 marks the presence of the property and 0 the absence of the property.

This framework captures the nature of causation as *difference-making*: if variable *X* is causally relevant for variable *Y*, changes in the value of variable *X* make a difference in the value of variable *Y* (in a range of circumstances). Interventionist causation is also essentially *contrastive*: It is *X*'s taking some value *x* instead of *x*' that causes *Y*'s taking value *y* instead of *y*'.

The interventionist account accords well with the way causal notions are employed in the special sciences (Woodward, 2000; 2003; 2008). The account has also received broad acceptance among both philosophers and scientists. However, it seems to provide exactly the kind of "weak" notion of causation that Kim finds unsatisfactory. Kim is after a productive or generative notion of causation that is more metaphysically robust.

The main problem with such a stronger notion is that it would have to be somehow grounded in physics. In the end, the metaphysical question that Kim wants to answer is how there could be mental causes in a fundamentally physical world. If the stronger notion of causation was *not* grounded in physics, it is hard to see what reason there would be to prefer it to the interventionist account, assuming that the latter captures the notion of causation as it is needed in science and everyday life.

The problem with grounding causation in physics is that notions like cause and effect do not really play a role in our best physical theories (as famously argued by Bertrand Russell (1912-13), and more recently by Ladyman and Ross (2007), Loewer (2007), Norton (2007), and many others). The fundamental laws of physics relate the totality of a physical state at one time to the totality of the physical state at later instants, but do not single out causes and effects among these states. If we want to find causes that "bring about" or "produce" their effects, or causes that are "sufficient" for their effects, we have to consider something like the entire state of the universe as the cause for even a small effect.[9]

Of course, we can put labels onto relata that appear in physical equations and call some of them causes and others effects, but this is entirely superfluous to the physics itself. There is no "principle of causality" that would in any way guide or restrict physical theory formation. Furthermore, there are cases even in Newtonian physics that go straight against our ideas of causation – for instance, effects that take place with no observable causes (Norton, 2007) – not to even speak of phenomena like quantum entanglement.

The interventionist account seems to capture the nature of causation both in special sciences and everyday life very well, and in fundamental physics, causal notions are unnecessary and superfluous.[10] It then seems that the interventionist account, insofar as it is successful, gives us all we want from an account of causation. A philosopher of mind could still insist that the question of what causation *really* is has to be answered. But from a scientific point of view, this search for the true nature of causation can be seen as just a metaphysical exercise. As Woodward (2008, 249) puts it: "We are thus left with possibility that the only people who think that vindicating the claim that mental states are causes requires showing that they are causes in a richer, more metaphysical sense are certain philosophers of mind."

3.3.2. Causal Exclusion in the Interventionist Framework

What are the consequences of the interventionist account for mental causation? Prima facie, it seems that mental causation is unproblematic in the interventionist framework. There are invariant psychological generalizations such that we can make interventions to mental states in order to change other mental states or physical behavior. For example, as Woodward (2008) points out, when you persuade someone, you manipulate her beliefs by providing information or material things, in order to change her other beliefs. Also many psychological and social science experiments involve intervening on the beliefs of the subjects, usually through verbal instruction, in order to change some other beliefs and observable behavior.

Upon closer philosophical analysis it appears that the interventionist account indeed vindicates mental causation. Several authors (e.g., Menzies, 2008; Raatikainen, 2010) have recently come up with an argument that claims to show that if the interventionist account is correct, mental states can be causes of physical behavior, and they are not excluded by their physical realizers. This is due to the fact that causation in the interventionist account is a matter of difference-making, and not a matter of physically producing or bringing about the effect. The difference-making cause of a physical event, like a hand movement, can be a mental cause, and it is not excluded by some physical cause. Therefore, the exclusion principle does not hold or turns out to be nonsensical in the interventionist framework. On the other hand, Michael Baumgartner (2010) and Vera Hoffmann-Kolss (unpublished manuscript) have argued

that there is an interventionist version of the exclusion argument, and thus adopting the interventionist account does not make the problem of exclusion go away.

Instead of going through the details of these arguments, I will argue that there is a deeper underlying problem that applies to the arguments of participants at both sides of the debate. The problem is that typical causal representations of the mental and the physical causes fail to satisfy the *Causal Markov condition*.[11]

According to one formulation that is the most fitting one in the present context, the Causal Markov condition states (CM): conditional on its direct causes, each variable is independent of every other variable except its effects.[12] In other words, variables that are not related as cause or effect or as effects of a common cause have to be uncorrelated. It is widely agreed that when the causal relationships in a system are correctly and fully represented, CM will be satisfied. Furthermore, the condition *follows* from the interventionist definition of causation and some other plausible assumptions[13] (Hausman and Woodward, 1999). Hence, in a full and correct interventionist causal representation of a system, CM has to be satisfied.

Typical representations of mental causation in philosophy of mind, including the one applied in Kim's exclusion argument (section 2), *fail* to satisfy CM (see Figure 1). In these representations, mental property *M* causes another mental property *M**, physical (or neural) property *P* causes another physical (or neural) property *P**, *M* supervenes on *P*, and *M** supervenes on *P**. Due to supervenience, the values of *M* and *P* are correlated, and *M* depends on *P*. Whenever *M* changes, *P* also changes, and when the value of *P* is fixed, the value of *M* is also fixed.[14] However, *M* does not cause *P*, *P* does not cause *M*, and they are not both effects of a common cause. Hence, CM is violated. This means that something has gone wrong in building the causal representation of the system.

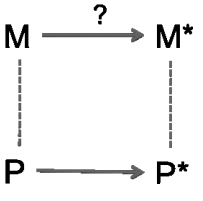


Figure 1: A typical representation of mental causation in philosophy of mind. The arrows represent causation, the dotted lines represent supervenience.

There are (at least) the following three ways of reacting to this problem complex. (1) The reductive solution: get rid of the mental variables, either by identifying them with physical variables or simply eliminating them. (2) The nonreductive solution: fix the level of analysis when building the causal representation, and never include supervenient variables in the same representation with their supervenient base variables. (3) Argue that this is a problem for the interventionist account, and that it needs to be replaced or revised (e.g., by adding some additional principles for dealing with supervenient variables).

The problem with the reductive solution is that if we accept it, we can just as well apply the same reasoning to nonmental variables, which leads to undesirable consequences. All that is required for the argument to work is that there is a supervenience relation between the variables, and supervenience relations can be found all over the place, also in biological, chemical, and even physical contexts. We can also consider the fact that the neural properties (variables) supervene on biochemical or some other lower-level properties (variables). Therefore, we can simply draw the same picture again, replacing mental variables by neural variables and neural variables by biochemical variables. Then it seems that since we got rid of the mental variables in the first case, we also have to get rid of the neural variables in the second case. Causation seems to be draining away towards some fundamental physical level, which is particularly strange if we consider the fact that there seems to be no causation at the fundamental physical level (see previous section).

This is a version of the *generalization argument* that has often been raised against Kim's exclusion argument (e.g., Block, 2003; Van Gulick, 1992). The generalization argument states that if Kim's reasoning about mental properties is correct, then we can apply it to all higher-level or nonfundamental properties, which then are excluded. However, this is an absurd conclusion, so there has to be something wrong with Kim's argument. Kim has provided several answers to the generalization argument, but it is widely agreed that none of them is satisfactory (see, e.g., Walter, 2008). What a defender of the exclusion argument (also the interventionist version) would have to show is that there is some principled reason why mental properties (variables) are excluded but other higher-level or macroproperties are not. Until such a reason is provided, the reductive solution is a nonstarter.

The *nonreductive* solution would be to allow higher-level causal representations, but not allow including the supervenient base variables in the same representation. For example, we would not include neural variables in the same representation as the mental variables. We would have a plurality of causal representation, but not representations that include both supervenient variables and their base variables. As Hausman and Woodward (1999, 531) put it in a different context: "One needs the right variables or the right level of analysis – variables that are sufficiently informative and that are not conceptually connected."

This solution is attractive and close to scientific practice, and I think ultimately something like this approach is the right way to go.[15] However, there are at least two problems. First of all, there seems to be an element of arbitrariness or *ad hoc* here, since the only reason for not including the supervenience base variables is that it would violate the Causal Markov condition. Secondly, there might be cases where we would like to include supervenient variables and their base variables in the same representation. If it turns out there are serious and scientifically relevant cases like that, it means trouble for the nonreductive solution.

Defending the nonreductive solution also requires showing what exactly goes wrong in the exclusion argument. At least one of the principles appealed to in the argument has to turn out false. Due to constraints of space, I cannot go into the details here, but the most likely candidate is the exclusion principle, which becomes highly problematic when it is formulated in interventionist terms (see Menzies (2008) and Raatikainen (2010) for more). This is again due to the fact that interventionist causation is a matter of difference-making, not of physically producing the effect.

The third solution is to argue that the interventionist account of causation is deficient, and that we need to replace it, or at least revise it, for example by adding some further rules or principles for dealing with cases of supervenience. Baumgartner (2010) argues that the exclusion argument is indeed a fundamental problem for the interventionist account, and is skeptical regarding possible revisions. However, one argument against this solution is that if the problem arises only in an abstract philosophical context, like the problem of mental causation in philosophy of mind, it might be that the abstract philosophical context needs to be revised, not the interventionist account, which takes us back to options (1) and (2). Again, it remains to be seen how common or relevant are the

cases where we want to include also supervenience base variables in the representation.

To summarize, Kim's argument does hold in one sense even in the interventionist framework: it shows that causal claims become very problematic when conjoined with supervenience claims. However, if it is understood as an argument to the effect that mental causation is not possible, or is more problematic than other kinds of causation, it does not hold.

Thus, with a correct understanding of causation, a large part of the motivation behind functional reduction disappears. Kim wanted to show that mental properties are functionally reducible in order to save mental causation. However, it seems that mental causation does not need such a rescue operation: mental causation in the interventionist sense is no more problematic than any other kinds of causation, and the search for more metaphysical (productive, generative, sufficient, etc.) mental causes is pointless.

What is then the motivation for reducing or reductively explaining the mental? I think the correct answer is that we want to reductively explain the mental because we want to explain everything there is to explain, and some kind of reductive explanation seems to be very fruitful in this context, as the success of neuroscience in recent decades shows. But what exactly is the nature of this explanatory enterprise?

4. Functional Reduction as Mechanistic Explanation

Perhaps the functional model could be revised, taking into account all that has been said above, in roughly the following way. We want to reduce mental property *M*. First, we have to find out what the functional role of *M* is. However, this is not done through conceptual analysis alone, but through the interplay of conceptual analysis and empirical research. Also, it is an ongoing process, and the initial definitions may be refined later. This first step is not necessarily temporarily prior to the next steps, and anyway the whole process is integrated and all the steps are intertwined. In the second step, we figure out what the neural mechanism that is the "realizer" of *M* is. *M* is neither identical to its realizer nor "just a concept" – the realizing mechanism is structured into levels, and *M* can be seen as a higher-level property of the mechanism. Third, we construct

the "theory" that explains why the mechanism is the realizer of *M* – that is, we show how the functioning of the mechanism results in *M* (i.e., how the mechanism performs the functional role of *M*).

This quickly sketched revised account of functional reduction looks very much like what in philosophy of science is known as *mechanistic explanation*. The key idea of the mechanistic explanation paradigm (Bechtel, 2008; Bechtel and Richardson, 1993; Craver, 2007; Machamer et al., 2000) is that if one takes into account actual scientific practice in neuroscience and many of the life sciences, it turns out that instead of focusing on laws or formalizable theories, practicing scientists formulate explanations in terms of mechanisms.

According to an often-cited definition, mechanisms are to be understood as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, 3). Or, as Bechtel (2008, 13) puts it, a "mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization." A mechanistic explanation then describes how the orchestrated functioning of the mechanism is responsible for the phenomenon to be explained.

This suggests that to functionally reduce a property *M* amounts to providing a mechanistic explanation for *M*. The upshot is that if we want to keep the model of functional reduction close to science, it turns out that there is no functional reduction over and above mechanistic explanation.

What does replacing the functional model with mechanistic explanation mean for the questions of reduction and causation? The mechanistic explanation model, conjoined with the interventionist account of causation, does not involve the kind of strong ontological reduction in terms of property identities or eliminations that Kim is after, since it emphasizes the *multilevel* nature of mechanisms, and the causal and explanatory relevance of higher-level things. However, it is important to remember that the main reason for being an ontological reductionist (at least for Kim) is the causal exclusion argument. If the exclusion problem does not arise when we understand causation in interventionist terms, then also the motivation for being a strong reductionist fades away.

Many philosophers (e.g., Bechtel, 2008; Sarkar, 1992; Wimsatt, 1976) have argued that the process of "looking downward" and invoking parts of the mechanism to understand the behavior of the mech-

anism as a whole is close enough to what scientists generally take to be a reductive explanation to warrant treating the downward-looking aspect of mechanistic explanation as a kind of reductive explanation. On the other hand, Craver (2007) considers the framework of mechanistic explanation antireductive. This issue is mainly a terminological one, but I see no harm done calling downward-looking mechanistic explanation reductive explanation, as long as it is clearly distinguished from stronger forms of reduction. Regardless of whether we want to call mechanistic explanation reductive explanation, this approach supports a kind of causal and explanatory pluralism: higher-level entities or properties (including psychological entities and properties) do have causal and explanatory relevance, and are not reducible in any strong sense to lower-level entities and properties.

To conclude, functional reduction fails as a general account of reduction in philosophy of mind: if we try to understand it in a scientifically credible way, it effectively gives way to mechanistic explanation, which in turn leads to causal and explanatory pluralism. Whether this is compatible with "physicalism, or something near enough" (Kim, 2005) is an open question that has still to be addressed.

Acknowledgements

I am very grateful to Dan Brooks, Vera Hoffmann-Kolss, Jani Raerinne, Max Seeger, Achim Stephan and an anonymous referee for comments on earlier versions of this article. The article is based on a presentation at the workshop "Reductionism, Explanation and Metaphors in the Philosophy of Mind" in September 2009 in Bremen, organized by Albert Newen and Raphael van Riel.

Notes

1 As an anonymous referee pointed out, one could argue that Nagelian reduction involves only the deduction of *laws*, which does not as such require formal theories. However, this only leads to a parallel problem: laws in the sense of generalizations that fill the traditional criteria for laws are not central in psychological and neuroscientific theories and explanations (Craver, 2007; Cummins, 2000; Machamer et al., 2000; Woodward, 2000).

2 Marras (2002; 2005), however, argues that functional reduction in fact collapses back to Nagelian reduction. I return to this in section 3.2.

3 Esfeld and Sachse (2007) have argued that by introducing functional subtypes we can have property identities and conservative functional reductions, multiple realizability notwithstanding.

4 This problem is obviously related to the issue of common-sense (analytical) vs. empirical functionalism (psychofunctionalism).

5 Perhaps one solution would be to argue that mental properties are some special kind of "abstract" properties. However, Kim does not appear to seriously consider such a solution. In any case, it would require developing or spelling out the metaphysics for such properties, which is no easy task.

6 In fact, Kim sometimes seems ready to reject the multiple realizability of mental properties and argues for "species-specific identities", such that "multiply realized properties are sundered into diverse realizers in different species and structures" (Kim, 1998, 105). This leads to problems if there is also multiple realizability within species or structures: it seems to follow that mental properties are spliced into properties restricted to very specific neural or physical structures, and it is hard to see how such properties could be relevant in scientifically explaining human behavior.

7 Wilson and Craver (2007) have recently defended a mechanistic approach to realization. They argue that this also comes close to how the term "realization" is often used in the cognitive sciences and neurosciences: when scientists say they are looking for the neural realization of memory consolidation, what they typically mean is that they are looking for the neural mechanism of memory consolidation. The approach of Wilson and Craver is promising, but remains rather provisional and schematic.

8 In section 4 I argue that we should understand reductive explanation as downward-looking mechanistic explanation. If there are one-to-many mappings from psychological properties or functions to the underlying mechanisms, this is no obstacle to downward-looking mechanistic explanation of those properties or functions. In these cases, different mechanisms can perform the same roughly defined function, and therefore there are different mechanistic explanations for this function. There is nothing problematic about this.

9 Perhaps it is sufficient to consider the state of the universe on the surface of a sphere with a radius of about 300000000 meters centered on the effect, assuming that the cause precedes the effect by one second – the speed of causal influence cannot be faster than the speed of light. Of course, this does not make the idea of productive physical causation any less problematic. See Loewer (2007) for more.

10 As an anonymous referee pointed out, not all philosophers of physics agree that there is no causation in fundamental physics (see, e.g., Frisch, 2009). However, even if it turns out that causal notions do play a role in fundamental physics, it is still the case that there is currently no metaphysically robust and physically grounded notion of causation that would be suitable for considering mental causation and a serious alternative to interventionist causation.

11 This was pointed out to me by Dan Brooks, for which I am very grateful.
12 See Hausman and Woodward (1999) for other formulations and an extensive discussion of the Causal Markov condition. Another condition that is also extensively covered in the same paper, and that could perhaps also be used as a basis for the arguments in this section, is modularity: a system consisting of several causal relationships is modular to the extent that these various causal relationships can be changed or disrupted while leaving the others intact. Both CM and modularity have been under intense discussion in recent years – see, for example, Cartwright (2002) or Steel (2006).
13 Alternatively, it could be said that the interventionist definition follows from CM and some other plausible assumptions. Without (something like) CM it is impossible to talk of variables being causal in the interventionist sense.
14 According to a standard definition, a set of A-properties supervenes on a set of B-properties if and only if two things cannot differ with respect to their A-properties without also differing with respect to the B-properties.
15 Recently Shapiro and Sober (2007) have also argued that supervenient causes are problematic in the interventionist framework and defended a nonreductive solution. Let us consider a situation where we want examine whether M, which supervenes on P, is a cause of physical behavior B. We have to make an intervention on M such that other causes of B, including P, remain unchanged. The problem is that this is impossible, since the value of P determines the value of M (due to supervenience). It is not acceptable or nomologically possible to wiggle M while holding P fixed. Hence, this must be a wrong way of conceptualizing the situation.

Bibliography

Baumgartner, Michael, 2010: Interventionism and epiphenomenalism. In: *Canadian Journal of Philosophy* 40, pp. 359–383.

Bechtel, William, 2008: *Mental Mechanisms*. London: Routledge.

Bechtel, William, and Robert C. Richardson, 1993: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bickle, John, 1998: *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, John, 2008: Real Reduction in Real Neuroscience: Metascience, Not Philosophy of Science (and Certainly Not Metaphysics!). In: Hohwy, J.; Kallestrup, J. (eds.) *Being Reduced*. Oxford: Oxford University Press, pp. 34–51.

Block, Ned, 2003: Do Causal Powers Drain Away? In: *Philosophy and Phenomenological Research* 67, pp. 133–150.

Cartwright, Nancy, 2002: Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. In: *British Journal for the Philosophy of Science* 53, pp. 411–453.

Chalmers, David J., 1996: *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, David J., and Frank Jackson, 2001: Conceptual Analysis and Reductive Explanation. In: *The Philosophical Review* 110, pp. 315–360.

Craver, Carl F., 2007: *Explaining the Brain*. Oxford: Oxford University Press.

Cummins, Robert, 2000: "How Does it Work?" versus "What Are the Laws?" Two Conceptions of Psychological Explanation. In: Keil, F.; Wilson, R. (eds.): *Explanation and Cognition*. Cambridge: MIT Press, pp. 117–144.

Esfeld, Michael, and Christian Sachse, 2007: Theory Reduction by Means of Functional Sub-types. In: *International Studies in the Philosophy of Science* 21, pp. 1–17.

Frisch, Mathias, 2009: 'The Most Sacred Tenet'? Causal Reasoning in Physics. In: *British Journal for the Philosophy of Science* 60, pp. 459–474.

Hardcastle, Valerie, 2001: The Nature of Pain. In: Bechtel, W.; Mandik, P.; Mundale, J.; Stufflebeam, R.S. (eds.): *Philosophy and the Neurosciences: A Reader.* Malden, MA: Blackwell, pp. 295–311.

Hausman, Daniel M., and James Woodward, 1999: Independence, Invariance and the Causal Markov Condition. In: *British Journal for the Philosophy of Science* 50, pp. 521–583.

Hoffmann-Kolss, Vera (unpublished manuscript). The Supervenience Argument Is Alive and Kicking.

Hooker, Clifford A., 1981: Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction. In: *Dialogue* 20, pp. 38–59, 201–236, 496–529.

Khalidi, Muhammad A., 2005: Against Functional Reductionism in Cognitive Science. In: *International Studies in the Philosophy of Science* 19, pp. 319–333.

Kim, Jaegwon, 1992: Multiple Realization and the Metaphysics of Reduction. In: *Philosophy and Phenomenological Research* 52, pp. 1–26.

Kim, Jaegwon, 1998: *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, Jaegwon, 1999: Making Sense of Emergence. In: *Philosophical Studies* 95, pp. 3–36.

Kim, Jaegwon, 2002: Mental Causation and Consciousness: The Two Mind-body Problems for the Physicalist. In: Gillett, C; Loewer, B. (eds.): *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, pp. 271–283.

Kim, Jaegwon, 2005: *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kim, Jaegwon, 2008: Reduction and Reductive Explanation: Is One Possible without the Other? In: Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 93–114.

Kim, Jaegwon, 2009: Mental Causation. In: McLaughlin, B; Beckermann, A.; Walter, S. (eds.): *The Oxford handbook of philosophy of mind*. Oxford: Oxford University Press, pp. 29–52.

Kistler, Max, 2005: Is Functional Reduction Logical Reduction? In: *Croatian Journal of Philosophy* 14, pp. 219–234.

Ladyman, James, and Don Ross, 2007: *Every Thing Must Go: Metaphysics Naturalised*. Oxford: Oxford University Press.

Levine, Joseph, 1993: On Leaving Out What It's Like. In: Humphreys, G.; Davies, M. (eds.): *Consciousness*. Oxford: Blackwell, pp. 121–136.

Lewis, David, 1972: Psychophysical and Theoretical Identifications. In: *Australasian Journal of Philosophy* 50, pp. 249–258.

Loewer, Barry, 2007: Mental Causation, or Something Near Enough. In: McLaughlin, B.; Cohen, J. (eds.): *Contemporary Debates in Philosophy of Mind*. Malden, MA: Blackwell Publishing, pp. 243–264.

Machamer, Peter K., Lindley Darden, and Carl Craver, 2000: Thinking About Mechanisms. In: *Philosophy of Science* 67, pp. 1–25.

Marras, Ausonio, 2002: Kim on Reduction. In: *Erkenntnis* 57, pp. 231–257.

Marras, Ausonio, 2005: Consciousness and Reduction. In: *British Journal for the Philosophy of Science* 56, pp. 335–361.

McCauley, Robert N., 2007: Reduction: Models of Cross-scientific Relations and their Implications for the Psychology-Neuroscience

Interface. In: Thagard, P. (ed.): *Handbook of the Philosophy of Psychology and Cognitive Science*. Amsterdam: Elsevier, pp. 105–158.
Menzies, Peter, 2008: The Exclusion Problem, the Determination Relation, and Contrastive Causation. In: Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 196–217.
Menzies, Peter, and Huw Price, 1993: Causation as a Secondary Quality. In: *The British Journal for the Philosophy of Science* 44, pp. 187–203.
Nagel, Ernest, 1961: *The Structure of Science*. London: Routledge & Kegan Paul.
Norton, John D., 2007: Causation as Folk Science. In: Price, H.; Corry, R. (eds.) *Causation, Physics, and the Constitution of Reality. Russell's Republic Revisited.* Oxford: Oxford University Press, pp. 11–44.
Pearl, Judea, 2000: *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.
Polger, Thomas W., 2004: *Natural Minds.* Cambridge: MIT Press.
Polger, Thomas W., 2007: Realization and the Metaphysics of Mind. *Australasian Journal of Philosophy* 85, pp. 233–259.
Polger, Thomas, and Lawrence Shapiro, 2008: Understanding the Dimensions of Realization. In: *The Journal of Philosophy* 105, pp. 213–222.
Putnam, Hilary, 1967: Psychological Predicates. In: Capitan, W. H.; Merrill, D. D. (eds.): *Art, Mind, and Religion*. Pittsburg: Pittsburg University Press, pp. 37–48.
Raatikainen, Panu, 2010: Causation, Exclusion, and the Special Sciences. In: *Erkenntnis* 73, pp. 349–363.
Russell, Bertrand, 1912–1913: On the Notion of Cause. In: *Proceedings of the Aristotelian Society* 13, pp. 1–26.
Sarkar, Sahotra, 1992: Models of Reduction and Categories of Reductionism. In: *Synthese* 91, pp. 167–194.
Shapiro, Lawrence A., 2004: *The Mind Incarnate*. Cambridge, MA: MIT Press.
Shapiro, Lawrence A., and Elliott Sober, 2007: Epiphenomenalism – the Do's and the Don'ts. In: Wolters, G.; Machamer, P. (eds.): *Thinking about Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press, pp. 235–264.
Spirtes, Peter, Clark Glymour, and Richard Scheines, 1993: *Causation, Prediction, and Search*. New York: Springer.

Steel, Daniel, 2006: Comment on Hausman & Woodward on the Causal Markov Condition. In: *British Journal for the Philosophy of Science* 57, pp. 219–231.

van Gulick, Robert, 1992: Three Bad Arguments for Intentional Property Epiphenomenalism. In: *Erkenntnis* 36, pp. 311–332.

Walter, Sven, 2008: The Supervenience Argument, Overdetermination, and Causal Drainage: Assessing Kim's Master Argument. In: *Philosophical Psychology* 21, pp. 671–694.

Walter, Sven, and Markus I. Eronen (forthcoming): Reductionism, Multiple Realizability, and Levels of Reality. In: French, S.; Saatsi, J. (eds.): *Continuum Companion to the Philosophy of Science*. Continuum.

Wilson, Robert A., and Carl F. Craver, 2007: Realization: Metaphysical and Scientific Perspectives. In: Thagard, P. (ed.): *Handbook of the Philosophy of Psychology and Cognitive Science*. Amsterdam: Elsevier, pp. 81–104.

Wimsatt, William C., 1976: Reductionism, Levels of Organization, and the Mind-Body Problem. In: Globus et al. (eds.): *Consciousness and the Brain. A Scientific and Philosophical Inquiry*. New York: Plenum Press, pp. 205–267.

Woodward, James, 2000: Explanation and Invariance in the Special Sciences. In: *The British Journal for the Philosophy of Science* 51, pp. 197–254.

Woodward, James, 2003: *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, James, 2008: Mental Causation and Neural Mechanisms. In: Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 218–262.

Woodward, James, and Christopher Hitchcock, 2003: Explanatory Generalizations, Part I: A Counterfactual Account. In: *Noûs* 37, pp. 1–24.

Albert Newen

Phenomenal Concepts and Mental Files: Phenomenal Concepts are theory-based[1]

Abstract

In this paper, it is argued that phenomenal concepts – conceived as a specific kind of mental representations – should be classified as *theory-based* concepts in contrast to *perception-based* concepts. Phenomenal concepts are acquired in ontogeny, and they are based on a set of theoretical principles, i.e. a mini-theory about the status of experiences as subjective and private. Building upon the notion of a mental file, this idea of phenomenal concepts is explicated in detail on the basis of antecedent physicalism, and its explanatory power is shown at work discussing the knowledge argument. So a new notion of phenomenal concepts in terms of mental files is spelled out that completely accounts for Mary's cognitive situation from a physicalists' view.

Zusammenfassung

In diesem Beitrag wird dafür argumentiert, dass phänomenale Begriffe (verstanden als eine spezifische Art mentaler Repräsentationen) als *theoriebasierte*, und nicht als *wahrnehmungsbasierte* Begriffe aufzufassen sind. Phänomenale Begriffe werden im Laufe der Ontogenese erworben und sie basieren auf theoretischen Prinzipien – nämlich auf einer Minitheorie über den Status von Erfahrungen als subjektiv und privat. Aufbauend auf der Begrifflichkeit von *mentalen Files* wird die Idee phänomenaler Begriffe im Detail expliziert. Dabei wird ein Physikalismus vorausgesetzt. Die explanatorische Kraft der entwickelten Theorie wird am Beispiel des Argument des unvollständigen Wissens (*knowledge-Argument)* demonstriert. Die hier ausbuchstabierte Idee eines phänomenalen Begriffs erlaubt eine Interpretation der kognitiven Situation von Mary aus der Sicht des Physikalismus.

Introduction

Phenomenality, i.e. the fact that we can taste the sweetness of a cookie, and see the redness of a tomato etc., can be described as the *what-it-is-like* aspect of our experience. As such, it plays a crucial role in issues

philosophia naturalis 47-48 / 2010-11 / 1-2

regarding functional reductionism, the idea of an explanatory gap, and the question of how knowledge of phenomenal experience relates to knowledge about the physiology underlying phenomenal experience. The aim of this article is to develop a new theory of phenomenal concepts and to use this theory to account for the knowledge argument in particular. First, however, I briefly situate the status of phenomenal experiences within the recent philosophical debates.

Sometimes, it is argued that phenomenal properties resist an analysis or characterization in functional terms (Chalmers 1996). If this is correct, then phenomenal properties resist functional reduction. The basic idea behind a functional reduction of phenomenal properties can be explained as follows: In a first step, we have to describe the functional roles of a phenomenal property and in a second step, we have to identify the entity instantiating these roles at the physical level (this is an empirical task). If, however, phenomenal properties cannot be given a description in terms of functional roles, then they cannot be reduced in this way. As a result, it would be difficult to find the physical patterns constitutive for the occurrence of a phenomenal experience (the so called *explanatory gap*). Furthermore, there are also situations conceivable in which *phenomenal properties* and *physical properties* are treated as independent, contrary to what one might expect from the physicalists' point of view. A good illustration of this is the knowledge argument introduced by Frank Jackson (1982 & 1986). This argument primarily focuses on the question of how knowledge of the physiology of the brain relates to knowledge about phenomenal experience. Jackson invites us to imagine a brilliant future scientist named Mary who knows all the scientific facts about vision and color-perception there are to know. The problem is: Mary grew up in a black and white room. In fact, she never saw any colors until, one day, she leaves the black and white room and perceives a red tomato. Does she, thereby, gain new knowledge, and if so, does this show that phenomenal properties are ontologically different from the properties she already knew about? Papineau (2002, 51) sketches how physicalism might address this challenge: Even though there is a difference between Mary before she saw the red tomato and after she saw it, this difference is not to be cashed out in terms of knowing about a new property. All we have to claim in order to consistently interpret this story is that she acquired a new way of conceiving of a property she was already able to refer to before she saw a red tomato. The difference is

that in the black and white room, she had to employ "material concepts" (Papineau 2002, 51), whereas after her release she has a (full-fledged) phenomenal concept of the same property. This reconstruction of the knowledge argument is also called *the phenomenal concept strategy*.

What has troubled many as an ontological problem is now often interpreted as a *conceptual* issue (Papineau 2002 & 1993a & b; Levine 1998; Tye 1999; Block & Stalnaker 1999; Carruthers 2000). Even David Chalmers could be incorporated in this list since he believes the gap to be conceptual in a very specific sense *and therefore* (what distinguishes his view from others) ontological (Chalmers 1996). Recent discussions about the reduction of phenomenal experiences crucially build upon the notion of a phenomenal concept.[2] Therefore, the current paper contributes to these discussions by putting forward a new theory of phenomenal concepts.

To develop a new account of phenomenal concepts I will proceed as follows: First I introduce an ontological framework that is presupposed by antecedent physicalism, one that also includes the different types of representations connected with the acquisition of knowledge. Then an argument for the orthogonality of phenomenality and content is briefly discussed: phenomenal experience cannot be explained as a specific type of content. It can best be characterized as the product of a special way of information processing. This supports a naturalistic view of phenomenality. In the main part of the article a novel explication of phenomenal concepts is presented. To do so, I first develop a distinction between perception-based concepts and theory-based concepts. Furthermore, concepts are developed in terms of mental files using different types of information. *The main claim to be illustrated and defended is that phenomenal concepts are theory-based concepts.* The analysis of the phenomenal concepts allows me to account for the knowledge argument on the basis of physicalism.

1 Antecedent physicalism and types of information about properties

The methodological starting point is antecedent physicalism (Perry 2001a). This means that physicalism is presupposed and on this basis I develop an explication of phenomenal concepts which allows me to

account for all relevant aspects of the knowledge argument. For antecedent physicalists it is commonplace to presuppose that the world consists of objects, properties, events, processes etc.[3] I accept that some physical objects can have physical as well as mental properties but it is presupposed that mental properties are identical with some complex physical properties (usually involving typical neural states).

Furthermore, a theory of representation of the real world is needed. Any theory of knowledge has to account for different ways of representing an entity that belongs to the real world because there are different special sciences developing special knowledge of the same object. Here I suggest that we should distinguish sensorimotor representations, image-like representations and propositional representations according to a principled analysis of types of knowledge (Jung/Newen 2010). An object or a property of an object can be represented by a person by means of one or more types of representations. One can think of this as different modes of presentation of the same entity. A round, red ball can be captured by a sensorimotor representation on the basis of a mere touch or by a visual image or a verbal description of it. The information we receive in these cases can respectively be called sensorimotor, image-like and descriptive information. Sometimes we receive these types of information independently from each other, and sometimes we have all of them available.

In addition to entities and representations of them, we need a background theory of the interrelation between the real entities and the representations of them. Presupposing an identity theory I build upon the idea that the phenomenal property someone experiences is identical with a physical process, especially with a specific sort of processing information (and not with a type of informational content; Vosgerau, Schlicht, Newen 2008). Having a phenomenal experience is then identical with a specific type of information processing.

2 How can we account for phenomenality as antecedent physicalists?

2.1 Speaking about phenomenality

Does our talk about phenomenal experiences commit us to the existence of qualia? I introduced the (object-) property of being red and the phenomenal property of a red-experience. If "F" refers to a phenomenal

property, e.g. the red-experience, then the question is how we interpret the sentence "Subject S has F". According to one interpretation, proposed, for example, by Locke (1965, II, 1, § 4) and Armstrong (1968, 92–99; 200; 323–338), sentences of this sort are true only if subject A has an inner sense by which F is experienced. According to a rival interpretation, which has been put forward by Frege (1966, 40) and Schlick (1979, chapter 20), it is not the case that for a sentence of this sort to be true, we should assume that there is an inner sense via which we have experience F. To use an example of Wolfgang Künne (2007, 64), such sentences are to be interpreted like "He is dancing a Tango". In this case, the person we talk about is not *standing in the dancing relation to something which is extrinsic to the dance* (like the object of a sensory experience to that experience), but rather *performing a specific sort of dance*. Similarly, when we are experiencing pain, or having a red-experience, we do not sense pain or redness, but rather, we have specific sorts of experiences. If the Lockian story were correct, then we had to assume that there are qualia that we are perceiving. From a more parsimonious ontological view that I prefer and defend, we do not need independent entities called qualia but can just do with different ways of experiencing or, more generally, registering information. Our talk of experiences does not commit us to the existence of qualia as nonreducible entities. Let me explicate this in more detail by illustrating a general idea how we can account for phenomenality.

2.2 The orthogonality of phenomenality and content

Before my positive claim is defended the zombie argument is shortly analyzed since it seems to deliver a strong argument for the nonreducibility of phenomenality: zombies are to be conceived of as psychophysical duplicates of ordinary humans, which differ from humans only with respect to the phenomenal dimension. Zombies lack the phenomenal experience that humans enjoy. If this is conceivable, then it seems that qualia do not reduce to representations which can be referred to in physical explanations of our behavior. But even if we grant that, we would have to deal with the following problem: An essential feature of qualia is their irrelevance with respect to behavioral output (epiphenomenalism): differences in qualia do not influence our behavior in any way since the underlying causal mechanisms in the case of a normal human and a zombie are supposed to be the same. The zombie argument

allows only for two positions: epiphenomenal dualism or physicalism including mental causation. Epiphenomenalism would have a number of counterintuitive consequences and decisively, a quale in a dualistic framework cannot contribute to any explanation of behavior. It would be scientifically redundant or superfluous (Perry 2001). Although this is not a knock-down argument, it suggests that a naturalistic view is more attractive. Furthermore, there are good reasons to accept that we are better off without the notion of qualia, conceived of as the contents of mental representations. Primarily, this is because there seem to be no specific qualia-contents.

My proposal is to conceive of the phenomenal aspects of cognitive processes to be tied to different ways of processing. It is not my aim to defend a specific positive view on phenomenal consciousness but only to undermine the claim that phenomenality could be understood as being a special kind of content. This should be sufficient to take seriously a naturalistic view of phenomenality, such that we can start to develop an understanding of phenomenal concepts in a physicalist framework.

The main arguments against the claim that a phenomenal experience is a special kind of content is given by a series of examples showing that one and the same content can be available for a subject unconsciously in one situation and consciously in a different situation. A well-known example is the pathological case of blindsight: People who suffer from blindsight are not able to consciously see objects due to a lesion in the V1 area of the visual cortex. However the semantic information about the kind of object in the blind field is transported by the early visual areas directly to the prefrontal cortex, triggering a correct judgment on the basis of a forced choice test. The information that this is a pencil and not a ball, for example, remains the same. The difference is that, for a blindsighted person, it is only available in forced choice situations without any conscious experience of the object. Another well-known case is the study of the patient D.F. suffering from visual agnosia: D.F. can only guess the orientation of a turnable letter box since conscious experience delivers only crude shades of colors. Nevertheless, if D.F. is given a letter and asked to post it into the box, she can do it perfectly. In other words, despite a lack of conscious information about the orientation, the same information is available unconsciously and this information is useable by the dorsal stream, i.e. for the motor behavior of posting the letter. These two examples have been discussed in greater detail in

Vosgerau, Schlicht, Newen 2008, and together with the additional arguments put forward there, they clearly support the claim that phenomenality should be separated from the dimension of content.

For the general line of argument, it will be very important to distinguish the object-property of being red (a dispositional property of a surface of an object which can be essentially characterized by the reflection of a specific wave length of light) from the subjective red-image (as a special kind of experience a subject has). But to do so within our perspective of antecedent physicalism does only involve different ways of information processing involved in registering the property of being red: either it is a type of information processing that remains unconscious or it is a different type connecting the registration of the property with a conscious experience. Concerning the ontological dimension a phenomenal experience can be analyzed as a special way of information processing that is connected with the registration of the property of being red. On the basis of this ontological background, I can now start to characterize concept formation in general and then to explicate the notion of a phenomenal concept in detail. All different types of concept formation – independent form the level of abstraction – can in principle be integrated into this physicalist framework.

3 Concepts as mental files

Before outlining reasonable constraints about concept possession I grant that there are nonconceptual representations: I accept the standard arguments for nonconceptual representations. 1. The argument from the richness of our experience, i.e. that standard visual experience involves much more than we are able to conceptualize (especially during the short period of having a visual experience); 2. The argument of the fine-grainedness of our experience, i.e. that we are able to distinguish different shades of red when we see them simultaneously near each other, but presented with a time delay we can no longer distinguish them (Raffman 1995). Furthermore, one has to concede that there are cases in which we have phenomenal experiences without phenomenal concepts. At least, that seems to be the best way to describe the phenomenon that animals and new-borns feel pain without having any conceptualization at all, and, thus, *a fortiori*, they suffer from pain without conceptualizing it. In this sense, experience comes first. It is only in a later step

that concepts come into play. In the following, the focus is the level of conceptual representations. According to my view there are two levels of conceptualization: children first acquire perception-based concepts before they can deal with theory-based concepts.

It has been shown elsewhere that this is a fruitful framework (Newen/Bartels 2007): a theory of concepts has a special explanatory value if we define the possession of concepts as an intermediate step between basic sensory patterns, on the one hand, and linguistic representations, on the other. Therefore, without defending a specific theory of concepts in this article, it is presupposed that from the wide range of possible theories of concepts (Laurence & Margolis 1999) we can safely exclude those which identify concepts by a language-like structure (e.g. Peacocke 1992) and those which already speak of concepts on the basis of sensory pattern (e.g. Fodor 1998).[4] For the purpose of this article it is sufficient if the reader grants three features of concepts which I will use in our characterization of concepts as mental files: (i) concepts can be characterized by content features that may include more than basic sensory pattern and they need not (but can) include content features with a linguistic structure, (ii) we can distinguish different types of information which constitute a concept and (iii) the organization of the information registered as content features are usually attached to objects or properties (or more general to an extension of the concept).

In this section a specific way of characterizing concepts in terms of mental files will be discussed. The basic idea of mental files is not new. In philosophy it was, e.g., used by Perry (1990) and in psychology it became prominent with Treisman's theory of temporary object files (Treisman 1998). Mental files have three central dimensions which are adopted from features of concepts in general: 1. mental files have content; 2. we can distinguish three types of information that constitute the input of the mental file, and 3. a mental file normally has an anchoring relation to a real entity (objects, properties, classes of objects, n-tuples etc.) which may be empty (i.e.: the "real world position" of the relation is not necessarily filled). We can construct mental files of several entities. Let us focus on objects and properties. A concept of an object, e.g. the red puppet on Karin's desk, can be represented using typical information, e.g. in the case of a red puppet we have *sensorimotor information (SI)* by grasping the puppet and *image-like information (II)* by seeing the red puppet.[5] This information is part of a mental file

referring to the red puppet, and can already be grasped by nonlinguistic children. With the acquisition of language, *descriptive information (DI)* also comes into play, e.g. *"my red puppet"*, *"my favorite toy"*. The first level of perception-based concepts:[6]

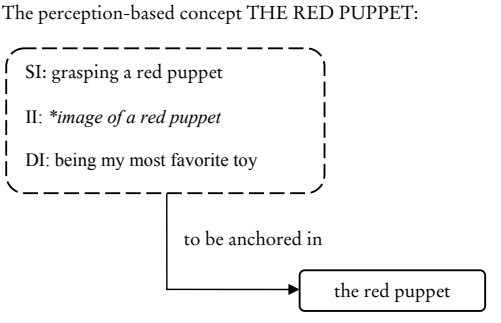


Figure 1

The mental file containing this information can be understood as presenting different concepts depending on the criterion of individuation: If the image of the red puppet (or a sensori-motor information) is the individuating information, we have a *perception-based concept*. But if the description "being my most favorite toy" uttered by Karin is the individuating feature, she developed a theory-based concept because the description presupposes an embedding in a mini-theory about children usually playing with toys, having a favorite toy and girls who often prefer puppets. The individuation criterion is marked by an asterisk and written in italics and now it is the descriptive information.[7]

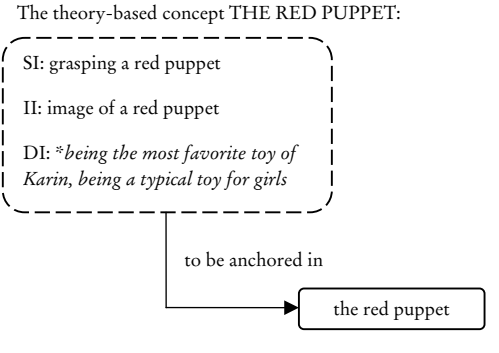


Figure 2

To explain how we acquire phenomenal concepts we have to say something about the development of mental files of properties. Although we can continue to engage with the same state of affairs in the world, we may develop a different type of concept, i.e. this time we may develop a concept referring to the property of being a red puppet:

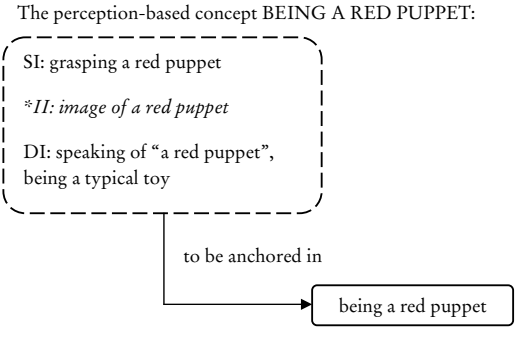


Figure 3

4 Phenomenal concepts as theory-based concepts

Now, I want to apply this idea of mental files to those concepts that can be associated with the word "red", since I want to discuss concepts that focus on phenomenal experience. To do so step by step, I first introduce a concept referring to the object-property of being red:[8]

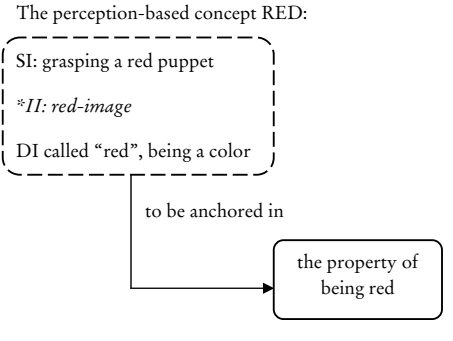


Figure 4

In what follows, I will argue that in the process of acquiring concepts which can be used to classify our experience, two stages can be distinguished: (i) We classify objects according to perception-based properties (constituting a perception-based concept), and (ii) we conceptualize our experiences by descriptive information against the background of some mini-theory (constituting a theory-based concept). The central claim concerning phenomenal concepts is that they are theory-based – they cannot be acquired without having a mini-theory about experiences. Again, this requires us to briefly address an additional issue: What is a theory-based concept? Having answered this question, I will turn to an argument which shows that in fact, phenomenal concepts are best conceived of as being theory-based.

4.1 Perception-based and theory-based concepts

Perception-based concepts rely on characteristic features of the extension which are available on the basis of perception and/or sensory information, while theory-based concepts rely on a description which is treated as a definition. Concept acquisition is a process in ontogeny that starts with classifying entities via characteristic features and develops into the understanding of defining features. Let us make this idea more precise by considering an example: the development of a full-blown concept ISLAND.

As I already introduced, there are three kinds of information one might want to distinguish. In order to characterize a representation of an island we have sensorimotor information which we may think of in terms of affordances (e.g. swimming in the water, wearing a swim suit, enjoying the sun), we have image-like information (e.g. images of palm trees, sandy beaches, sea shells, and sunshine), and finally there is descriptive information represented in natural language (e.g. being a piece of land near water, being the place of our summer holidays, having a beach). If we are language-competent much of this information available can also be referred to by description.

Relying on the different kinds of information we can construct the mental file of the island Gran Canaria:

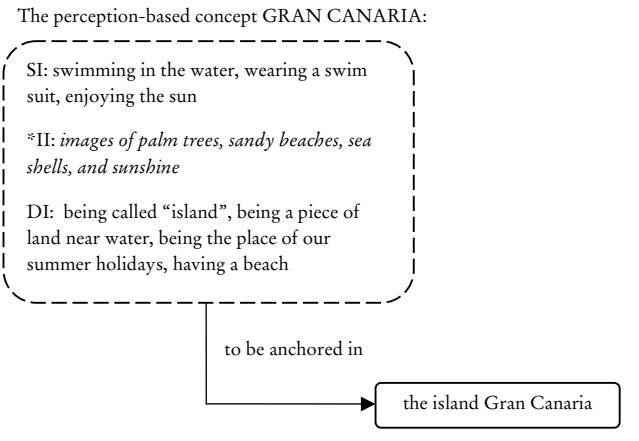


Figure 5

A mental file of an object or a property is normally characterized by the content in the file and by an anchoring relation to the entity in the world. It remains a perception-based concept as long as the individuation information is sensorimotor and/or image-like information. If descriptive information is taken to be the individuating element then we develop a theory-based concept. That is exactly what happens when kids learn to understand the concept of an island by a definition: x is an island $\leftrightarrow_{\text{def.}}$ x is body of land surrounded by water. In developmental psychology there is a well-documented shift from understanding a concept relying only on characteristic features to the understanding of the definition (Keil, 1989). In the following a typical dialogue is presented with a three (or four) year old child that has a concept of ISLAND that is relying on some characteristic features, e.g. people are not wearing clothes, being summertime, being near water (Keil 1989, p. 62–63):

"Experimenter: Who lives on an island?
Child: People … yup, people without clothes on … […]
E: Is there an island in Ithaca?
C: No.
E: Why not?
C: 'Cause it's not summertime yet.
E: You mean there can only be islands during the summer?
C: Yeah. […]
E: Is an island near water?
C: Yeah.

E: But Ithaca is near water. What's the difference?
C: You know what? We went to Danny's house and swimmed there."

This understanding of the concept of ISLAND is far away from the correct understanding by definition since the characteristic experiential feature of being summertime is treated as a necessary component. However in the first stage there is not yet one single individuating feature, but rather a bundle of features that is relevant for the extension. In the ontogenetic process of understanding the definition of "island", first the relevant feature "being a body of land surrounded by water" is represented as one of the relevant necessary components before it is treated as a definition. On this basis the picture is simplified in the following discussion, presupposing that we always have an individuating feature (or bundle of features). We now focus on the shift from image-like (or sensorimotor) features to descriptive features as individuating elements of the mental file. The individuating information is marked with an asterisk and italics. After learning the definition the kid has a modified singular concept of the island Gran Canaria.

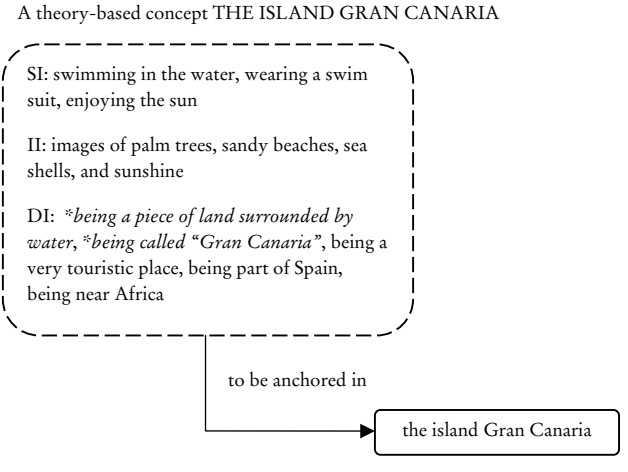


Figure 6

The sensorimotor as well as the image-like information can remain part of the concept but need not be. We sometimes learn a new concept only by description. If a person acquires the concept THE ISLAND GRAN CANARIA only by learning the definition of "island" and understand-

ing that there is an island called "Gran Canaria" then the mental file may have the following structure:[9]

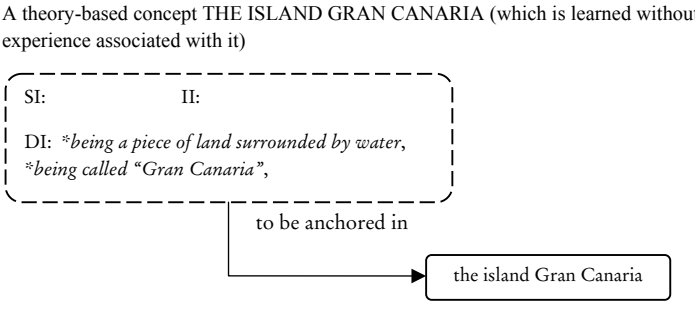


Figure 7

What I'm characterizing here are individual concepts in the sense defined by Carnap, i.e. the reference of the concept is an individual entity. According to the ontological framework the focus is on objects and properties as referents.[10] The anchoring relation is a relation to the object as an entity in the world and this relation is usually a causal relation, i.e. I can experience the object or I can be connected by a causal chain of communication as described by Kripke. Including a referent into the box means that the person is causally connected with the object, e.g. via perception or a causal chain of communication. If there is no causal connection established, the box of the referent is empty. This is of course the case if we have concepts of fictional persons like Sherlock Holmes.[11] For the sake of argument we are presupposing that we deal with real properties when we discuss the problem of phenomenal concepts. In the real world the property of being red exists and it is instantiated by fire engines, upper traffic lights, tomatoes etc. More in particular, being red is a property of these objects: we know that reflecting a specific wave length of light is a characteristic feature of objects instantiating this property. Another typical feature is that it normally produces a red-experience in humans. So we presuppose that there is the property of having a red-experience which is instantiated as my subjective experience when I see a tomato. In the following I want to apply this framework of mental files to explicate the development of phenomenal concepts. Before doing so, let me mention some advantages of this framework: Mental files can easily account for the phenomenon of

learning that the morning star is identical with the evening star (Perry 1990, 2001b). The Frege phenomenon that was discussed in the context of informative identity statements starts with two different mental files attached to two different singular terms:

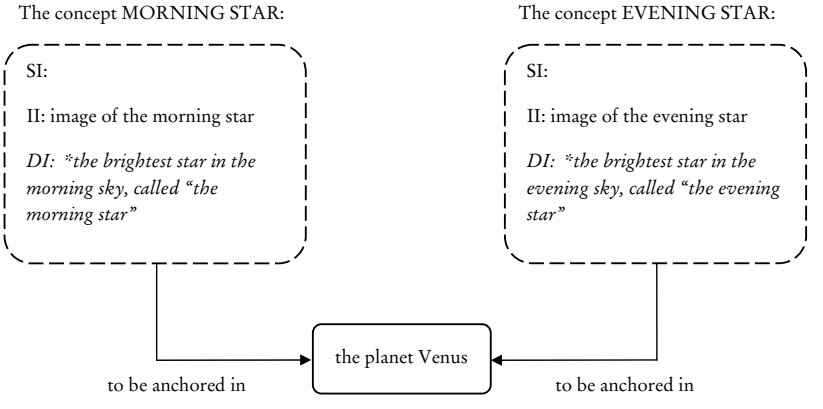


Figure 8

Both mental files actually have the same referent but the thinker does not know this. When she becomes aware of the co-reference of two singular terms, she unifies the content of both files into one file:

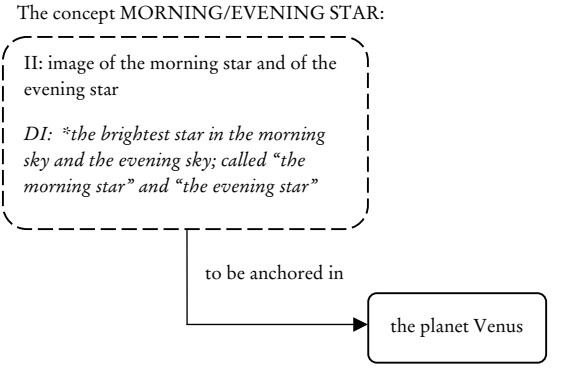


Figure 9

Furthermore, we can account for the important psychological evidence concerning the development of concepts: children start to acquire con-

cepts on the basis of registering affordances and image-like information. Then descriptive information becomes more and more important throughout development. We sketched the shift from a feature-based to a definitional understanding of concepts. All this can be nicely modeled in this framework. Importantly, we are able to model the shift of the criterion of individuation, e.g. from an image-like information of islands (images of beaches, palm trees, etc.) to a definitional description "a large piece of land surrounded by water". Although the mental file may still contain the same information, its individuating criterion has changed. This allows us to explain why we can easily understand utterances which seem to be contradictory, e.g. "This sausage <pointing at a vegetarian sausage> is not a sausage".[12] The same word is used to express different concepts. The underlying structure of this sentence is the following: It is said that the objects falling under a *perception-based concept* are not part of the extension of the *theory-based concept*.

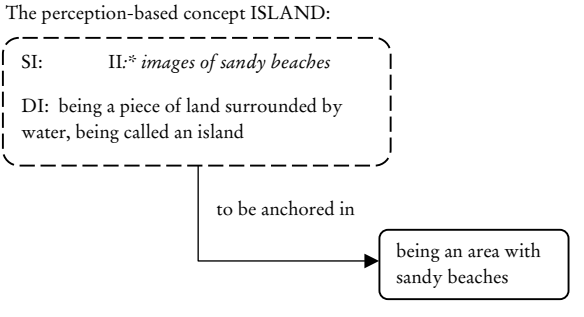


Figure 10

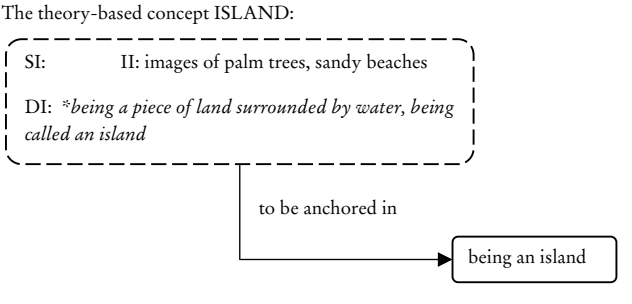


Figure 11

One can easily transfer this to the case of a perception-based concept of SAUSAGE with an *image of a sausage* as the criterion of individuation while the theory-based concept is individuated by the description "being made of finely minced meat, esp. pork or beef, mixed with fat, cereal or bread, and seasonings usually stuffed into a prepared animal intestine". This not only demonstrates another fruitful application of the model, but also shows that perception-based concepts can have different referents vis-à-vis theory-based concepts despite the fact that we use the same word in our language. Now all the necessary aspects of the mental file-framework that are needed to apply it fruitfully to the case of phenomenal experiences and phenomenal concepts I propose, are introduced.

4.2 Phenomenality and conceptualization

Let me first deal with the conceptualizations of pain and then explicate how I propose to conceptualize color-experiences. When children learn to conceptualize pain-experience they first develop a perception-based concept of PAIN. This is a very general concept of PAIN which includes a lot of characteristic features, and the characteristic feeling of pain is here only one element among many others. This view is closely connected with Wittgenstein's view of phenomenal concepts as public concepts which can be explicated by patterns of characteristic features. According to such a Wittgensteinian view, the concept PAIN includes 1. typical behavioral patterns (including speech behavior), 2. typical physiological symptoms and/or bodily properties, and 3. typical phenomenal experiences. If someone is suffering from tooth-ache the perception-based concept includes not only the characteristic pain-experience but also the typical facial pain-expression, the holding of one's cheek etc. Typical behavioral features are essentially connected with the pain-experience to form a pattern constitutive for classifying the phenomenon as falling under the concept PAIN. The perception-based concept of PAIN normally takes the observable features to be the individuating criteria:

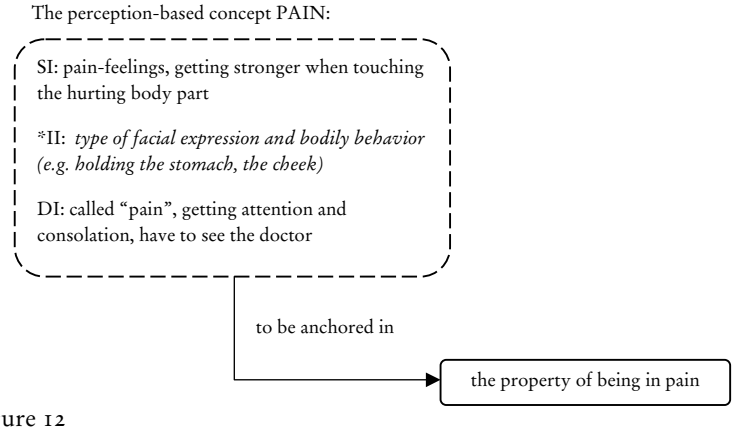


Figure 12

In ontogeny we acquire the cognitive ability to form concepts that are defined only by one feature (even if they include other features). To do so children have to learn to shift from feature-based conceptualization to conceptualization on the basis of defining features, as was illustrated above. Then they are able to learn concepts of PAIN-BEHAVIOR, PAIN-PHYSIOLOGY and of PAIN-EXPERIENCE. We have to assume that these concepts are theory-based concepts because in our everyday life these aspects normally appear together. Thus, they have to be cognitively separated with the help of a theoretical context. To develop a concept of PAIN-EXPERIENCE children have to learn a mini-theory of the mind. In our culture this is at least the folk-psychological mini-theory according to which persons have a subjective and private experience. To develop this minimal picture of the mind children must be able (i) to distinguish appearance and reality, (ii) to explicitly distinguish properties, objects, on the one hand, and subjective experiences on the other etc. and (iii) to think of persons as having a mental world that is different from one's own (Theory-of-Mind ability) and that this mental world is private and (iv) to evaluate the behavior of others on the basis of recognizing mental states. This allows us to establish a standard folk psychological theory of mind.

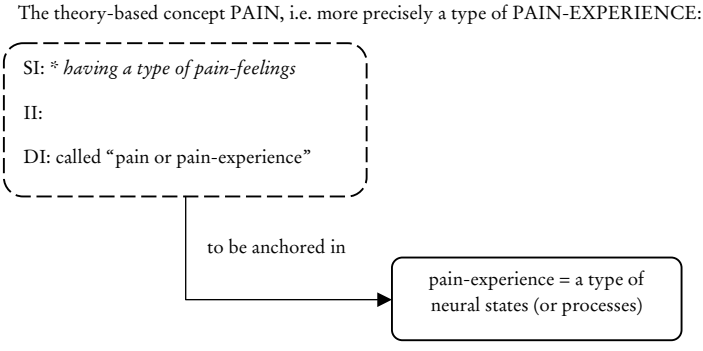


Figure 13

In our everyday language we mostly understand the word "pain" with the perception-based concept i.e. thinking of a cluster of characteristic features: if we observe how someone shuts the fingers of a child in a door, we immediately know that child is in pain. We see the red fingers that start to get blue even though under shock the child did not cry but only turned white. We immediately start to do something that will stop the pain and heal the fingers. The theory-based concept of pain is mostly used in special contexts, e.g. when I have a permanent headache and the doctor asks me to describe the type of headache in detail or when we are in a philosophy seminar.

The main point here is that we can distinguish between perception-based concepts and theory-based concepts although they are related to one and the same world. The ontology does not change when we learn to describe our pain-experiences as subjective experiences with a special phenomenal quality. It is only our conceptualization, the way we conceptualize the world that develops. This view can also be applied to color-experiences.

4.3 Color-experiences and our conceptualizations

When I see a tomato, my experience can be analyzed by distinguishing two properties: the tomato has the property of *being red* (a property of the object) that causes the property of *having a red-experience*.[13] When we start to deal with colors we acquire a perception-based concept that refers to the object-property of being red. Furthermore, I acquire a concept that refers to my red-experience. This requires a mini-theory about the subjectivity and privacy of my experiences such that I am developing a theory-based concept. Let me illustrate this point. The perception-

based concept of being red can be acquired by seeing ripe tomatoes, cherries, Ferraris and so forth. The related image-based information is the individuating element in the context of the implicit knowledge.

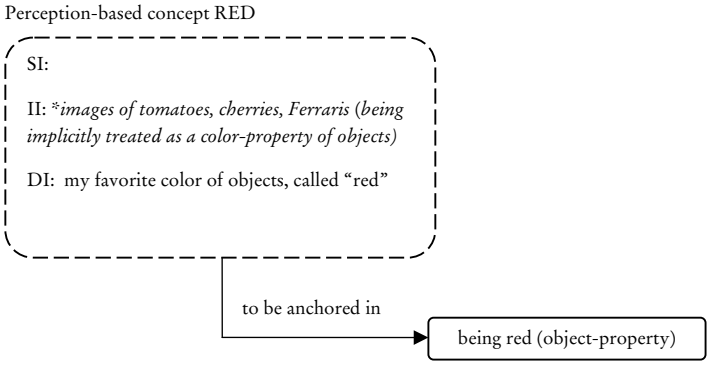


Figure 14

In addition to the perception-based concept of RED we are able to develop a concept of RED-EXPERIENCE (by changing the reference). This is referring to the property of having a red-experience (= a red image). We conceptualize it as a subjective and private color-image of my experience when seeing tomatoes (although it is at the same time identical with a neural correlate according to antecedent physicalism).

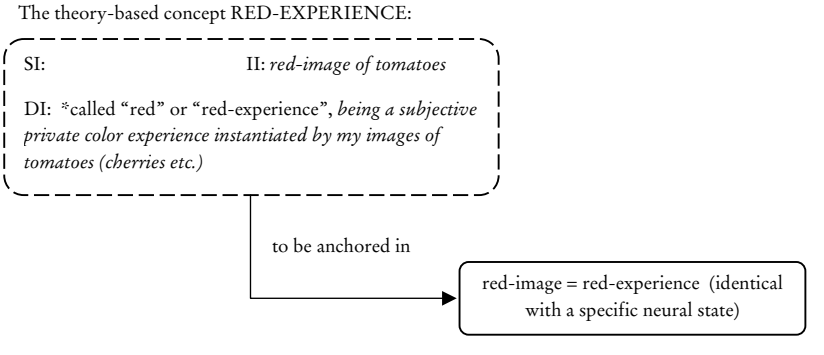


Figure 15

This concept is individuated by the description "*being a subjective private color experience instantiated by my images of tomatoes*" while this description refers to the red-image of the person which thereby is part

of the individuation criterion (albeit indirectly). Here the red-image is part of the individuation criterion on the background of this explicit mini-theory about subjective and private experiences. This is exactly the way philosophers use the word "red". In our everyday speech we are often using loose talk. Then we use the word "red" to speak about both, the perception-based concept RED and the theory-based concept of RED-EXPERIENCE. But it is important to be aware what one is talking about and this helps us to adequately analyze the knowledge argument.

5 Phenomenal Concepts and the Knowledge Argument

In the introduction, the phenomenal-concept strategy was already outlined and it shall be defended now (for independent arguments in favor of the strategy to account for mental properties by analyzing the relevant concepts, see van Riel, in this volume). I shall now apply the interpretation of phenomenal concepts as being based on (mini-) theories to three aspects of the knowledge argument which play a crucial role in the context of reduction: the case of Mary, the dualist intuition and a modified case of Marianna.

5.1 The case of Mary and the dualist intuition

Let me recall the main point: When Mary is released from her black and white room, she receives new information: she learns what it is like to see red by having a red-experience. But an antecedent physicalist would simply insist that the red-experience is identical with a physical fact, probably a neural correlate of the experience. Then the basic analysis runs as follows: Mary as a brilliant neuroscientist had knowledge about the neural correlate of a red-experience already before her release. When she first had a red-experience she learned a different epistemic access to the same ontological entity she already had knowledge about. Given our analysis of concepts, it can be shown that Mary develops her concepts by including essential information but this does not include any change in the ontology.

Already in the black and white room Mary is in the position to form a perception-based concept RED which refers to the property of being red (see the relevant mental file above). Although this concept is lacking a central information since she has never seen a red object. Since

she understands that the other humans have color experiences, she can also develop theory-based concepts of RED-EXPERIENCE. Since she observes that philosophers are fighting about the ontology she knows that there are different ontological theories concerning the same physical phenomenon of having a red-experience. Mary is able to build two theory-based concepts, i.e. a *phenomenal concept* and a *neural concept* of the red-experience. Both have different criteria of individuation: while the phenomenal concept takes the red-experience to be essential, the neural concept takes the neural correlate as the concept's individuating feature. At the beginning both mental files lack the information of having a red image. But when Mary acquires this information she does not learn a new concept but only supplements the concepts which she already was able to construct. Here are more details about this view:

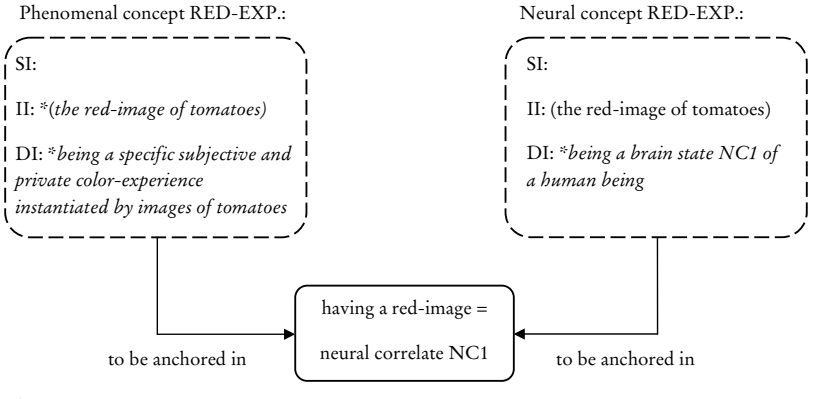


Figure 16

At the beginning both theory-based concepts of the red-experience lack the information of a red-image in the file. How can we analyze the acquisition of the new information by Mary? She fills in the new information, the red image, in both concepts. The theory-based phenomenal concept of RED-EXPERIENCE then involves it as an essential part because the red-experience is the individuating criterion while the neural concept can include the additional information as an inessential part of the concept. For an antecedent physicalist who knows that the red-experience is identical with a neural correlate it would be natural to start to merge the files. This is analogous to the merging of two files into one in typical Frege cases (if you learn a new informative identity). So as an

antecedent physicalist Mary would end up with one file combining all the information and having one anchoring relation. Since the information is known to be of one and the same entity, Mary can have a plurality of criteria of individuation.

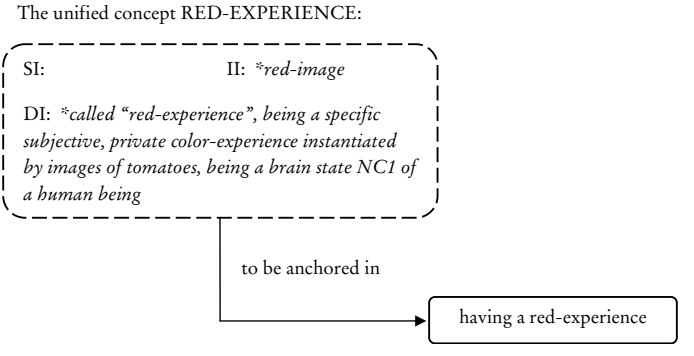


Figure 17

This final picture is only available for the antecedent physicalist. This illustrates again that this is a theory-based concept. Furthermore, I can nicely illustrate the dualist intuition. Since our folk psychological theory of phenomenality and recent neurophysiological data are still not very well interconnected up to now, we have a strong and rational tendency to develop two theory-based concepts, a phenomenal and a neural one. The dualist intuition is simply the intuition that since there are these two concepts with different contents there must be two different underlying entities referred to. A physicalist denies this conclusion. An antecedent physicalist simply presupposes from the beginning that there is only one referent but given our epistemic situation one may start with two theory-based concepts since one does not know the relevant informative identity concerning an experience and the underlying neural correlate but de facto we have two concepts of the same entity right from the start. Mary's cognitive situation has to be described using a theory-based phenomenal concept and a theory-based neural concept of RED-EXPERIENCE. But the thought experiment tells us nothing that allows us to decide whether the anchoring relation involves only one or two entities. This remains an open scientific question and not a philosophical one. An antecendent physicalist can account for all aspects of Mary's cognitive situation in the naturalist framework modeling her knowledge with mental files.

5.2 The case of Marianna

In the literature the knowledge argument was developed further with the so-called Marianna case (Nida-Rümelin 1995) which is a modification of the Mary story. I can also account for this thought experiment within the framework. Marianna has the same starting conditions as Mary but when she is released she first has a color experience and does not know how to match colors and color-names and which standard objects cause her new experience. So she opens a mental file that is individuated by the specific color experience lacking any description of the typical cause. This is the concept THIS COLOR EXPERIENCE; and she already has a concept of RED-EXPERIENCE on the basis of theoretical information:

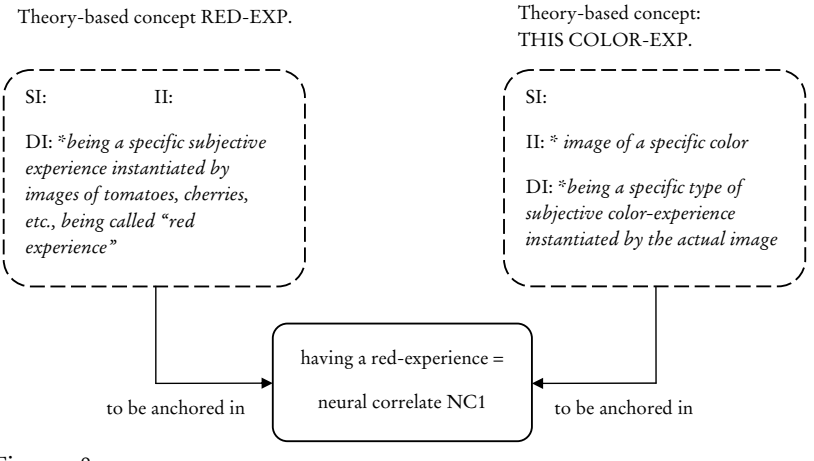


Figure 18

When she is informed that her experience is caused by looking at a tomato and called "red-experience", she is able to merge both files into one.

Concluding remarks

The theory of phenomenal concepts is based on two constraints: first, I am presupposing antecendent physicalim and second, we have to distinguish the property of an object of being red and the property of a subject of having a red-experience. This leads to a distinction of different

concepts related to the everyday expression "red". To prevent misunderstandings I speak of the acquisition of two kinds of concepts, i. e. the concept RED referring to the property of being red and the concept RED-EXPERIENCE referring to the property of having a red-experience. There are two understandings of *phenomenal concepts*: in a loose sense RED is a phenomenal concept because the mental file involves the red-image as a characteristic feature of the concept. But since the red-image is not the referent it is not a phenomenal concept in a strict sense. This is true for the concept RED-EXPERIENCE which is a theory-based concept while RED remains a perception-based concept. A theory-based concept of RED-EXPERIENCE involves not only the red-image as part of the mental file but it also refers to it. All philosophical discussions about color-experiences rely on a theory-based concept of RED-EXPERIENCE. To account for the case of the brilliant neuroscientist Mary we need to distinguish two theory-based concepts, the phenomenal concept of RED-EXP. referring to the red-image as individuated by the image itself and a neural concept of RED-EXP. referring to the red-image as individuated by a specific neural state. This allows us to analyze the knowledge argument: Mary acquires a red-image when she is released into the colored world. This red-image is not only integrated into her perception-based concept RED but also in both theory-based concepts. Now, in the neural concept of RED-EXP. the red image can be integrated as a characteristic feature but not as the defining component. But it is integrated into the phenomenal concept RED-EXPERIENCE as an essential part of the individuating criterion. Since this can be described consistently and fruitfully within a physicalist framework, a naturalistic view of phenomenal concepts is offered which shows that the knowledge argument does not imply dualism. Phenomenal concepts are theory-based concepts that are developed to classify our phenomenal properties, e. g. red images. This is an advanced cognitive ability which indicates a complex cognitive organization of the concepts without any implications for the underlying ontology of experiences.

Notes

1 I would like to express my special gratitude to Raphael van Riel who supported the completion of this article with very detailed critical comments and helpful constructive suggestions. Furthermore, I would like to thank

the members of my research group for further critical comments, especially Leon de Bruin.

2 While accepting phenomenality I try to avoid the ontologically loaden speech of *qualia* and prefer to use the terms "phenomenality" or "phenomenal experience" as neutral terms to speak of the character of our everyday experiences of objects, events, and processes.

3 I'm open-minded to a discussion of what exactly the basic building blocks in an ontology are but at one level it is unproblematic to presuppose objects and properties and this is all I need for my purposes. If all objects and properties are integrated in absolute basic ontology, e.g. reduced to processes, this does not touch the argument. It only shows that I am dealing with an intermediary ontology that is further reducible.
+ explication would only be an intermediate stage to all full explication of phenomenal concepts in a physicalist story. So we aim at an explication at the level of presupposing objects and properties which we take to be a very plausible starting point for independent reasons.

4 This is developed and defended in detail arguing for an epistemic theory of concepts including the constraint that concepts are involving some systematic organizing of the representations (e.g. involving an object-property-structure allowing a flexible use in new situations and involving a minimal holistic interconnection) (Newen, Bartels 2007).

5 I am not dealing with the question when exactly the organization of this information is such that we actually have a concept and not just a basic sensory generalization. For the sake of argument I simply presuppose that we are now dealing only with concepts, leaving aside nonconceptual representations.

6 The anchoring relation to a property can be understood as an anchoring relation to an instantiation of a property. The causal relations can of course be of different kinds, e.g. based on perception, on a causal chain of communication, etc.

7 With this tool the difference between referentially and attributively used singular terms can also be illustrated. In the latter case it is descriptive information that is individuating the concept while in the former an anchoring relation to one special entity is the criterion of individuation. The anchoring relation can be understood as a causal relation to an object in a wide sense including Kripke's causal chain of communication as one way of being causally related.

8 There is an ongoing debate concerning the question how closely related the acquisition of color concepts and color terms are. For the purpose of this article this question is not a central aspect. So I can grant that during ontogeny of understanding colors the acquisition of concepts and terms is taking place in parallel (Pitchford 2006).

9 This should only illustrate the extreme case of having no other associations which rarely is realized.

10 We can easily widen this picture in the direction of general concepts. Then the referents may be understood as classes of n-tuples. To develop the core idea of phenomenal concepts we need not enfold this dimension of mental

files here. We simplify the view by understanding universals as abstract objects, i.e. the universal red is understood as a type of property that you also can characterize by the class of red objects. Important is only that we have an understanding of objects and properties which takes them to be part of the real world.

11 We are not going to deal with the special problems of empty names and concepts of fictional characters in this paper.

12 Further examples are: "This bottle of beer <pointing at an object made of chocolate> is not a bottle of beer." "This gun <pointing to a fake gun> is not a gun."

13 It is commonplace that we have to distinguish the physical property of an object to reflect a special type of light with wave length x and the mental property of having a color experience of type E. Color experiences are constructed by the brain accounting for a lot of contextual information in addition to the wave length.

Bibliography

Alter, T., Walter, S., 2007: *Phenomenal Concepts and Phenomenal Knowledge – New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Armstrong, David, 1968: *A Materialist Theory of Mind*. London: Routledge.

Block, N., Stalnaker, R., 1999: Conceptual Analysis, Dualism, and the Explanatory Gap. In: *Philosophical Review* 108, pp. 1–46.

Carruthers, P., 2000: *Phenomenal Consciousness*. Cambridge, UK: Cambridge University Press.

Chalmers, D., 1996: *The Conscious Mind*. Oxford, UK: Oxford University Press.

Fodor, J. A., 1998: *Concepts: Where Cognitive Science went wrong*. New York: Oxford University Press.

Frege, Gottlob, 1966: *Logische Untersuchungen*. Göttingen: Vandenhoeck & Ruprecht.

Jackson, F., 1982: Epiphenomenal Qualia . In: *Philosophical Quarterly* 32, pp. 127–136.

Jackson, F., 1986: What Mary Didn't Know. In: *Journal of Philosophy* 83, pp. 291–295.

Jung, E.-M., Newen, A., 2010: Knowledge and Abilities: The need for a new understanding of knowing-how. In: *Phenomenology and Cognitive Sciences* 9, 1, pp. 113–131.

Jung, E.-H., Newen, A., 2011: Understanding Knowledge in a New Framework, in: Newen, A., Bartels, A., Jung, E.-H. (eds.): Knowledge and Representation, Stanford, CSLI publicatiohns, pp. 80–105.

Keil, F. C., 1989: *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kripke, S., 1972: Naming and Necessity. In: Davidson, D., Harman, G. (eds.): *Semantics of Natural Language*. Dordrecht: D. Reidel, pp. 253–355, pp. 763–769.

Künne, W., 2007: Fiktion ohne fiktive Gegenstände. Prolegomenon zu einer Fregeanischen Theorie der Fiktion. In: Reicher, M. E. (ed.): *Fiktion, Wahrheit, Wirklichkeit. Philosophische Grundlagen der Literaturtheorie*. Paderborn: mentis, pp. 54–72.

Levine, J., 1998: Conceivability and the Metaphysics of Mind. *Noûs* 32, pp. 449–480.

Locke, John, 1965: *An Essay Concerning Human Understanding*. Edited by Yolton, John W., London: Dent.

Margolis, E., Laurence, S. (eds.), 1999: *Concepts – Core Readings*. Cambridge, MS: MIT Press.

Newen, A., Bartels, A., 2007: Animal minds and the possession of concepts, *Philosophical Psychology*, 20, 283–308.

Nida-Rümelin, M., 1995: What Mary Couldn't Know: Belief About Phenomenal States. In: Metzinger, T. (ed.): *Conscious Experience*. Paderborn, München, Wien, Zürich: Ferdinand Schöningh.

Papineau, D., 1993a: *Philosophical Naturalism*. Oxford, UK: Blackwell.

Papineau, D., 1993b: Physicalism, Consciousness, and the Antipathetic Fallacy. *Australasian Journal of Philosophy* 71, pp. 169–183.

Papineau, D., 2002: *Thinking about Consciousness*. Oxford, UK: Oxford University Press.

Papineau, D., 2007: Phenomenal and Perceptual Concepts. In: Alter, T., Walter, S. (eds.): *Phenomenal Concepts and Phenomenal Knowledge – New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Peacocke, C., 1992: *A Study of Concepts*. Cambridge, MA: MIT Press.

Perry, J., 1990: Self-Notions. *Logos* 11, pp. 17–31.

Perry, J., 2001a: *Knowledge, Possibility and Consciousness*. Cambridge, MA: Bradford-MIT.

Perry, J., 2001b: *Reference and Reflexivity*. Stanford: CSLI Publications.

Pitchford NJ., 2006: *Reflections on how color term acquisition is constrained*. J Exp Child Psychol. 2006 Aug; 94(4):328–33.

Raffman, D., 1995: On the Persistence of Phenomenology. In: Metzinger, T. (ed.): *Conscious Experience*. Paderborn, München, Wien, Zürich: Ferdinand Schöningh, pp. 293–308.

Schlick, Moritz, 1979: *Allgemeine Erkenntnislehre*. (Originally published in 1925) Frankfurt a. M.: Suhrkamp.

Treisman, A., 1998: Feature binding, attention and object perception. *Phil.Trans. R. Soc. Lond.* B 353, pp. 1295–1306.

Tye, M., 1999: Phenomenal Consciousness: the Explanatory Gap as Cognitive Illusion. *Mind* 108, pp. 705–725.

Tye, M., 2003: A Theory of Phenomenal Concepts. In: O'Hear, A. (ed.): *Minds and Persons*. Cambridge: Cambridge University Press, pp. 91–105.

Vosgerau, G., Schlicht, T., Newen, A., 2008: Orthogonality of Phenomenality and Content. *American Philosophical Quarterly* 45, pp. 309–328.

Raphael van Riel

Identity-Based Reduction and Reductive Explanation[1]

Abstract

In this paper, the relation between identity-based reduction and one specific sort of reductive explanation is considered. The notion of identity-based reduction is spelled out and its role in the reduction debate is sketched. An argument offered by Jaegwon Kim, which is supposed to show that identity-based reduction and reductive explanation are incompatible, is critically examined. Based on the discussion of this argument, some important consequences about the notion of reduction are pointed out.

Zusammenfassung

In Zentrum dieses Aufsatzes steht die Relation zwischen identitätsbasierter Reduktion und einer spezifischen Sorte reduktiver Erklärung. Der Begriff identitätsbasierter Reduktion wird ausbuchstabiert und seine Rolle in der Reduktionsdebatte wird nachgezeichnet. Ein Argument Jaegwon Kims, nach dem identitätsbasierte Reduktion und reduktive Erklärung inkompatibel miteinander sein sollen, wird einer Kritik unterzogen. Die Diskussion dieses Arguments wirft Licht auf einige interessante Aspekte des Reduktionsbegriffs.

Introduction

Ever since Ernest Nagel cashed out his model of reduction in terms of explanation (Nagel, 1961, section 11-II), it was common among philosophers to regard reduction as being closely tied to explanation. However, there is no agreement about what sort of explanation reduction amounts to. Schaffner (1993, 429), Sklar (1967) and Patricia Churchland (1986, 283) describe theory reduction as explanation of why a theory worked as well as it did; Friedman (1982) described reduction as being tied to explanation of the reduced theory's *phenomena* by the reducing theory. Obviously, Nagel's remarks on reduction as an instance

philosophia naturalis 47-48 / 2010-11 / 1-2

of a covering law model of explanation point into the direction that he regarded reduction to be explanation of the reduced theory by the reducing theory (this interpretation can be found in Schaffner (1967) – a view which is repeated by Patricia Churchland (1986, 283)). Moreover, philosophers suggested causal (Enç, 1976), unificatory (Churchland, 1986, 279f) and functional and mechanistic (Fodor, 1974, 107; Kim, 2008) interpretations of reductive explanation. Although philosophers suggested these alternatives, the question of how reduction relates to which sort of explanation has not been addressed in a systematic way yet. One might get the impression that within the reduction debate, the tensions between the different interpretations of reductive explanation just escaped the discussants' attention.

In addition, it is hard to uncover what the exact relation between reduction and reductive explanation is supposed to be. Nagel's model suggests that the reduction relation just is an explanation relation. A more modest interpretation would suggest that if a reduction relation holds between theories or models, then there is a corresponding explanation relation, such that reduction is sufficient for reductive explanation, but that they should not be identified (this interpretation can be found in Schaffner's General Reduction Replacement Model (Schaffner, 1993)). That reductive explanation is, in turn, sufficient for identity-based reduction, as spelled out in a moment, has been denied by a number of philosophers.

Craver (2007), for example, argued that mechanistic explanation is at least not sufficient for reduction (as based on identities). A similar idea seems to be pertinent in Chalmers (1996, 43) and Fodor (1974, 107), who assume that reductive explanation is compatible with the denial of reductionism. In a far more radical spirit, Jaegwon Kim (2008) argued that identity-based reduction is *incompatible* with a specific interpretation of reductive explanation (conceived of as mechanistic explanation). Kim thus stands in opposition to the widespread agreement that reduction and reductive explanation stand and fall together. In the present paper, Kim's argument will be critically examined, and a general conclusion about specific notions of reduction and reductive explanation will be drawn.

Reductive explanation is, according to Kim's reading, intuitively captured by formulations of the following sort: An entity A is reductively explained by an entity B iff A is explained in terms of B, or iff the occur-

rence of A's is explained in terms of the occurrence of B's, or iff B is the mechanism which explains how A functions. Here is an intuitive and utterly preliminary formulation of Kim's argument: Given that the relata of the reduction relation form the explanans and the explanandum of a corresponding reductive explanation, then identity-based reduction is incompatible with reductive explanation. This is so because according to identity-based reduction, there just are no two distinct relata of the reduction relation. Since, as Kim argues, nothing can be explained referring to itself only – which is his interpretation of the assumption that explanations must not be *trivial*, or take the form of 'x explains x' – identity-based reduction is incompatible with reductive explanation. If x is shown to be identical to y, then x cannot explain y. A discussion of this argument forms the core of the present paper. Showing in which respect it fails, we can derive an interesting conclusion about the notion of identity-based reduction.

Having explicated the idea of identity-based reduction and having distinguished between several kinds of explanation which have been associated with reduction in the debate (sections 1 & 2), I will present Kim's argument (section 3) and argue that it is fallacious: Identity-based reduction is perfectly compatible with the relevant kind of reductive explanation (section 4). This intermediate conclusion will form the basis for an elaboration on the notion of identity-based reduction (section 5). It can be shown that this notion has to be reconsidered: A sentence of the form 'A reduces to B' expresses a truth only if (i) the referent of the instance of A is identical to the referent of the instance of B, and (ii) the instances of 'A' and 'B' express different conceptual contents, or have different meanings. A similar conclusion can be drawn for theory-reductions. Thus, the paper aims at the clarification of one specific interpretation of reductive explanation and its relation to identity-based reduction. Moreover, it draws conclusions about the nature of identity-based reduction gained from the insights of the previous discussion.

1 Identity-Based Reduction

The term 'identity-based reduction' is supposed to cover several reduction-relations which share a common feature: They partly rely on the notion of identity.[2] The aim of this first section is to introduce notions

of identity-based reduction and to give a short idea of the fundamental role they play in the reduction debate by giving a list of accounts which rely (at least partly) on a notion of identity-based reduction.

1.1 Two notions of identity-based reduction

The core idea of identity-based reduction is easy to grasp. Identity-based reduction comes, roughly, in two forms. Either the reduction predicate is flanked by terms which refer to entities which have to be identical in order to reduce. This is the case for kind- or phenomena-reduction. In this sense of 'reduction' a kind or phenomenon A reduces to a kind or phenomenon B only if A=B. Under an alternative reading, the reduction predicate is flanked by expressions which refer to representational items which are said to reduce, like a set of laws (conceived of as a constituent of a theory), or an entire theory, or a scientific model, or a description. In such cases, the theory should better not be identical to the theory it reduces to. However, the reduction of a theory might hold in virtue of identities among kinds or phenomena the theories deal with. One way of spelling out this idea is this: A theory T_R reduces to a theory T_B only if for any (relevant) referential term of T_R, t, there is a term t* in T_B such that the referent of t=the referent of t*. Put in more familiar terms for theory-reduction: A model of theory reduction in the spirit of Nagel would be an *identity-based* model of reduction if bridge-principles were to be characterized as stating identities. We thus get two variants of identity-based reduction, one of which could be labeled *ontological* and one of which could be labeled *representational*:

Identity-based reduction (ontological)
'A reduces to B' expresses a truth only if 'A = B' expresses a truth.

Identity-based reduction (representational)
A reduces to B only if for any kind x of A, there is a kind y of B such that x = y.

The substitution class for 'A' and 'B' in the characterization of the ontological variant contains only expressions which stand for kinds, phenomena, or properties, roughly: for non-representational, worldly entities science talks about. In the characterization of the representational version, instances of 'A' and 'B' stand for representational items like theories or scientific models. I chose a schematic way of presenting

these claims for reasons that will become apparent in this paper's last section. Roughly, it will be argued that reduction statements of the form 'A reduces to B' express a truth only if certain complex semantic constraints on instances of 'A' and 'B' in such statements hold true.

Note that if these conditions were sufficient (that is, if the conditionals could be replaced by corresponding bi-conditionals), then we would get a problem for both conditions. The ontological criterion would imply that if A reduces to B, then B reduces to A. Reduction would turn out to be a symmetric relation. The representational criterion would imply that at least possibly, there are two representational items, A and B, which are such that A reduces to B and B reduces to A. Accordingly, reduction would turn out to be a non-asymmetrical relation. We cannot go into the details here, but it should be obvious that this consequence should be avoided if it could be avoided. The discussion of Kim's argument will point into a direction of how to set the stage for a treatment of identity-based reduction as an asymmetric relation.

Now, both these kinds of reduction do play a crucial role in the reduction-debate. Here is a summary of what I believe to be the most important occurrences of these kinds of reduction. I proceed in three steps. First, I give an overview on models of reduction which are normally regarded as being developed in the Nagelian spirit (Ernest Nagel's and Kenneth Schaffner's models). Then I will turn to models of identity-based reduction implicit and explicit in the philosophy of mind, which are commonly described as *identity-theories*. In a final step, I shall consider models of reduction which can be found in the writings of New-Wave reductionists.[3]

1.2 Nagelian Versions of Reduction and Identity

Let us consider Nagel's remarks on reduction and identity first and then turn to a model of reduction which was, among Nagelian models of reduction, the most influential one: Schaffner's *General Reduction Replacement Model*.

Nagel

Nagel's model of theory reduction can roughly be summarized as follows: A theory or set of laws T_R reduces to another theory or set of laws T_B if and only if T_R is deducible from T_B, sometimes with the help of bridge-laws (Nagel, 1961, 352) and boundary conditions (Nagel, 1961,

434). A special case of reduction in the Nagelian model is reduction which does not require bridge-laws because the reducing and the reduced theories employ the same vocabulary. These are labeled 'homogeneous reductions'. Homogeneous reductions will not gain special attention here. It should be obvious that if sameness of vocabulary requires sameness of semantics, then these cases are cases of identity-based reduction. The other cases gain considerable attention in Nagel's discussion of reduction. Bridge-laws state the relevant connection between the distinct theories' kinds (and, thus, between the distinct theories' vocabularies). Some philosophers assume that bridge-laws in Nagelian models, or models which use a covering law strategy to explicate reduction and are, in this sense, developed in a Nagelian spirit, should be interpreted in terms of identities (Causey, 1972 & 1977; Schaffner, 1993, section 9.4.2), whereas others focus on characterizations other than in terms of identities (see, for example, (Richardson, 1979)). In addition, Nagel himself explicitly states that in some specific sort of reductions, bridge-laws can be interpreted as stating identities (Nagel, 1961, 340; Nagel 1970, see also: van Riel, forthcoming).[4]

The idea that bridge-laws state identities is explicitly defended by Kenneth Schaffner.

Schaffner

The *General Reduction Replacement Model* (henceforth: GRR model) proposed by Schaffner in his (1993, chapter 9) is the most recent version of the model he developed in a series of papers before (1967, 1974 & 1976). The definition or model is rather complex (*cf.* Schaffner, 1993, 429), comprising several disjuncts which are supposed to guarantee that a continuum of relations ranging from identity-based reduction over different sorts of stronger and weaker forms of replacement to mere replacement or elimination is covered.

The motivation for an attempt to cover these latter cases of replacements consists in the assumption that an idealized version of straightforward reduction does not adequately match most cases of actual theory succession. These less straightforward theory successions are modeled as follows (ignoring the technical details): Schaffner suggests that we could construct corrected versions within the vocabulary of the reduced theory, such that this corrected version reduces to the reducing theory in the strict sense (based on cross-theoretic or cross-model identities), or that

we can construct a theory using the replacing theory, such that this construction becomes the reducer of the reduced theory. In fact, correction seems to be involved in many actual theory-successions, and, moreover, replacement issues became a general concern after Feyerabend suggested that there might be general problems in getting cross-theoretical identities due to the incommensurability of theoretical expressions (see, for example, (Feyerabend, 1962)). The relevance identity-based reduction plays even in Schaffner's liberal model can easily be pointed out.

Identity-based reduction enters the GRR model in two independent ways. Schaffner defines the GRR model such that it yields Nagelian reduction as a limiting case (Schaffner, 1993, 430). In this respect, identity-based reduction enters the game (and Schaffner is ready to accept that identity is the basis for Nagelian bridge-laws (Schaffner, 1993, section 9.4.2)). Moreover, cross-model or cross-theoretic identity is the basis for specific sorts of reductions in which correction is involved: In these cases, the question of identity-based reduction is shifted from the original pair of reduced and reducing theory to another pair which contains at least one dummy for one of the initial pair's members (*cf.* Schaffner, 1993, 429). Assume that we have a case of replacement, such that the replaced theory is possibly transformed into a corrected version which, in turn, stands in the relevant reduction relation to the replacing theory. In this case, identity-based reduction may re-occur: What is the relation between the corrected version of the replaced theory and the replacing theory? Schaffner's idea is that here we have, again, a case of reduction in the Nagelian spirit.

Thus, even though philosophers do not agree about whether or not Nagelian versions of reduction should be spelled out in terms of identity to yield an *appropriate* model of reduction, it seems obvious that prominent versions of Nagelian reduction (namely, Nagel's model as well as the GRR model) are supposed to comprise reductions which are based on identity. The idea of reduction as being based on identity is, however, more common in the philosophy of mind tradition.

1.3 Identity-theories and Reduction

In this section, three different sorts of identity-theories will be introduced and their place in the reduction debate will shortly be discussed. We start with classical identity theories, and then, in a next step, turn to the moves proposed by disjunctivists and contextualists (those who

suggest that we should identify physical kinds with either disjunctive or contextualized high-level kinds).

Feigl, Smart, Place

Identity-based models of reduction are usually associated with early type-identity theorists. Herbert Feigl gives the following characterization of type-identity theory as a version of reductionism (Feigl, 1967, 71 ff.), answering the question of "whether the mental and the physical can in some sense be identified":

> [I]t is proper to speak of "*identification* […]" Concepts of molar behavior theory like habit strength, expectancy, drive, instinct, memory trace, repression, superego, etc., may yet be identified in a future psychophysiology with specific types of neural-structure-and-process-patterns. (Feigl, 1967, 77)

Talking about *concepts*, Feigl seems to have *expressions* in mind, such that it is the kinds referred to by these expressions which are to be identified with the relevant *scientific kinds*. Combining the claim of identity with the list of examples given in this citation, we get the following claim: *Mental kinds are identical with physical kinds*, which is offered as an explication for the claim that *the mental and the physical can be identified,* or, perhaps less misleadingly: *The mental is to be subsumed under the physical.* Construed as a general thesis and ignoring modalities, type-identity theory about the mental thus states that *any mental type is identical to a physical type.* Similar interpretations of reductionism can be found in Smart (1959 & 1963) and Place (1956 & 1960), and it is this sort of reductionism which was famously attacked by Putnam (1967) and Fodor (1974), using a multiple realizability strategy. The concept of reduction which underlies this debate is obviously a concept which relies on the assumption that identification is at least a necessary condition for reduction: The reductionism defended by Feigl and his allegiance requires the mental to be identical to (part of) the physical.

Disjunctivism and Contextualized Kinds

More recent forms of this interpretation can be found in the writings of disjunctivists like Clapp (2001) and Walter (2006) and in texts of defenders of variants of contextualized-kind-approaches to reductionism. Building on earlier versions of functional reduction as defended in (Lewis, 1972), this idea can be found in different versions in Kim (1992),

Esfeld & Sachse (2007) and in Sachse (2007). Both positions respond to the well known multiple-realizability arguments against type-identity theory developed by Putnam (1967) and Fodor (1974). Disjunctivists argue that we should go disjunctive: Instead of subscribing to the claim that any mental kind is identical to a unique physical kind, we could get reduction of the mental by arguing that mental kinds are identical to the disjunction of kinds which realize our initial mental kind. Adherents of the contextualized-kind move argue that instead, we should talk about contextualized mental kinds as being identical to physical kinds. For example, if pain can be realized in octopuses and humans in different ways, then we could, maybe, identify human pain with one specific pain-realizer, and octopus pain with another.

These descriptions are utterly brief, and they do not come even close to capture the rich debate on these issues. However, these short remarks should suffice to point to the relevance identity-based reduction still plays in the philosophy of mind. Disjunctivists and those who opt for a version of the contextualized kind approach seem to rely on a model of reduction which comes still close to the initial model underlying classical type-identity theory, at least in the following respect: Arguing that we can possibly identify disjunctive or contextualized mental kinds with physical kinds, these camps follow the general understanding of reduction as being based on identity.

Let us now turn to the final dominant version of models of reduction. These models bear striking similarities to Schaffner's GRR model. Identity is, again, a crucial ingredient.

1.4 New-Wave Reductionism

The model (or better: the family of similar models) has been developed in a series of articles and books by Clifford Hooker (1981), Paul Churchland (1979 & 1985) and, more recently, by John Bickle (esp. 1998), and it builds on some aspects of Schaffner's model (1993) and earlier versions of it (Schaffner, 1967 & 1976 & 1977). Here is a formulation of Hooker's definition:

> Within T_B construct an analog, T^*_R, of T_R under certain conditions C_R such that T_B and C_R entail T^*_R and argue that the analog relation, AR, between T^*_R and T_R warrants claiming (some kind of) reduction relation, R, between T_R and T_B. Thus (T_B & CR $\rightarrow$ T^*_R) and (T^*_R AR T_R) warrants (T_B R T_R). (Hooker, 1981, 49)[5]

The conditions, CR, will consist of limiting assumptions and boundary conditions which guarantee that if T_B is more comprehensive than T_R, the application of elements of T_B's vocabulary is restricted to the domain relevant for T_R. The idea is to effect reductions which include corrections by working on the reducing theory (here 'T_B') and not by working with corrections formulated in the vocabulary of the reduced theory. Ronald Endicott nicely sums up the relevant features of New Wave reductionism (I quote):

(i) New-wave construction: the basic reducing T_B, not the original reduced T_R, supplies the conceptual resources for constructing the corrected T_R*.
(ii) New-wave deduction: the corrected T_R*, not the original reduced T_R, is deduced from the basic reducing T_B.
(iii) New-wave relation: there is a required analogical relation, not bridge laws, between the reduced T_R and the corrected T_R*.
(iv) New-wave continuum: there is a continuum of strong to weak analogies between the reduced T_R and the corrected T_R*, with the strong relations justifying retention and the weak relations justifying replacement of the ontology of T_R. (Endicott, 1998, 56)

Thus, the general framework is in accordance with Schaffner's GRR-model: Models of reduction should be able to cover cases of replacements based on correction within the base theory. Like the GRR-model, it aims at covering a continuum of relations between theories which make for reduction or replacement. Within certain boundary conditions (under which T_R* is derived), the reducing theory reflects aspects of the reduced theory. Reflection of aspects comes in degrees: The analogical relation mentioned in condition (iv) can come in various ways, some of which make for straightforward identity-based reductions whilst others do not. Identity (a limiting case for the continuum) ensures that the ontology of the reduced theory is preserved by the reducing theory. Let me make this idea more precise.

Officially, in New-Wave reduction, bridge-laws are replaced by *ordered pairs* of elements of the descriptive parts of the vocabularies of the two theories, which enable us to judge the degree of similarity between the derived image (which is formulated in the language of the reducing theory) and the reduced theory (Bickle, 1992, 223). Reduction is associated with a space of theory-relations ranging from "perfectly smooth" or "retentive" reductions to "bumpy" reductions, which

are best understood as mere replacements (Bickle, 1992, 223; Hooker, 1981, 45). For our concerns, the former case is more interesting. Bickle describes it as follows: (I_N is the derived image and T_O is the reduced or the "old" theory):

> In cases lying at or near the retentive endpoint, the I_N is the exactly equipotent isomorphic image of the T_o, and no counterfactual limiting assumptions or boundary conditions are required for the derivation of I_N. (Bickle, 1992, 223)

In this case, the pairing of the terms depends on identity of referents (Bickle, 1992, 224 & Bickle, 1998, 230) and I_N can be directly obtained from T_R (the reducing theory), whereas in cases on the opposite point of the spectrum, pairing is achieved only by reference to counterfactual limiting assumptions and boundary conditions. In this case, ontology is eliminated, like in the case of reduction of phlogiston-theory. Note, however, that any similarity or analogical relation should be given a reading which allows for it to come in degrees, such that there is a spectrum of reductions and replacements. In cases of perfectly smooth reductions, we thus have identity. In less stringent cases, we clearly have replacement and, thus, no interesting candidate for our present concerns.

In summary, Schaffner's GRR model and New-Wave reductionism rely on notions of reduction which do not qualify as being based on identity, according to the criteria given above. We could, however, isolate a specific sub-type of reduction which depends upon identity in the relevant sense, using the resources given by these models. That there is such a sub-type of reduction is implied by the respective models – according to each of them, identity-based reductions are possible, and they count as proper reductions. A similar interpretation is suggested in Nagel (1961) – at least some reductions seem to rely on cross-level identities. Variants of identity theories obviously rely on a model of identity-based reduction.

So, models of identity-based reduction occur in different fields of the reduction debate. If it could be shown that identity-based reduction is incompatible with reductive explanation, it could be shown that the models discussed above do cover cases of reduction which do not amount to reductive explanation. Let us now turn to *reductive explanation*.

2 Reductive Explanation

Reduction is traditionally assumed to be intimately related to explanation. In this section, one notion of reductive explanation which is, maybe, the most intuitive, and which is crucial for an understanding of Kim's argument, will be considered in more detail, and it will be distinguished from other interpretations of the notion of reductive explanation.

2.1 Reductive Explanation as Mechanistic Explanation

The model of reductive explanation Kim assumes to be incompatible with identity-based reduction can be regarded as what is often discussed under the heading of *mechanistic explanation*. The notion of *mechanistic explanation* has gained considerable attention in recent years (Kauffman, 1970; Bechtel & Richardson, 1992; Bechtel, 1994; Glennan, 1996: Machamer et al., 2000; Craver, 2006& 2007; Bechtel, 2008). Given the rich debate on this issue, and given the different interpretations of mechanistic explanation on the market, we should rely on a very general understanding of the issue at hand. That is: We should rely on the intuitive characterizations upon which models of mechanistic explanation as, for example, proposed by Bechtel and Craver, rely. Here is Kim's characterization of this sort of explanation:

> Suppose we explain an M-phenomenon in terms of P-phenomena. We now understand why, and how, this M-phenomenon arises from certain P-phenomena. It is because these particular P-phenomena constitute an underlying mechanism whose operations yield phenomena of kind M. (Kim, 2008, 94)

We should briefly reflect upon this short passage. Mechanistic explanation is tied to an understanding of constitution, and it relies on a notion of dependence which is intuitively captured by the concept of *arising from*. Now, consider the two pairs: *water* and H_2O and *temperature* and *mean kinetic energy*. In a sense, the molecular constituents of water are *constituted* by structures of H and O atoms, and temperature arises from the kinetic energy of collections of molecules. In the case of events, it may make sense to talk about a mechanism: Intuitively, the mechanism of the concerted activation of constituents of gases makes for the gases' specific temperatures, although the teleological connotations of the term 'mechanism' (which are explicit in Craver's definition, see, for example,

(Craver, 2005, 385)) might be misleading in the case of temperature and mean kinetic energy. Mechanistic or constitutive explanation should, however, not be tied to the explanation of the occurrence of events only (as it is in Craver's model, which crucially relies on an understanding of the explanandum of a mechanistic explanation in terms of a schema of the following sort 'S φ-ing' where 'φ' stands for expressions signifying events). For example, explaining the properties of water in terms of properties of H_2O – say, why frozen water has properties which are different from those of non-frozen water in terms of the properties of the lattice structures of I_h or I_c structures of sums of H_2O molecules – we do not explain the occurrence of events, but rather the occurrence of (non-temporally structured) *properties* of instances of water. Thus, the term 'mechanistic explanation' might be misleading. So let me explicitly state that this term is here used to cover explanations of occurrences of events as well as other explanations.

Interestingly, this sort of explanation seems to fit the example Nagel uses in order to give an idea of reductive explanation:

> [Once] the detailed physical, chemical, and physiological conditions for the occurrence of headaches are ascertained [...] an explanation will have been found for the occurrence of headaches. (Nagel, 1961, 366)

Here, we have an explanation of the occurrence of headaches in terms of *conditions for the occurrence* of headaches. The (physical, chemical, or physiological) conditions for the occurrence of headaches can also be described as the (physical, chemical, or physiological) mechanism which underlies headaches, or from which headaches arise.

This sketch should suffice to get access to Kim's argument against the compatibility of reduction and reductive explanation. It is, however, important to bear in mind that this sort of explanation is not the only one which figures under the heading of *reductive explanation*. Let us briefly reflect on rival notions of reductive explanation to avoid misunderstandings.

2.2 Alternative Interpretations of Reductive Explanation

The closest relative of the model just discussed is, seemingly, the idea of phenomena-explanation in the context of theory reduction: Friedman (1982) and Schaffner (1993, 469) assume that the reducing theory explains the reduced theory's *phenomena*. If this is to be interpreted

in the Nagelian spirit, namely, that the reducing theory should explain how the reduced theory's phenomena occur, we have, again, a case of reductive explanation as mechanistic explanation, although the idea is now tied to sets of phenomena of whole theories. Note that this sort of reductive *theory*-explanation is to be distinguished from the following interpretations:

For theory reduction, we find the assumption that it is concerned with *unification* which is, under at least one interpretation (see, for example, Kitcher, 1982&1989) a criterion for explanation. That theory-reduction is tied to unification is advocated by Nagel (1961, 354), Patricia Churchland (1986, 279), and Hooker (1981).

The more famous Nagelian idea that reduction is tied to explanation of theories by other theories is defended by Nagel (1961, 11-II), Churchland (1986, 283), Schaffner (1993, 429), and (Sklar, 1967). This interpretation comes in two forms: First, it can (according to, for example, Nagel) be conceived of as *direct* theory explanation, or, alternatively, as explanation of why the reduced theory worked as well as it did (Sklar, 1967; Schaffner, 429).

In (Enç, 1976) and in (Schaffner, 1993, 469), who in a first step repeats and then tentatively rejects this view, we find the idea that bridge-laws state causal laws and that, therefore, reduction is concerned with causal explanation. Note that given the three examples briefly discussed above, it should be obvious that mechanistic or constitutive explanation is distinct from causal explanation. Effects do not arise from their causes in this specific sense, nor are they constituted by their causes.

In addition, reductions are sometimes conceived of as closing *explanatory gaps*. Once we have reduced one entity (in the wide sense of the term) to another, an explanatory gap has been closed. As has been pointed out in the literature (Kim, 2008; Tye, 1999; Block&Stalnaker, 1999) this can be done in two different ways: Either by giving an explanation which closes the gap, or by showing that we just do not need an explanation.[6]

We are now in a position to get a sufficiently precise idea of the relevance and the meaning of Kim's claim: That identity-based reduction is incompatible with reductive explanation (as mechanistic explanation). Let us now turn to Kim's argument.

3 The (alleged) Incompatibility of Identity-Based Reduction and Reductive Explanation

It is difficult to judge the relation between *reduction* and *reductive explanation* (conceived of, from now on, as mechanistic explanation). Craver (2007), discussing a model of mechanistic explanation and mechanistic levels, shifts between reductionist and anti-reductionist descriptions of mechanistic (here figuring as 'constitutive') explanations. On one occasion, he writes:

> There are two dominant and broad traditions of thought about constitutive explanation: the reductive tradition and the systems tradition. My view is a development and elaboration of one strand on the systems tradition. (Craver, 2007, 108)

Thus, "his view" opposes reductive interpretations of mechanistic explanation. Some pages later, however, the following intuitive characterization of the system approach is given: It is 'to reduce a capacity to the programmed exercise of sub-capacities' (Craver, 2007, 110). So, is mechanistic explanation reductive, or isn't it? One might suggest solving this tension in the spirit of Kim (2008) and others (Chalmers, 1996, 43, Fodor, 1974, 107) who give an idea of reductive explanation in terms of mechanistic explanation. According to this view we can have reductive explanation *without* reduction. Kim goes a step further, claiming that we cannot have both at the same time: identity-based reduction and reductive explanation (of some specific phenomenon) contradict each other. This would nicely fit Craver's way of talking – that mechanistic explanations do not yield reduction, but, nevertheless, do reductively explain some capacity. So, what about Kim's argument?

Kim discusses this issue in a broader context of comparing three sorts of reduction (or three sorts of explications of the reduction predicate) – *bridge-law* reduction, *identity-based* reduction and *functional* reduction – with respect to their relation to reductive explanation. These distinctions will not be questioned, although it should be obvious that functional reduction, as, for example, defended by (Lewis, 1972), may turn out to be identity-based reduction as well. Moreover, the notion of identity-based reduction introduced above is covered by some accounts to reduction which describe the reduction relation in terms of bridge-laws. Kim uses the term 'identity reduction' in a slightly different way.[7]

He focuses on what has been called the *ontological* version of identity-based reduction above. For the moment, we should follow him in this and turn to representational reduction later on. Under one plausible interpretation of how the representational version gives rise to the relevant sort of reductive explanation, Kim's argument could be used to show that even the representational version is incompatible with reductive explanation.

So, what about Kim's argument that reduction which is based on identity precludes the possibility of reductive explanation? This is Kim's worry: Given that A=B, and A reduces to B, and A is to be explained in terms of B, we have a problem: Since nothing can be explained by itself (Kim, 2008, 102 f. & 106) – we cannot have both at once, identity-based reduction and the corresponding reductive explanation.[8] Note that others obviously have intuitions other than Kim. Consider the following quote:

> What reduction needs [...] is the idea that the 'reduced phenomenon' is made more comprehensible or intelligible by being shown to be identical with the 'reducing phenomenon'. (Crane, 2001, p. 54)

If, say, water is made more intelligible by being shown to be identical to H_2O, then it seems that water (or its behavior) is thereby somehow explained. Similarly, in Salmon, we find the idea that reductive explanations go together with identity:

> When, for example, we explain optical phenomena in terms of Maxwell's electromagnetic theory, the explanation is constitutive. Light waves are the electromagnetic waves (in a particular part of the spectrum) treated in Maxwell's theory (Salmon, 1984, 270)

So, who is right? In what follows, I will argue that we can have both at once-type-identity (conceived of as a symmetric relation) as a prerequisite for identity-based reduction and corresponding reductive explanation. In order to show this, we should explicitly state Kim's argument. Here is my reading of Kim's argument.

I assume that the necessary condition for the ontological version of identity-based reduction is, from Kim's point of view, sufficient to give rise to the relevant problem. Thus, any sort of identity-based reduction (which is such that according to it, 'A reduces to B' is true only if A=B) is incompatible with a corresponding reductive explanation of A

in terms of B. This partial explication of identity-based reduction could thus be used as a first premise:

P1) 'A reduces to B' is true only if A=B.

Secondly, Kim assumes that nothing explains itself, that is: That explanations must not be trivial. I suggest giving this premise the following form:

P2) An explanation is appropriate only if its *explanans* is not identical to its *explanandum*.

The terms 'explanans' and 'explanandum' should, for the moment, be interpreted in a wide sense. The correct reading of this premise will play a crucial role in locating the fallacy involved in Kim's argument. But what exactly does Kim try to show? Kim tries to show that if 'A reduces to B' expresses a truth such that A=B, then necessarily, a corresponding reductive explanation is false. In this sense, reductive explanation is incompatible with identity-based reduction. But what is the 'corresponding reductive explanation' of a statement of the form 'A reduces to B'?

For P1) and P2) to imply that identity-based reduction is incompatible with reductive explanation, we have to rely on the assumption that the pair of the *explanans* and the *explanandum* of a reductive explanation which corresponds to a reduction statement of the form 'A reduces to B' (ontologically understood) is formed by the instances of A and B. (Otherwise, it would be hard to see how identity-based reduction contradicts reductive explanation in virtue of being identity-based – intuitively, this requires identity of A and B to be sufficient for identity of *explanans* and *explanandum* of the corresponding explanation). Unfortunately, this interpretation faces a grammatical problem – instances of 'A' and 'B' in statements of the form 'A reduces to B' are singular terms, whereas well formed sentences of the form 'C because D' and 'C by D' require at least one of the instances of 'C' and 'D' to be a sentence (only the instance of 'D' in the second alternative need not be a sentence.) Thus, it seems more appropriate to describe *explanans* and *explanandum* of a reductive explanation to be the instances of 'A' and 'B' in sentences of the following form: 'B (reductively) explains A'.[9] This allows us to move from 'A reduces to B' to a corresponding statement ('B reductively explains A'), which states that a specific explanatory link

holds true, and which must, according to Kim, be always false, if the reduction-predicate in 'A reduces to B' is interpreted as being identity-based. (This move, which is at odds with ordinary use of 'explanation', 'explanans' and 'explanandum', is merely of heuristic value to follow Kim's argument.)

Now, it should be obvious how we can introduce the relevant third premise, giving the relevant connection between the *relata* of the reduction relation and the *explanans* and the *explanandum* of the corresponding explanation.

P3) For any statement of the form 'A reduces to B', A and B form the *explanandum* and the *explanans* of the corresponding explanation (-statement).

From this, we obtain the conclusion that for any reduction statement of the form 'A reduces to B' which requires 'A=B' to be true, its corresponding explanation is false. True sentences stating identity-based reductions require their corresponding explanations to be such that its *explanans* is identical to its *explanandum* (P1 and P3). Thus, the truth of the reduction statement guarantees the falsity of the corresponding explanation (P2). This is what Kim seems to have had in mind, claiming identity-based reduction to be incompatible with reductive explanation. Based on a more precise interpretation of premise P2), we will be able to attack premise P3) later on (ignoring the problem of describing sentences of the form 'x explains y' as explanations). Before doing so, let us connect this idea to the representational version of identity-based reduction.

Under one plausible interpretation of the relation between the representational version of identity-based reduction and the relevant kind of reductive explanation, a similar problem would reoccur for the representational version of identity-based reduction. This would be the case if for any bridge law L of the form A=B, which partly connects a given A-theory T_R to a given B-theory T_B, such that partly in virtue of L, T_R reduces to T_B, there was a reductive explanation of the following sort: A is reductively explained in terms of B. If this were the case, then we could run Kim's argument for representational versions of identity-based reduction. The bridge-principles would function in a way similar to the reduction statements of the ontological version of identity-based reduction.

The argument I am about to put forward against Kim's conclusion takes two steps. In a first step, I discuss the example of the mirror-neuron mechanism as the reducer of the mechanism of social cognition, which is supposed to illustrate the odd consequences Kim's argument would have. In a next step, I consider premise P2): It will be argued that although it is true that instances of 'p because p', or 'A explains A' are inappropriate, it might very well be the case that some instances of 'p because q' or 'A explains B' turn out to be true even if the state of affairs that p= the state of affairs that q, or A=B. An appropriate understanding of P2) is an understanding which individuates an *explanans* and an *explanandum* on the level conceptual contents expressed by instances of 'p', 'q', 'A', and 'B'. This interpretation is incompatible with Kim's premise P3), which has, thus, to be rejected (or the terms 'explanans' and 'explanandum' become ambiguous, such that the argument turns out to be invalid).

4 Why the Argument Fails

Let us now turn to the example. It will be shown that *prima facie*, we have a case of reductive explanation which is compatible with identity-based reduction. In the next step, a different interpretation of Kim's principle that nothing explains itself will be suggested, and it will be argued that this reading is plausible.

4.1 An Example

I will present an idealized reconstruction of the reductive strategy of giving an explanation of social cognition in terms of simulation and the mirror neuron mechanism. This reductive strategy will be regarded as being based on identity – once we come to see that the mirror-neuron mechanism (the kind) makes for social cognition, we come to see that the mental mechanism which grounds the capacity to cognize the other as social (a kind) is identical to the mirror neuron mechanism (the kind). Based on this interpretation, it can be shown that *prima facie*, identity-based reduction is compatible with reductive explanation: Given the reductionist interpretation of this case, it can be shown that the mirror neuron mechanism reductively explains our mental capacity to cognize the other as social. In order to do so, let us stipulate that the human

mechanism of social cognition reduces to and is, thus, identical to the mirror neuron mechanism. Note that Kim would presumably not buy this premise. However, this does not pose a problem: We just assume that they are identical and then check whether or not this assumption is compatible with the relevant explanatory link.

Let me firstly introduce the *questions* associated with social cognition as it is treated in the intersection of cognitive science and neuroscience. In a next step, I shall briefly introduce one version of how to make the *physiology* of mirror neuron mechanisms explicit (following de Bruin (2010)). In a third step, the sort of *explanation* of social cognition in terms of mirror neuron processes will be considered.[10]

4.1.1 Some Questions of Social Cognition

The question a theory of social cognition tries to answer can be put as follows: *How* does the relevant part of our mind contribute to the cognition of others as social beings? Or: *What* is the mechanism which enables us to cognize others as social beings? Or, once we have found a plausible candidate for this mechanism, we may ask: *Why* is this specific mechanism the mechanism which enables us to cognize others as social beings? Several answers have been given in the literature, among the most prominent of which figures *Theory-Theory* (Gopnik & Wellman, 1992; Carruthers, 1996) and *Simulation Theory* (Goldman, 2006; Gallese, 2000; Gallese et al., 2004). Whereas Theory-Theory states that in some sense or another, we develop or are born with a theory containing law-like structures which link an observational input to beliefs about the emotions or intentions underlying this input, Simulation Theory claims that we should think of mind-reading as being implemented by a mechanism of simulation – we simulate the other, and, because we are relevantly similar to her, we get a grasp of her as a social being. Let us now turn to the physiology of the mirror neuron mechanism.

4.1.2 Aspects of the Physiology of Mirror-Neuron Systems

Basically, mirror neuron systems are expected to be located in the motor system (Gallese, 2000). Mirror-neurons are visuomotor neurons. They are active when an action is *observed* as well as during the *execution* of an action. Their activation during action-observation of others supposedly explains the immediate, automatic understanding of others as agents. Gallese (2001) puts it as follows:

[W]hen we observe goal-related behaviors [...] specific sectors of our premotor cortex become active. These cortical sectors are those same sectors that are active when we actually perform the same actions. In other words, when we observe actions performed by other individuals our motor system 'resonates' along with that of the observed agent (Gallese, 2001, 38f.).

The existence of mirror neurons was first hypothesized on the basis of studies with macaque monkeys, in three cortical regions: The superior temporal sulcus (superior temporal cortex), area F5 (inferior frontal cortex) and area PF (posterior parietal cortex) (Keysers and Perrett, 2004). On the assumption that human imitation should be a good guide to discover functionally similar regions in the human brain (Graften et al., 1996; Rizzolatti et al, 1996), researchers hypothesized that during imitation, the activity of mirror neurons should be approximately the sum of activity of neurons during action observation and execution. Given that such high-activity during imitation could be found, there would be reason to assume that there are mirror neuron systems in the human brain. Iacoboni et al (1999) actually found two areas which showed this increase in activity during imitation: one was found in the pars opercularis of the inferior frontal gyrus (inferior frontal cortex), the second was found in the posterior parietal cortex. We do not have to go into further details concerning systems modeling goal-directedness of behavior or movements. For a detailed discussion, see de Bruin (2010). We are now in a position to evaluate the explanatory power of the mirror neuron system for an understanding of social cognition.

4.1.3 From Simulation to the Mirror-Neuron System

The explanatory pattern underlying the move from talk about a mechanism of social cognition to a detailed description of the mirror-neuron mechanism can best be understood when the different levels of description are compared to each other. We should thus consider examples of high-level talk and its connection to lower level talk about the mechanism which underlies the human capacity to cognize the other as social.

The first quote is taken from Alvin Goldman (2005). It is a pre-theoretical description of what Goldman takes to be the basis of social cognition:

1. First, the attributor creates in herself pretend states intended to match those of the target. In other words, the attributor attempts to put herself in the target's "mental shoes."

2. The second step is to feed these initial pretend states, e.g., beliefs, into some mechanism of the attributor's own psychology … and allow that mechanism to operate on the pretend states so as to generate one or more new states, e.g., decisions.
3. Third, the attributor assigns the output state to the target … e.g., we infer or project the decision to the other's mind (Goldman, 2005, 80f).

It is noteworthy that this description employs only psychological, functional and ordinary language expressions. Even the notion of a mechanism can be understood such that it does not involve the concept of a *neural* mechanism (even though, in the context of the book, it seems clear that at least reference to some neural mechanism is intended). However, even if the relevant description were 'some neural mechanism', the only neuroscientific terminology being employed was the predicate '_is neural'. Thus, we have a description which does not contain any neuroscientific meat, so to speak: It is alluded to, but it is not used in the description. What the mechanism does is given by some high-level description, and what this mechanism consists in is explained moving to lower levels of description.

These lower level descriptions are taken from Gallese's, Keysers' and Rizzolatti's (2004) paper. Here I present them in a structured way, which is supposed to help us differentiating between two levels (part a) and part b)), and understanding how these levels supposedly relate (part c)):

a) Here we will argue, however, that the fundamental mechanism that allows us a direct experiential grasp of the mind of others is not conceptual reasoning but direct simulation of the observed events
b) through the mirror mechanism.
c) The novelty of our approach consists in providing for the first time a neurophysiological account of the experiential dimension of both action and emotion understanding. (Gallese et al., 2004)

Here is a description of the "fundamental mechanism" in more detail (taken from (Rizzolatti et al., 2000)):

b*) [W]e propose that the mirror system is a basic system for the recognition of action …. According to this view, there are "vocabularies" of motor actions at the core of the cortical motor system. Neurons forming these vocabularies store both knowledge about an action and the description … of how this knowledge should be used. The ensemble of neurons related to a given action forms the global motor schema of that action. When an appro-

priate stimulus is presented, the relevant schema is activated. (Rizzolatti et al., 2000, 549 f)

In the following passage, we are confronted with a psychological description, which is then used to define the term „simulation". Simulation, in turn, is used to specify a neural mechanism which is said to underlie the psychological states:

d) What makes social interactions so different from our perception of the inanimate world is that we … carry out similar actions and we experience similar emotions. There is something shared between our first- and third-person experience of these phenomena: the observer and the observed are both individuals endowed with a similar brain-body system. A crucial element of social cognition is the brain's capacity to directly link the first- and third-person experiences of these phenomena (i.e. link 'I do and I feel' with 'he does and he feels'). We will define this mechanism 'simulation'. (Gallese et al., 2004, 397)

In this light, we can, abstracting from the details, describe the reductive move as follows:

'Mechanism of an F':	(*Goldman, Gallese et al. – a)* (see quote above)): where 'F' is a general term the meaning of which consists of psychological and ordinary language concepts only. A rough outline of the mechanism's behavior in high-level terminology is given, and its functional role is described. For the input, we would get (not quoted here): some observed behavior (in context). For the output, we get: projection or inference.)
'Simulation':	(*Gallese et al. – d)* (quote above) Paralleled in Goldman, above): Here, high-level talk is employed to define the term 'simulation', even though the mechanism is explicitly located in the brain. These statements are best interpreted as stating some identity between simulation and the relevant mechanism. So is c).

'Mirror neuron mechanism': (*Gallese et al. – b)* and *Rizzolatti et al. – b*)* (quotes above)): The mechanism is here described in neuroscientific terminology. It is described in terms of location (motor cortex), type of information stored (motor schema), and in terms of the causal role it plays (stimulus presentation/schema activation).

Given that a variant of type-identity is possible at least in principle (a pre-requisite for the notion of reduction we are interested in), the idea underlying this move can be conceived of as follows: At least for humans, the mechanism of social cognition (or some relevant part of this mechanism) is identical to a neural mechanism located in the motor-cortex and definable in neuroscientific terms. An identity-based reductionist reading of these steps could thus be phrased as follows:

W*hatever enables us to conceive the other as a social object (high-level behavioral description)* is identified via *some functional description of being a simulation which is supposed to cover the relevant functional aspects of the underlying mechanism* with the *mirror-neuron-mechanism (lower-level neuroscientific description)).*

Now, consider a set of reductive explanations associated with mirror-neuron theory, and then recall again Kim's argument. In virtue of the (allegedly identity-based) reduction of social cognition (in humans) to simulation understood as a mirror-neuron-based process, we are supposed to answer the questions listed above (the answers given here are dummies for appropriate answers):

Question: How does the relevant part of our mind contribute to the cognition of others as social beings?
Answer: It does so by engaging in events of the following sort (given the relevant input of an action of type A by person x): the mirror-neuron system's stored information about (a motor-schema of) an action of type A is activated, which simultaneously constitutes a perception (or recognition) of an instance of type A realized by x.
Question: What is the mechanism which enables us to cognize others as social being social?
Answer: It is the mirror-neuron-system, a part of the motor system, which is triggered by of an event of…

Question: Why is this mechanism the mechanism which enables us to cognize others as social beings?
Answer: Because this mechanism is the mirror neuron system which …

What does this show? I think it shows that in a sense, the mirror neuron mechanism perfectly explains the mechanism which enables the human to cognize the other as social. At least, it is not the case that these explanations are trivial, even under an interpretation according to which the mirror neuron mechanism is identical to the mechanism which underlies the human capacity to cognize the other as social. It seems that here, we have a clear case of mechanistic explanation: A phenomenon (our brain's capacity which enables us to cognize the other as social) is explained by the constituents and their interaction of the relevant part of the brain. So, we have *prima facie* reason to assume that identity does not preclude mechanistic explanation, and that mechanistic explanation at least covers some reductive explanations.

4.2 Reconsidering Kim's Argument

However, we do not only have *prima facie* reason to doubt that Kim's argument is sound. Evaluating the reason for why we should believe in the truth of P2), we see that it should be given a specific interpretation which differs from the interpretation it is given in Kim's argument, such that premise P3) turns out to be false, even under the liberal understanding of 'explanation', 'explanans' and 'explanandum' I suggested.

Kim's principle seems to rely solely on the requirement that explanations must not be trivial. That is: 'p because p' or 'A explains A' do not have proper explanations as instances, and therefore the *explanans* and the *explanandum* of an explanation must not be identical. But this requirement is perfectly compatible with identity-based reduction as being tied to reductive explanation, such that from the requirement that explanations must not be trivial, it does not follow that one thing cannot be explained by itself. Let us thus try to evaluate which interpretation of P2), that an *explanans* and an *explanandum* must not be identical in a correct explanation, is supported by the assumption that explanations must not be trivial, or that 'p because p' or 'A explains A' have incorrect instances only.

In these cases, identity of *explanans* and *explanandum* concerns at least three linguistic levels: The expression (type) of *explanans* and

explanandum is the same on both occasions; assuming that the expressions are used in an identical way, their semantic values (their designata or what is signified by a sentence or a predicate) are the same; and, under the same assumption, they have the same meaning, or conceptual content. These latter notions will remain unexplained here. Conceptual content comes close to Fregean sense. It is what we understand when we understand an expression, and two synonymous expressions express the same conceptual content, whereas co-referential expressions may express different conceptual contents. For Kim's argument to go through, the principle must reflect the fact that the semantic values (the designata, or what is signified) of an *explanans* and an *explanandum* must not be identical. But is this interpretation plausible?

The example of mirror neuron systems and social cognition seems to show that identity of semantic values (such as kinds, or maybe facts) does not form a problem. We have identity of referents and, at the same time, reductive explanation. And this seems to be correct: That explanations must not be trivial should be regarded as being concerned with the epistemology of explanations. Trivial explanations are trivial because in such cases, what explains is informationally equivalent to what is explained. The example discussed above shows that one and the same thing can be described in distinct ways, such that the explanations given are not (informationally, or epistemically) trivial.

It seems more plausible that the principle that explanations must not be trivial is concerned with conceptual contents. Assume that the expressions 'the mirror neuron system' (or any neuroscientific description of the mirror neuron system) and 'the mechanism of social cognition' were synonymous. This would dramatically alter the example; it seems that the relevant questions (like: What is the mechanism of social cognition?) could not be answered properly using an expression which is synonymous (and, thus, informationally equivalent) with the expression used to describe the target of the explanation in the question. If so, meaning-identity or sense-identity would indeed form a problem.

Given that explanations of the form 'p because p' or 'A explains A' do not form a problem solely in virtue of the fact that the *explanans* and the *explanandum* are spelled identically, but are problematic because the pairs of expressions flanking 'because' or 'explains' respectively have the same meanings, then principle P2) should be interpreted as follows: That an *explanans* and an *explanandum* of an explanation must not be

identical is true iff *explanans* and *explanandum* are individuated on the level of meanings, or Fregean sense, or conceptual content. Thus, the referents of the instances of 'A' and 'B' in true sentences of the form 'A is reductively explained by B' do not form the *explanans* and the *explanandum* of the corresponding explanation. If this is correct, Premise P3) is false – it suggests that explanans and explanandum are what is required to be identical by the reduction statement. But what is required to be identical by the reduction statement are the *referents* of the instances of 'A' and 'B', rather than their conceptual contents.

What Kim misses is this: Explanations are not individuated by the properties and individuals their constituents pick out in order to explain, but rather by the way they present us with these properties and objects. This is pointed out by Ned Block (who uses the term 'fact' in the way I would use 'proposition'):

> Just as knowledge of the fact that freezing happened is not knowledge of the fact that lattice-formation happened, so also an explanation of the fact that freezing happened is not an explanation of the fact that lattice-formation happened. By contrast: just as the time at which freezing happened is the time at which lattice-formation happened, so the cause of freezing is also the cause of lattice formation. (Block, forthcoming)

Block's point is that we can give different explanations which both refer to the same objects (cause) in the *explanans*. He argues that the cause of freezing=the cause of lattice formation. However, the explanation of why lattice-formation happened is different from the explanation of why freezing happened. This is so because the *explanans* and the *explanandum* have, in both cases, *different meanings*, although they give access to the same state of affairs, or law-like connection. For an explanation to be *trivial*, meaning identity, or identity of conceptual content, or identity of Fregean sense is required. This is Block's point, or so it seems: knowing that freezing happened is not knowing that lattice-formation happened, because 'lattice formation' and 'freezing' have different meanings. So, the principle that *nothing* explains *itself* seems to require quantification over entities *as presented by a certain meaning* (or, maybe, quantification of pairs of entities and concepts under which these entities are presented). Similarly, it should be clear that conceptual difference between *explanans* and *explanandum* is a *requirement* for the explanation of the human capacity to cognize the other as social

in terms of mirror neuron mechanisms to turn out acceptable. Similarly, nothing is water because it is water, though one might suggest that it is water *because* it is H_2O (Mulligan, 2006). Thus, conceptual difference is, again, the relevant point here: Even though water is identical to H_2O, H_2O can, in some sense, explain water. *H_2O* comes in the appropriate conceptual shape to do so. Thus, difference in meaning is a pre-requisite for reductive explanation. But conceptual difference does not necessarily translate into ontological difference. One and the same thing can be characterized by different descriptions, and it can be given in different ways. The principle Kim seems to allude to, that nothing explains itself, is, as he interprets it, misguided. Therefore, there is, at least *prima facie*, no reason to assume that reductive explanation is incompatible with identity-based reduction. It is, at least, conceptually possible that the freezing of water is just the lattice formation of H_2O-molecules, or that the mechanism which enables humans to cognize the other as social reduces to the mirror neuron mechanism. This does not render the relevant reductive explanations false. P2) is correct only under a narrow interpretation of the notions of an *explanans* and an *explanandum*, namely, as conceptual contents of the expressions used stating this explanation. Under this narrow interpretation, it is possible that A explains B even if B=A.

5 A Lesson to be Learned

What can be learned from Kim's misunderstanding? It should be clear that given that explanations must not be trivial in the sense specified above, for identity-based reduction to yield reductive explanation, the relata of the reduction relation must comprise concepts, in addition to the relevant kind. If a sentence of the form 'A reduces to B' is true such that 'B reductively explains A' is true, or 'A can be explained in terms of B' is true, then (i) A=B and (ii) the concept expressed by an appropriate instance of 'A' differs from the concept expressed by an appropriate instance of 'B'. This perfectly matches paradigmatic cases of reductions: water reduces to H_2O, water=H_2O, and the conceptual contents of 'water' and 'H_2O'differ (even if water is a rigid designator lacking conceptual content, they would differ, given that 'H_2O' has a conceptual content). Similarly, the conceptual contents of 'pain' and 'C-fiber

stimulation' differ, pain just is (by assumption) C-fiber stimulation, and the former reduces to the latter. I assume that we can generalize here. A general requirement for an explication of reduction is this:

If an instance of 'A reduces to B' expresses a truth, then
(i) The referent of the appropriate instance of A= the referent of the appropriate instance of B, and
(ii) the conceptual content of the appropriate instance of A ≠ the conceptual content of the appropriate instance of B.

(To give a counter-instance, we would need an instance of the schema 'A reduces to B' which is such that the instances of 'A' and 'B' were synonymous. Little reflection on this point seems to show that it would be absurd to claim that such an instance could be found, which expresses a truth.)

Here are a few remarks on the form and on the content of this requirement. Let us stipulate that the conceptual content of an expression A ≠ the conceptual content of an expression B iff either both have a conceptual content, and these conceptual contents are not identical, or only one of them has a conceptual content. Thus, according to this interpretation, at least one of the expressions of an appropriate instance of 'A reduces to B' may be a rigid designator lacking conceptual content.

This is given in a schematic, meta-linguistic fashion, because we have to talk about an expression's referent and its conceptual content at the same time. Using schematic expressions makes it relatively easy to do so.

So, what does this mean for identity-based reduction? It means that any model of identity-based reduction should take this feature into account. Strictly speaking, even in cases of identity-based reduction, nothing reduces to itself. The relata of the reduction relation comprise more than just a certain kind or phenomenon – it is conceptual contents which play a crucial role in this context. Similarly for bridge-principles: If 'A=B' is a bridge-principle in virtue of which one theory reduces to another (in a non-homogeneous case of reduction), then 'A' and 'B' must express different conceptual contents. Note that this is in accordance with Nagel's assumption that bridge-principles must not be analytic – if 'A' and 'B' express the same conceptual content then they are analytic and, moreover, evidently true.

Conclusion

We have seen that the principle that nothing explains itself, or that nothing can be explained by itself does not apply to a case of identity-based reduction. It would apply (in general) if identity-based reduction required the relata of the reduction relation to be kinds only, or if it required the conceptual contents which seemingly play a role in reduction to be identical too. This gave reason to reconsider the relata of identity-based reductions. Identity-based reduction is partly defined by the conceptual contents under which the relevant kinds are presented.

Notes

1 I would like to express my gratitude to Albert Newen, Stephan Hartmann, Leon de Bruin, and an anonymous Referee for helpful sugesstions and discussions. I would also like to thank the audience at the workshop in Bremen in 2009, where a previous version of this paper was intensively discussed, and the group in Bochum for the stimulating atmosphere in which this paper emerged.

2 For an overview on the candidates which might allow us to adequately model the notion of reduction, like identity, supervenience, derivation etc, see Van Gulick 1992.

3 Note that this distinction is, to some degree, arbitrary. For example, it has been argued that models of reduction underlying New-Wave reductionism collapse into more recent versions of Nagelian reduction (Endicott 1998&2001; Dizadji-Bahmani et al., 2010). Moreover, Nagel's model played a crucial role in the philosophy of mind (it is, for example, alluded to in (Kim 1993, 150 & 248)). And finally, the idea of type-identity theory as a variant of reduction is nowadays discussed within a wider context, that is: it is not only pertinent in the philosophy of mind, but also in related areas (cf. Sachse, this volume). However, the structure of the following paragraphs should be sufficient to point to the importance of identity-based reduction in the reduction debate.

4 To be sure, Nagel also suggests a number of different interpretations, ranging from syntactic characterizations as bi-conditionals or mere conditionals to epistemological characterizations (like *postulating a hypothesis* (Nagel 1961, 354)).

5 This is equivalent to Paul Churchland's (1985) and Bickle's (1992) model of reduction.

6 In fact, Kim argues that identity-based reduction can merely close explanatory gaps without explaining anything, namely, by showing that we do not need an explanation. See footnote 7 for a brief discussion.

7 Kim labels this sort of reduction simply 'identity reduction'. The gram-

matical structure of this expression allows for a misleading reading: It is not the reduction of identity this type of reduction is concerned with. This is why I prefer talking about *identity-based* reduction.

8 In section III of his (2008) paper, when identity reduction is officially considered, Kim argues that identity-based reduction does not yield reductive explanation because it does not give an answer to questions of the form 'Why does x correlate with y?'. Kim focuses on the explanatory gap problem and suggests that identity-reduction, in principle, cannot close the explanatory gap by giving a reductive explanation, but rather by showing the question to be misguided (it seems to me that this is roughly the idea pertinent in (Tye 1999) and (Block&Stalnaker 1999)). Unfortunately, Kim's paper suffers from a serious problem. When identity-based reduction is considered, Kim introduces the notion of reductive explanation as mechanistic explanation. In a first step, however, he gives an argument which is independent of this notion of explanation. He rather argues that the scientific value of identity-based reduction just is not explanation, but rather showing that where we believed to be an explanatory gap, or where we believed to be a possible explanation, there just is none. Note that there is a fundamental difference between the types of questions associated with explanatory gaps on the one hand and mechanistic explanation on the other: Once we have shown a given A to be identical to some B, we stop questions of the sort: Why do A and B co-occur? Or: Why is A correlated with B? In Block's and Stalnaker's terms:

"If we believe that heat is correlated with but not identical to molecular kinetic energy, we should regard as legitimate the question of why the correlation exists and what its mechanism is. But once we realize that heat is mean kinetic energy, questions like this will be seen wrongheaded." (Block & Stalnaker 1999, 24)

These questions are not questions associated with mechanistic explanations, like: How does S φ? (Maybe: How does heat occur? Or: How does a gas' heat increase? Or: How does gas change its temperature?) The *mechanism* Block and Stalnaker talk about is a mechanism allegedly underlying a mere correlation between heat and molecular kinetic energy. The mechanisms asked for in the 'How' questions just listed does not concern the *connection* between heat and molecular kinetic energy. These 'How' questions are used to ask for the mechanism of the occurrence of heat or the change of temperature.

9 It might very well be the case that this explanation-relation is derivative upon explanation relations expressed by 'because' or 'by'. For example, one might hesitate to assume that H_2O *explains* water in a non-derivative sense, although it can be explained why water has specific properties by referring to specific properties of H_2O.

10 Note that the model of simulation is here treated as an *actual* model as opposed to being a *mere* model – it is treated as being literally true of the relevant system (for an elaboration on this distinction, see (Craver, 2006)). This is a questionable assumption (see (Gallagher, 2007) for a discussion), but for the sake of simplicity, I shall stick to this interpretation. Moreover,

alleged problems of the explanatory power of the mirror neuron mechanism will be ignored. The most important criticisms are Jacob & Jeannerod (2005). In particular, they argue that recognition of an agent's intention which is prior to her action (and not to be conflated with her motor-intention), her social and her communicative intention are not captured by the model of a mirror-neuron system of social cognition.

Bibliography

Bechtel, William, 2008: *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Bechtel, William, 1994: Levels of descriptions and explanation in cognitive science. In: *Minds and Machines* 4, pp. 1–25.

Bechtel, William & Richardson, Robert, 1992. In: Emergent phenomena and complex systems. In: Kim, Jaegwon, Beckermann, Ansgar, and Flohr, Hans (eds.): *Emergence or Reduction? Essays on the prospects of nonreductive physicalism*. Berlin: de Gruyter, pp. 257–288.

Bickle, John, 2001: Understanding Neural Complexity: A Role For Reduction. In: *Minds and Machines*, 11, pp. 467–481.

Bickle, John, 1998: *Psychoneural Reduction: The New Wave*, Cambridge, MA: MIT Press.

Bickle, John, 1992: Mental Anomaly and the New Mind-Brain Reductionism. In: *Philosophy of Science*, 59, pp. 217–230.

Block, Ned, forthcoming: Functional Reduction. In: Sabatés, M; Sosa, D.; Horgan, T (eds.): *Supervenience in Mind: A Festschrift for Jaegwon Kim*. Cambridge, MA: MIT Press.

Block, Ned, 2007: *Consciousness, Function, and Representation*. Cambridge, MA: MIT Press.

Block, Ned, 2002: The Harder Problem of Consciousness. In: *The Journal of Philosophy* 99, pp. 391–425.

Block, Ned and Stalnaker, Robert, 1999: Conceptual Analysis, Dualism, and the Explanatory Gap. In: *Philosophical Review* 108, pp. 1–46.

Block, Ned, 1997: Anti-reductionism Slaps Back. In: *Philosophical Perspectives* 11, pp. 107–132.

Block, Ned, 1995: The Mind as the Software of the Brain. In: Smith, E. E. and Osherson N. (eds.): *An Invitation to Cognitive Science 3: Thinking*. Cambridge, MA: MIT Press, chapter 11, pp. 377–426.

Block, Ned, 1978: Troubles with Functionalism. In: *Minnesota Studies in the Philosophy of Science* 9, pp. 261–325.
Block, Ned and Fodor, Jerry A., 1972: What Psychological States Are Not. In: *Philosophical Review* 81, pp. 159-181.
de Bruin, Leon (2010), *Mind in Practice*. Veenendaal: Universal Press.
Causey, Robert, 1977: *Unity of Science*. Dordrecht, NL: Reidel.
Causey, Robert, 1972: Attribute Identities in Microreductions. In: *Journal of Philosophy* 64, pp. 407–422.
Chalmers, David, 1996: *The Conscious Mind*. Oxford, UK: Oxford University Press.
Churchland, Patricia, 1986: *Neurophilosophy*. Cambridge, MA: MIT Press.
Churchland, Paul M., 1985: Reduction, Qualia and the Direct Introspection of Brain States. In: *Journal of Philosophy* 82, pp. 8–28.
Churchland, Paul M., 1981. Eliminative Materialism and the Propositional Attitudes. In: *The Journal of Philosophy* 78, pp. 67–90.
Churchland, Patricia, 1980: A Perspective on Mind-Brain Research. In: *The Journal of Philosophy* 77, pp. 185–207.
Clapp, Lenny, 2001: Disjunctive properties and multiple realizations. In: *The Journal of Philosophy* 98, pp. 111–136.
Crane, Tim, 2001: *Elements of Mind*. Oxford, UK: Oxford University Press.
Craver, Carl, 2007: *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford, UK: Oxford University Press/ Clarendon Press.
Craver, Carl, 2006: When Mechanistic Models Explain. In: *Synthese* 153, pp. 355-376.
Craver, Carl, 2005: Beyond Reduction: Mechanisms, Multified Integration and the Unity of Neuroscience. In: *Studies in the History and Philosophy of Biological and Biomedical Sciences* 36, pp. 373–395.
Dehaene, S. and Naccache, L., 2001: Towards a Cognitive Neuroscience of Consciousness: Basic evidence and a Workspace Framework. In: *Cognition* 79, pp. 1–37.
Dizadji-Bahmani, F., Frigg, R. and Hartmann, S. 2010: Who is Afraid of Nagelian Reduction. In: *Erkenntnis* 73, pp. 393–412.
Enç, Berent, 1976: Identity Statements and Microreductions. In: *Journal of Philosophy* 73, pp. 285–306.

Endicott, Ronald, 2001: Post-Structuralist Angst – Critical Notice: John Bickle, Psychoneural Reduction: The New Wave. In: *Philosophy of Science* 68, pp. 377–393.

Endicott, Ronald, 1998: Collapse of the New Wave. In: *The Journal of Philosophy* 95, pp. 53–72.

Esfeld, Michael. A. and Sachse, Christian, 2007: Theory reduction by means of functional sub-types. In: *International Studies in the Philosophy of Sciences* 21, pp. 1–17.

Feigl, Herbert, 1967: *The "Mental" and the "Physical". The Essay and a Postscript*. Mineapolis: University of Minnesota Press.

Feyerabend, Paul K., 1962: Explanation, Reduction, and Empiricism. In: Maxwell, G.; Feigl, H. (eds.): *Scientification Explanation, Space and Time*. pp. 28–97.

Fodor, Jerry A., 1974: Special Sciences: Or the Disunity of Science as a Working Hypothesis. In: *Synthese* 28, pp. 97–115.

Fodor, Jerry A., 1997: Special Sciences: Still Autonomous After All These Years. In: *Noûs Supplement: Philosophical Perspectives, 11, Mind, Causation, and World* 31, pp. 149–163.

Friedman, K., 1982: Is Intertheoretic Reduction Feasible? In: *The British Journal for the Philosophy of Science* 33, pp. 17–40.

Gallese, V., 2001: The "Shared Manifold" Hypothesis: from mirror neurons to empathy. In: *Journal of Consciousness Studies* 8, pp. 33–50.

Gallese, V., 2000: The inner sense of action: agency and motor representations. In: *Journal of Consciousness Studies* 7, pp. 23–40.

Gallese, V.; Keysers, C. and Rizzolatti, G., 2004: A unifying view of the basis of social cognition. In: *Trends in Cognitive Sciences* 8, pp. 396–403.

Gallagher, Shaun, 2007: Simulation trouble. In: *Social Neuroscience* 2, pp. 353–365.

Glennan, S., 1996: Mechanisms and the Nature of Causation. In: *Erkenntnis* 44, pp. 49–71.

Goldman, Alvin, 2005: Imitation, Mind reading, and Simulation. In: Chater, Nick; Hurley, Susan (eds.): *Imitation, Human Development, and Culture*. Cambridge, MA: MIT Press, pp. 79–94.

Goldman, Alvin, 2006: *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford, UK: Oxford University Press.

Gopnik, A. & Wellman, H., 1992: Why the Child's Theory of Mind Really is a Theory. *Mind and Language* 7, pp. 145–171.

Grafton, S. T., Arbib, M, Fogassi, L., and Rizzolatti, G., 1996: Localization of grasp representations in humans by positron emission tomography 2. Observation compared with imagination. In: *Experimental Brain Research* 112, pp. 103–111.

Van Gulick, Robert N., 2001: Reduction, Emergence and Other Recent Options on the Mind/Body Problem: A Philosophic Overview. In: *Journal of Consciousness Studies* 8, pp. 1–34.

Gutschmidt, Rico, 2009: *Einheit ohne Fundament. Eine Studie zum Reduktionsproblem in der Physik*. Frankfurt/M: Ontos-Verlag.

Hooker, Clifford, 1981: Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction. In: *Dialogue* 20, pp. 38-59, pp. 201-236, pp. 496–529.

Jacob, Pierre and Jeannerod, M., 2005: The Motor Theory of Social Cognition: A Critique. In: *Trends in Cognitive Sciences* 9, pp. 21–25.

Kauffman, S. A., 1970: Articulation of parts explanation in biology and the rational search for them'. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1970, pp. 257–272.

Keysers, C. and Perrett, D. I., 2004: Demystifying social cognition: a Hebbian perspective. In: *Trends in Cognitive Sciences* 8, pp. 501–507.

Kim, Jaegwon, 2008: Reduction and Reductive Explanation. Is One Possible Without the Other? In: Kallestrup, J.; Hohwy, J. (eds.): *Being Reduced. New Essays on Reduction, Explanation and Causation*. Oxford, UK: Oxford University Press, pp. 93–114.

Kim, Jaegwon, 1993: *Supervenience and Mind. Selected Philosophical Essays.* Cambridge, UK: Cambridge University Press.

Kim, Jaegwon, 1992: Multiple Realization and the Metaphysics of Reduction. In: *Philosophy and Phenomenological Research* 52, pp. 1–26.

Kitcher, Philip, 1989: Explanatory Unification and the Causal Structure of the World. In: Salmon, Wesley; Kitcher, Philip (eds.): *Scientific Explanation*. Minneapolis: University of Minnesota Press, pp. 410–505.

Kitcher, Philip, 1981: Explanatory Unification. In: *Philosophy of Science* 48, pp. 507–531.

Lewis, David, 1972: Psychophysical and Theoretical Identifications. In: *Australasian Journal of Philosophy* 3, pp. 249–258.

Johnson, M. H.; Munakata, Y. & Gilmore, R., 1993: *Brain Development and Cognition*. Oxford, UK: Blackwell Publishing.

Iacoboni, M.,Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., Rizzolatti, G., 1999: Cortical mechanisms of human imitation. In: *Science* 286, pp. 2526–8.

Machamer, P., Darden, L. and Craver, C. F., 2000: Thinking about Mechanisms. In: *Philosophy of Science* 67, pp. 1–25.

Mulligan, Kevin, 2006: Ascent, Propositions and other Formal Objects. In: *Grazer Philosophische Studien* 72, pp. 29–48.

Nagel, Ernest, 1970: Issues in the logic of reductive explanations. In: Kiefer, H. E.; Munitz, M.K. (eds.): *Mind, Science, and History*. Albany: State University of New York Press, 117–137.

Nagel, Ernest, 1961: *The Structure of Science. Problems in the Logic of Explanation*. New York: Harcourt, Brace & World, Inc.

Neurath, Otto, 1959: Sociology and Physicalism. In: Ayer, A.J. (ed.): *Logical Positivism*. New York: Free Press, pp. 282–317.

Place, U., 1960: Materialism as a Scientific Hypothesis. In: *Philosophical Review* 69, pp. 101-104.

Place, U., 1956: Is Consciousness a Brain Process. In: *British Journal of Psychology* 47, pp. 44-50.

Putnam, Hilary, 1967: Psychological Predicates. In: Merrill, D. D.; Capitan, W. H. (eds.): *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, pp. 37–48.

Richardson, Robert C., 1979: Functionalism and Reductionism. *Philosophy of Science* 45, pp. 533–558.

van Riel, Raphael, forthcoming: Nagelian Reduction Beyond the Nagel Model. *Philosophy of Science.*

Rizzolatti, G.; Fadiga, L.; Gallese, V. and Fogassi, L., 1996: Premotor cortex and the recognition of motor actions. In: *Cognitive Brain Research* 3, pp. 131–141.

Rizzolatti, G., Fogassi, L. and Gallese, V., 2000: Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In: Gazzaniga, M.S. (ed.): *The New Cognitive Neurosciences*. Cambridge, MA: MIT Press, pp. 539–552

Sachse, Christian 2007: *Reductionism in the philosophy of science*. Frankfurt/M: Ontos-Verlag.

Salmon, Wesely, 1984: *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schaffner, Kenneth, 1993: *Discovery and Explanation in Biology and Medicine*. Chicago: The University of Chicago Press.

Schaffner, Kenneth, 1976: Reductionism in Biology: Prospects and Problems. In: Michalos, A.; Cohen, R. S. (eds.): *Proceedings of the 1974 meeting of the Philosophy of Science Association* 1974, pp. 613–632.

Schaffner, Kenneth, 1974: The Peripherality of Reductionism in the Development of Molecular Biology. In: *Journal of the History of Biology* 7, pp. 111–139.

Schaffner, Kenneth, 1969: Correspondence Rules. In: *Philosophy of Science* 36, pp. 280–290.

Schaffner, Kenneth, 1967: Approaches to Reduction. In: *Philosophy of Science* 34, pp. 137–147.

Sklar, Lawrence, 1967: Types of inter-theoretic reduction. In: *British Journal for Philosophy of Science* 18, pp. 109–124.

Smart, J., 1978: The Content of Physicalism. In: *Philosophical Quarterly* 28, pp. 339–341.

Smart, J., 1963: Materialism. In: *Journal of Philosophy* 60, pp. 651–662.

Smart, J., 1959: Sensations and Brain Processes. In: *Philosophical Review* 68, pp. 141–156.

Tye, Michael, 1999: Phenomenal Consciousness: the Explanatory Gap as Cognitive Illusion. In: *Mind* 108, pp. 705–725.

Walter, Sven, 2006: Multiple Realizability and Reduction: A Defense of the Disjunctive Move. In: *Metaphysica* 9, pp. 43–65.

Verzeichnis der Autoren

Dr. Markus I. Eronen
Ruhr Universität Bochum
Institut für Philosophie II
Universitätsstraße 150
44780 Bochum
Raum: GA 3/139
Phone: +49-234-32-24724
Fax: +49-234-32-14963
Email: maeronen@uos.de

Prof. Dr. Michael Esfeld
University of Lausanne
Department of Philosophy
Quartier UNIL-Dorigny,
Bâtiment Anthropole 4074
1015 Lausanne
Switzerland
Phone: +41-21-692 29 23
Fax: +41-21-692 30 45
Email: Michael-Andreas.Esfeld@unil.ch

Prof. Dr. Robert Van Gulick
Department of Philosophy, 541 HL
Syracuse University
Syracuse, NY 13244-1170
USA
Phone: +1-15-443-5828
Fax: +1-15-443-5675
Email: rnvangul@syr.edu

Dr. Douglas Kutach
Philosophy Department
Brown University
Box 1918
Providence, RI 02912
USA
Email: douglas_kutach@brown.edu

Prof. Dr. Albert Newen
Ruhr-Universität Bochum
Institut für Philosophie II
Universitätsstr. 150
44780 Bochum

Raphael van Riel
Ruhr-Universität Bochum
Institut für Philosophie II
Universitätsstraße 150
44780 Bochum
Raum GA3/139
Phone: +49-234-32-24724
Fax: +49-234-32-14963
Email: Raphael.vanRiel@rub.de

Dr. Christian Sachse
Department of Philosophy
Bureau 4074,
Section de Philosophie,
Université de Lausanne,
1015 Lausanne
Switzerland
Email: christian.sachse@unil.ch

PHILOSOPHIA NATURALIS

Eingereichte Beiträge dürfen weder schon veröffentlicht worden sein noch gleichzeitig einem anderen Organ angeboten werden. Mit der Annahme des Manuskriptes zur Veröffentlichung in der *Philosophia naturalis* räumt der Autor dem Verlag Vittorio Klostermann das zeitlich und inhaltlich unbeschränkte Nutzungsrecht im Rahmen der Print- und Online-Ausgabe der Zeitschrift ein. Dieses beinhaltet das Recht der Nutzung und Wiedergabe im In- und Ausland in körperlicher und unkörperlicher Form sowie die Befugnis, Dritten die Wiedergabe und Speicherung des Werkes zu gestatten. Der Autor behält jedoch das Recht, nach Ablauf eines Jahres anderen Verlagen eine einfache Abdruckgenehmigung zu erteilen.

Richtlinien zur Manuskriptgestaltung

Bitte jeden Beitrag mit *Titelblatt* abgeben, das folgende Angaben enthält: Name und Vorname des Autors / der Autorin (mit akad. Titel), Titel des Beitrags, vollständige Adresse (inkl. Telefon-Nummer), nähere Bezeichnung der Arbeitsstätte.

Die *Manuskripte* sollten 3-fach und als WORD-File auf Diskette oder CD eingereicht werden und ein deutsch- und englischsprachiges Abstract enthalten. Das Manuskript sollte einen breiten Rand haben.

Der *Umfang* (einschließlich Anmerkungen und Bibliografie) soll bei den Aufsätzen nicht mehr als 30 maschinengeschriebene Seiten (ca. 2.000 Anschläge, 2-zeilig) betragen.

Für *Abbildungen* im Text bitte die Originalvorlage einreichen. Abbildungen müssen numeriert und mit Autorennamen versehen sein.

Zitate im Text sollten vom Haupttext durch eine Leerzeile abgehoben werden. Nach dem zitierten Text stehen Name des zitierten Verfassers, Erscheinungsjahr und Seitenangaben in Klammern, z.B.: (Elkana 1974, S. 34). Bei mehreren Autoren werden die jeweiligen Namen durch Schrägstriche getrennt, z.B.: Krantz/Luce/Suppes/Tversky 1971, S. 8). Wird auf mehrere Publikationen desselben Autors im selben Erscheinungsjahr verwiesen, so sollen sie numeriert werden: (Ludwig 1970 a) bzw. (Ludwig 1970 b).

Die *Anmerkungen* sind im Manuskript fortlaufend zu numerieren; sie stehen am Schluss des Beitrags in numerischer Reihenfolge.

Für das anschließende *Literaturverzeichnis* in alphabetischer und chronologischer Reihenfolge gilt folgendes Muster:

Elkana, Y., 1974: *The Discovery of the Conservation of Energy*. London: Huchinson.

Clausius, R., 1850: Über die bewegende Kraft der Wärme. In: *Annalen der Physik und Chemie*, 79, S. 500–524.

Klein, M.J., 1978: The Early Papers of J. Willard Gibbs: A Transformation of Thermodynamics. In: E.G. Forbes (Hg.): *Human Implications of Scientific Advance*. Edinburgh: University Press, S. 330–341.

Korrekturen: Die Autoren erhalten vom Verlag die Fahnen ihres Beitrags mit der Bitte, die korrigierten Fahnen *innerhalb von zwei Wochen* an den Herausgeber zu schicken. In den Fahnen sollen nur noch Satzfehler berichtigt werden.

Nach Erscheinen des Heftes erhalten die Autoren 3 Belegexemplare des jeweiligen Heftes.

Located at the crossroads between natural philosophy, the theory and history of science, and the philosophy of technology, *philosophia naturalis* Journal for the Philosophy of Nature has represented for many decades – not only in the German speaking countries but internationally – a broad range of topics not addressed by any other periodical.
The journal has a highly interdisciplinary focus. Articles with systematic as well as historical approaches are published in German and English. Their quality is assured by a strict peer review policy.

Inhaltlich an der Schnittstelle zwischen Naturphilosophie, Wissenschaftstheorie, Wissenschaftsgeschichte und Technik-Philosophie angesiedelt, vertritt die Zeitschrift *philosophia naturalis* Journal for the Philosophy of Nature seit mehreren Jahrzehnten nicht nur im deutschen Sprachraum, sondern auch im internationalen Vergleich, einen weiten Themenbereich, der von keinem anderen Publikationsorgan vertreten wird.
Die Zeitschrift ist ausgesprochen interdisziplinär ausgerichtet. Sie veröffentlicht Aufsätze in deutscher und englischer Sprache, die sowohl systematisch als auch historisch orientiert sind. Deren Qualität wird durch ein besonders strenges Begutachtungsverfahren gesichert.

ISSN 0031-8027 www.klostermann.de
ISBN 978-3-465-04126-9