Université de Montréal

# Strategies of Empirical Justification in Experimental Science

Par

**Tudor Baetu**

Département de philosophie
Faculté des arts et des sciences
Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de doctorat
en philosophie
Mai, 2008

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Strategies of Empirical Justification in Experimental Science

présenté par :

Tudor Baetu

a été évalué(e) par un jury composé des personnes suivantes :

MARQUIS, JEAN-PIERRE
président-rapporteur

GANTHIER, YVON
directeur de recherche

LAURIER, DANIEL
membre du jury

ZEMAN, VLADIMIR
examinateur externe

.....
représentant du doyen de la FES

**Résumé :**

Le but de cette thèse est de défendre la conception hypothético-déductive de la pratique scientifique. Cette thèse traite de problèmes fondamentaux concernant la formulation et la confirmation des hypothèses explicatives.

En règle générale, les hypothèses scientifiques sont déterminées à la fois par les données empiriques disponibles et par des considérations théoriques. La première partie de la thèse traite des composantes du donné empirique. L'analyse d'exemples tirés de la psychologie cognitive, la microbiologie et l'immunologie me permet de conclure qu'il est possible d'établir un réseau de corrélations expérimentales sans l'aide d'une structure interprétative surajoutée. Ce savoir purement expérimental garantit la coréférence de diverses interprétations théoriques, et, par conséquent, une continuité extra-théorique ou infra-théorique du savoir scientifique.

La deuxième partie traite des composantes théoriques dans la formulation des hypothèses. La théorie gravitationnelle de Newton est utilisée comme étude de cas afin d'investiguer le holisme et les problèmes qu'il pose, plus précisément en ce qui concerne les hypothèses auxiliaires et l'idéalisation dans leur rapport avec la confirmation, la falsification et la méthode hypothético-déductive en général.

Dans la troisième partie de la thèse j'explore l'interaction entre les ingrédients empiriques et théoriques des hypothèses scientifiques. La transition de la génétique mendélienne à la génétique classique, caractérisée par l'interprétation physique d'une explication conceptuelle, me permet de discuter certains aspects du réalisme sémantique. La transition de la génétique classique à la biologie moléculaire est à l'origine d'un débat sur la continuité théorique du savoir scientifique; par rapport à ce débat, je conclus que l'élucidation des mécanismes moléculaires contribue d'une façon significative à l'explication génétique, et que, malgré la complexité du problème, il est possible et nécessaire de redéfinir la notion classique de gène en termes de structures moléculaires. Dans le dernier chapitre, je tente de montrer comment différents modèles explicatifs peuvent être combinés afin de générer de nouvelles hypothèses.

**Mots clé :**
confirmation, épistémologie, explication, idéalisation, mécanisme, méthode hypothético-déductive, philosophie de la biologie, philosophie des sciences

**Abstract:**

The primary purpose of this thesis is to defend and update the hypothetico-deductive account of the scientific practice. It treats of fundamental issues concerning the formulation and confirmation of explanatory hypotheses.

Typically, scientific hypotheses are constrained by both available empirical data and more general patterns of theoretical explanation. Accordingly, Part I deals with empirical constrains on hypothesis formation. Based on study cases drawn from cognitive psychology, microbiology and immunology, I argue that it is possible to establish a web of empirical and experimental correlations in the absence of higher-level theoretical interpretations. Experimental knowledge ensures that different theoretical interpretations can continue to corefer to the same body of experimental data, thereby granting an extra-theoretical continuity of scientific knowledge. Also an attempt is made to revivify the observable – theoretical distinction by analysing actual examples of experimental knowledge and how this knowledge constrains higher-level theoretical interpretations in the context of hypothetico-deductive approaches to scientific methodology.

In Part II, I concentrate on the theoretical constraints shaping the formulation of scientific hypotheses. Newton's gravitational model is used as a study case to investigate holism and the problems it entails, most notably issues concerning background assumptions and idealisations, in respect to confirmation, falsification and the hypothetico-deductive method in general.

In Part III I explore the interplay between the empirical and theoretical constraints shaping the development of genetic explanations. The transition from Mendelian genetics to classical genetics is characterised by the physical interpretation of conceptual explanation and therefore provides an excellent opportunity for discussing issues related to semantic realism. The subsequent transition from classical genetics to molecular biology is at the origin of a rather heated debate concerning the theoretical continuity of scientific knowledge. I argue that the elucidation of the mechanisms underlying the properties of genes hypothesised by classical genetics contributes directly to the genetic explanation and that, despite the complexity, it is possible and necessary to redefine the classical 'gene' in molecular terms. In the final chapter, I show how different explanatory models combine together in order to yield new hypotheses and open new avenues of research.

**Key words:**
confirmation, epistemology, explanation, hypothetico-deductive method, idealisation, mechanism, philosophy of biology, philosophy of science

## Contents

## Table of Figures

## *Introduction*

The primary purpose of this thesis is to defend and illustrate the use of the hypothetico-deductive method in the scientific practice. It provides an introduction to some of the philosophical issues concerning the formulation and confirmation of explanatory hypotheses.

The rise of and subsequent demise logical positivism had a most peculiar effect on contemporary philosophy of science. By the end of the 19$^{th}$ century, the basic requirement of empirical investigation and empirical confirmation led Mach to believe that scientific theories are summaries of experimental data constructed for the purpose of organising data into a readily comprehensible format. He held that scientists should restrain from appealing to unobservables, and that, in the rare cases when it proved useful to introduce them, they should be allowed a purely instrument role. In the 1920s, logical positivists pushed matters a step further and proposed a rigorous formal programme aiming to reduce all assertions about unobservables to assertions about observables. According to positivist accounts, there are no such things as explanatory hypotheses. To be more precise, the semantic content of hypotheses is deemed to be completely

determined by available empirical data, while the excess theoretical content assumes the role of a form whose sole purpose is to organise the empirical content.

The positivist project failed. Three decades later, the demise of the logical positivist programme was followed by the rise of realism about unobservables. Explanatory hypotheses postulating the existence of entities, mechanisms and processes were once again reinstated – not in the scientific practice, which never ceased making use of them, but rather in philosophical accounts of the scientific practice. However, scientific realism didn't reign for long. The social revolutions of the 60s and 70s intruded philosophy of science by favouring a socio-historical understanding of science according to which the excess theoretical content of explanatory hypotheses is, for the most part, a social construction.

Today, philosophy of science is marked by a heated debate between realists and antirealists about the semantic, epistemic and ontological status of the unobservables postulated by scientific explanations. Realists and antirealists disagree on whether unobservables refer to something existing outside scientific explanations or whether they are mere human constructs. With very few exceptions, both camps agree that unobservables are indispensable, that they transcend the realm of direct empirical confirmation and that, one way or another, every statement, even a purely perceptual one, is theory-laden. Explanatory hypotheses are nowadays thought to be primarily theoretical in nature, to the extent that the autonomy of the experimental practice in science has become a notion so outlandish that nobody cares to defend it anymore in philosophical circles.

As a former scientist, I firmly believe that scientific hypotheses are always constrained by both available empirical data and more general patterns of theoretical explanation. Empirical investigation and experimentation are just as important today as they were a century ago, if not in

physics, then at least in newly emerging sciences, such as biology and psychology. The point I want to make is that philosophy of science may have changed a lot, but core scientific practice didn't. This is not to say that Mach was right. Quite on the contrary, I think that the reaction to positivism is justified. I think however it is an extreme reaction to an extreme programme. Even if science, as a whole, does not reduce to the experimental practice and the knowledge it generates, it is nevertheless the case that experimentation, and empirical investigation in general, have a well established place in science. Accordingly, my first concern is to reaffirm the autonomy of experimental knowledge. Part I deals with empirical constrains on hypothesis formation. An attempt is made to revivify the observable – theoretical distinction by analysing examples of experimental knowledge and by showing how this knowledge constrains higher-level theoretical interpretations in the context of hypothetico-deductive approaches to scientific methodology. I begin by arguing that the introduction and subsequent use of theoretical terms should remain contingent upon the satisfaction of certain observational conditions in order to avoid a needless proliferation of hypotheses. In chapter 2, I further argue that such a requirement can also provide a means to fix reference by correlating theoretical terms to a set of laboratory procedures and protocols. Experimental knowledge ensures that different theoretical interpretations can continue to corefer to the same body of experimental data, thereby granting an extra-theoretical continuity of scientific knowledge. In chapter 3, I defend the observable – theoretical distinction against theory-ladenness arguments. Against social constructionism, I argue that even though some aspects of perception are relative to the empirical experience and the social background of the researcher, it does not follow that access to empirical reality is exclusively mediated via learning mechanisms and that learned perception, such as categorisation, is exclusively a matter of theoretical interpretation. I also challenge the Duhemian view according to which strictly observa-

tional data pertains to a lesser form of knowledge proper to younger sciences by discussing the example of Gram staining in microbiology. I conclude that theory-free experimental investigation of empirical phenomena is not only possible, but also a fruitful and quite common way of conducting science, especially when science is motivated by pressing practical concerns. Based on an example from molecular research in immunology, I further show in chapter 4 that it is possible to establish an extensive web of empirical and experimental correlations in the absence of any significant higher-level theoretical interpretation of the experimental techniques. This shows that it is conceivable that experimental data posits an initial (or 'input') set of constraints on the formulation higher level theoretical interpretations, as opposed to only theory positing a constraint on the interpretation of empirical data. Given the conclusions reached in chapters 3 and 4, I argue in the final chapter of the section that while individual pieces of data can be interpreted in a variety of ways, more extensive webs of empirical and experimental correlations posit more stringent constraints on the number of possible interpretations consistent with data and techniques proper to several domains of investigation. By limiting the number of possible interpretations, extensive webs of experimental data can limit the number of possible explanatory hypotheses and thus give some weight to claims to abduction or inference to the best explanation.

In Part II, I concentrate on the theoretical constraints shaping the formulation of scientific hypotheses. The underlying assumption is that hypotheses are tied down to experimental data, and, at the same time, they are also connected to more general patterns of explanation (which are, in part, responsible for the interpretation of the empirical data). Chapter 6 provides a brief overview of the most common accounts of explanation available in the philosophical literature: the deductive-nomological, the causal and causal-mechanistic, the statistical relevance, the conceptual unification and the manipulationist accounts. In chapter 7, I establish a connection be-

tween deductive-nomological explanations and the theory-model distinction in science. Explanatory models are treated as attempts to extend the domain of application of pre-existing explanations, which count as theories, to new phenomena and domains of investigation. Since it is not always possible to derive the empirical laws describing the target phenomenon from the laws of the theory alone, I argue that deductive explanations must consist of 'fundamental hypotheses', derived from the theory, and auxiliary assumptions counting as 'model-specific hypotheses'. However, since model-specific hypotheses are required in order to derive the target empirical law from the fundamental hypothesis, confirmation (or corroboration) becomes holistic. In the remaining chapters of the section, Newton's gravitational model is used as a study case to investigate holism and the problems it entails, most notably issues concerning background assumptions and idealisations. In chapter 8, I provide evidence that confirmation holism does not necessarily entail that theories cannot be falsified. For example, if the inverse square law is verified on an individual basis, then it would take classical mechanics to be wrong about planetary motion in order to obtain a true conjunction of the principles of classical mechanics as established by local experiments, Newton's inverse law, the right distribution of massive objects and the presence of frictional forces, as demonstrated by the existence of a significantly dense inter-planetary medium. If such a conjunction of observations were ever shown to be true empirically, it would mean that the law of addition of forces holds true on Earth, but not in Heavens. Then the fundamental hypothesis would be explicitly and unambiguously falsified, classical mechanics shown to be false about planetary motion and, as an immediate and unavoidable consequence, all mechanical models for planetary motion falsified. Using a similar strategy, I argue in chapter 9 that some idealisations can be treated as 'model-consistent hypotheses', that is, as consequences deductively granted by the model if certain initial conditions hold true. The end result is that in-

stead of having to justify the assumptions themselves as additional, independent propositions required for the derivation – and therefore external to the theory –, it is only required to justify the conditions under which they follow as certain limit cases of the fundamental hypotheses used to formulate the model.

In Part III I explore the interplay between the empirical and theoretical constraints shaping the development of genetic explanations. The notions of semantic and epistemological realism are introduced in chapter 10. Instrumentalists argue that scientific theories and models serve solely pragmatic purposes and should not be taken literally, but rather as means to summarise and organise empirical data, for instance, as reliable methods or computational algorithms for predicting phenomena. In contrast, semantic realists believe that the entities, mechanisms and structures hypothesised by scientific theories and models underlie and determine the manifestation of empirical reality at the observable level. The transition from Mendelian genetics to classical genetics is characterised by the physical interpretation of conceptual explanation and therefore provides an excellent opportunity for discussing issues related to semantic realism. In chapter 11 I argue that even though some early geneticists seem to have adopted an instrumentalist stance, further research in the field of genetics was motivated by adopting a realist interpretation of the genetic explanations. The downside of instrumentalism is that, once an empirically adequate description is provided, research has no logical reason to continue. In contrast, to ask "How genes are inherited?" or "How genes determine phenotypes?" implies that there is more to be said, that a 'mechanism' of some sort must be specified, in short, that a an instrumental or 'black-box' use of the genetic explanation is unsatisfactory or incomplete. A realist interpretation of Mendel's talk about 'genetic elements' and 'alleles' made possible the chromosomal explanation of linkage and recombination. The subsequent transition from classical genetics to molecu-

lar biology is at the origin of a debate concerning the theoretical continuity of scientific knowledge. In chapter 12, I argue that a progression towards truth can be achieved when some parts of the theory acquire an empirical significance they didn't have before. In the case of genetics, the elucidation of the mechanisms underlying the properties of genes hypothesised by classical genetics contributes to the genetic explanation by taking into account additional empirical constraints, such as knowledge about the chemical makeup of the genes. For example, differences in the chromosomal makeup of an organism correlate with differences in phenotype; furthermore, experiments showed that any interference with the chromosomal makeup of an organism leads to radical changes in phenotype. This data served at the time as evidence for the existence of 'genetic elements' and as a further empirical constraint on future genetic theories: future explanations must take into account the fact that the fate of genes and their ability to determine the phenotype is tightly linked to the fate of chromosomes. In chapter 13, I attempt to show that, despite the complexity, it is possible to redefine the classical 'gene' in molecular terms.

Epistemological realists believe that the inferences and justification methods employed by scientists are reliable and tend to yield the truth. On the antirealist end of the debate, it is argued that even if theoretical terms refer, scientists cannot legitimately claim that they are in the possession of truth. Van Fraassen argues that talk about the experimental properties of an object considered outside the experimental context responsible for rendering these properties manifest is bound to refer to counterfactual states of affairs, meaning that experimental data does not contribute to the empirical adequacy of explanations concerning the actual state of the object. I address this objection in the final chapter of my thesis by arguing that there is a variable overlap between 'naturally-occurring situations' and 'experimental setups' that doesn't fit very well van Fraassen's rigid actual-counterfactual distinction between experimental and passive observation

data. Furthermore, since data obtained in a particular experimental setup is typically not thought to contribute directly to the empirical adequacy of explanations about how a phenomenon occurs in natural conditions, the data obtained in the lab is not used counterfactually. For example, data obtained *in vitro* and/or on cell lines is not thought to contribute directly to the empirical adequacy of models about *in vivo* primary cells. Rather, studies on cell lines contribute to the understanding of the functioning of the human body indirectly, by opening the possibility to formulate further hypotheses and more complex models. Based on an example drawn from my own research in molecular oncology, I argue that different explanatory models combine together in order to yield new hypotheses and open new avenues of research. Confirmation of individual models is only partial and the fact that some elements of the model are confirmed provides insufficient grounds for inferring that the model is true, most likely true or true to a high degree of approximation. Things change for the better when, instead of having to make a judgement about the truth of a single model, it becomes possible to assess the truth of a collection of models cross-referencing each other. Once several partially confirmed models combine together in order to yield new confirmed predictions, the conjunction of the models receives a higher degree of confirmation than each individual model.

# PART I

# EXPERIMENTAL CONSTRAINTS ON HYPOTHESIS FORMATION

# CHAPTER 1

# THE HYPOTHETICO-DEDUCTIVE METHOD: A GENERAL FRAMEWORK FOR UNDERSTANDING SCIENTIFIC REASONING

## *1.1 The Hypothetico-Deductive Method and Its Immediate Relative, Falsificationism*

One of the few philosophical accounts of scientific reasoning that captured the attention of practicing scientists is the so-called 'hypothetico-deductive method' (*HD*). Already present in the writings of Newton and Descartes, the account gained widespread popularity in the second half to the 20th century due to the efforts of Popper, whom is usually credited for it.

It is important however to realise that the version of the *HD* account endorsed by scientists, that is, the version found in the introduction of most science textbooks, diverges from Popper's falsificationism initially responsible for popularising the *HD* account in scientific circles. The former is officially defined as the method of proposing hypotheses and testing their accept-

ability by determining whether their logical consequences are consistent with observed data[1], and allows, at least in principle, for induction, inference and confirmation. In contrast, the latter accepts the validity of the Humean critique of induction and explicitly refutes the notion of confirmation (Popper, 1959 p. 315).

### *1.2 Falsificationism, in Principle and in Practice*

Lakatos points out that falsificationism can be understood both epistemologically and methodologically. For the time being, I will concentrate on the epistemological aspect. According to epistemological falsificationism (Popper, 1959; 1965), a hypothesis ($H$) is first conjectured – 'out of nowhere' so to speak, for theories cannot be inductively inferred from experience – then observational consequences ($E$), counting as predictions entailed by the hypothesis are derived as dictated by the standard rules of inference ($H \rightarrow E$). Popper takes $H$ to be a universal law or statement, that is, a general proposition from which a more particular proposition $E$ can be derived. However, since Hume's critique of induction, it is generally acknowledged that a universal statement cannot be empirically justified. The problem is twofold. Not only it is impossible to prove the truth of a universal proposition by confirming a few cases in which the proposition holds true, but, quite often, the domain of reference of the universal proposition is not explicitly defined, meaning that we can never know whether we covered all the cases subsumed under the universal. The same comment applies to laws and theories, the underlying reasoning being that a law or theory can never be proved to hold universally and cannot be shown to be true about all phenomena, entities or situations of a given type.

---

[1] Encyclopaedia Britannica defines the hypothetico-deductive method as the "*procedure for the construction of a scientific theory that will account for results obtained through direct observation and experimentation and that will, through inference, predict further effects that can then be verified or disproved by empirical evidence derived from other experiments*".

Given this difficulty, Popper adopts a strong anti-inductivism according to which the occurrence of $E$ cannot possibly prove $H$. Then how is scientific knowledge justified? The following quote summarises Popper's solution to the problem:

> "[W]*e seek a decision as regards these (and other)* [theory-] *derived statements by comparing them with the results of practical applications and experiments. If this decision is positive, that is, if the singular conclusions turn out to be acceptable, or verified, then the theory has, for the time being, passed its test: we have found no reason to discard it. But if the decision is negative, or in other words, if the conclusions have been falsified, then their falsification also falsifies the theory from which they were logically deduced.*
>
> *It should be noted that a positive decision can only temporarily support the theory, for subsequent negative decisions may always overthrow it. So long as a theory withstands detailed and severe tests and is not superseded by another theory in the course of scientific progress, we may say it has 'proved its mettle' or that it is corroborated by past experience.*
>
> *Nothing resembling inductive logic appears in the procedure here outlined. I never assume that we can argue from the truth of singular statements to the truth of theories. I never assume that by force of 'verified' conclusions, theories can be established as 'true', or ever as merely 'probable'* ".
>
> (1959 p. 33)

In sum, if $E$ occurs as predicted, $H$ is merely 'corroborated', allocution which simply states that $H$ survived the attempts to falsify it, hence Popper's claim that experience cannot determine scientific knowledge by telling us which theories are true, but only delimit it by showing which theories are false.

The epistemology of falsificationism rests entirely on *modus tollens*:

$$((P \rightarrow Q) \wedge \neg Q) \vdash \neg P$$

The logical schema is simple, yet this does not mean that the falsification procedure is always a simple operation. Granted, in principle, the falsification of a single observational consequence suffices to overthrow the more general statement, law or theory from which it is derived ($\mathfrak{N}E \rightarrow \mathfrak{N}H$). In practice however, given the possibility of observational error, it is virtually never the case that a single falsifying instance suffices to overthrow a general law or theory.

Several authors further pointed out that Popper's 'universal law → basic [i.e., singular/observational] statement' strategy for deriving predictions needs to be revised. Quine and Grünbaum challenged Popper's account on the grounds that laws of nature are non-existential ("All S are P" is to be understood along the lines "Whatever is S is also P") and therefore cannot deductively entail basic statements, which refer to the occurrence of specific phenomena at specific locations in space-time. Thus understood, a universal law can imply predictions about singular phenomena only in conjunction with a set of further statements specifying 'the initial conditions' or 'parameters' of the system under investigation (Wedeking, 1976). In fact, most of the time, predictions are not derived from the statement under test, but from a conjunction of general statements plus a set of 'initial conditions' plus a set of additional propositions needed for the derivation (Putnam, 1991). As Popper himself eventually recognized, while testing a particular statement, scientists must often assume that a whole set of 'background assumptions' holds true. Thus, although simple in principle, in practice, falsification often turns out to be exceedingly complicated as there is always a worry that the falsifying/corroborating observation is mistaken, or that the assumed background knowledge is faulty or defective (Popper, 1976).

Set aside these complications pertaining to the falsification procedure, which I will address in the second part of this book, there are other, more immediately obvious shortcomings of falsificationism as a general approach to science. First, unlike most versions of *HD* available in

science textbooks, falsificationism does not allow for corroboration to function as some form of empirical justification, weaker than logical proof, but stronger than mere 'absence of falsification'. And second, it does not tell us how hypotheses are formed and what their relationship to experimental data is prior to the epistemological justification of their observational consequences. The main purpose of the present chapter is to show that, in the actual scientific practice, induction to low-level generalisations is a common procedure and that the formulation of higher-level hypotheses is often constrained by lower-level empirical generalisations and experimental correlations.

### *1.3 The Lessons of Confirmationism*

In order to better understand falsificationism as an epistemological thesis, it is profitable to contrast it with its immediate competitor, the verificationism promoted by logical empiricism. According to logical positivism a scientific theory is

1) formulated in a first order mathematical language comprising the five standard truth functions (□, □, →, ↔, ℜ), the two standard quantifiers (□, □) and an identity sign (=) required to express co-reference; the language comprises
2) a logical vocabulary comprising logical constants and mathematical terms;
3) an observational vocabulary $V_O$ whose terms refer directly to observables entities, properties, events, etc.;
4) a theoretical vocabulary $V_T$; and (5) explicit definitions of $V_T$ in terms of $V_O$, or correspondence rules (C-rules) having the form $\Box x(Tx \leftrightarrow Ox)$, where $T$ is a theoretical term (t-

term) and *Ox* contains solely observational terms (o-terms) and logical vocabulary (Suppe, 1977 pp. 16-17).

In sum, the positivist picture of scientific knowledge looks as follows:

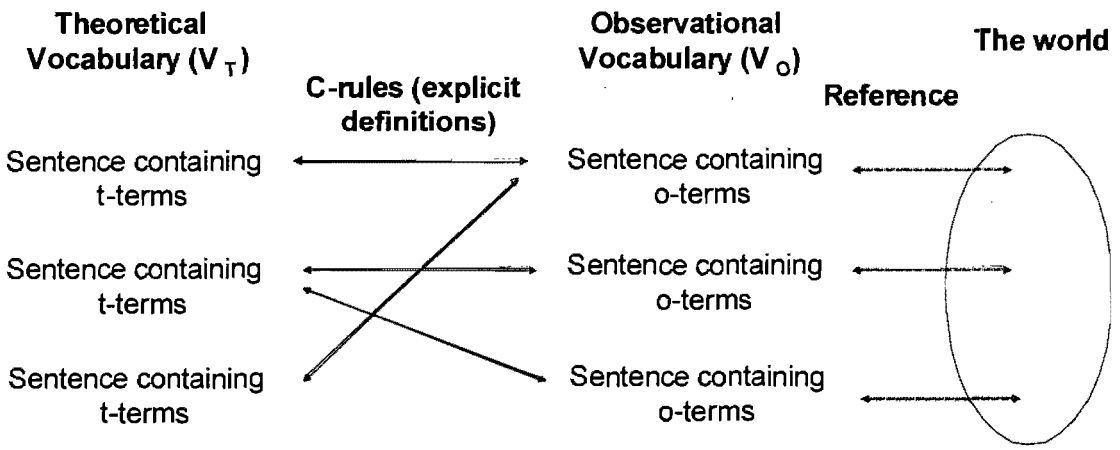



**Figure 1. Verificationism**

How does this compare with *HD* falsificationism? Presumably, the hypothesis *H* introduces t-terms, while the prediction *E* contains only o-terms. At the very least, *H* involves a low-level generalisation, such as the extrapolation of an empirically observed co-variation between two observables for values beyond the reach of actual measurements or observations, or for moments situated in a remote past or future, etc. Since a low-level generalisation extends a correlation beyond current observations and potentially trespasses into the realm of the unobservable, the extrapolation may be viewed as a t-term.

Carnap's verificationism requires that all t-terms are explicitly defined via o-terms (Carnap, 1928). The t-term *T* applies to or is a property of some entity or phenomenon *x* if and only if, under a test condition *C*, *x* displays observable property *O* ($Tx \leftrightarrow (Cx \rightarrow Ox)$). For example, object *x* is said to have a certain temperature if and only if, once put in contact with a thermometer, a certain change in the height of the red marker is observed. It follows from here that *T* (in this case, temperature) is equivalent and can be replaced without any loss of meaning by $C \rightarrow O$

(the operation whereby the temperature of the object is read from a thermometer). Verificationism clearly states that the meaning of t-terms is completely exhausted by various explicit definitions involving only o-terms. T-terms have no 'excess content' and as such they are always dispensable. In the end, it is best to think of them as handy abbreviations for observational assertions.

From the standpoint of the *HD* method, it is immediately obvious that verificationism severely restricts the kind of hypotheses we can form and, as it has been acknowledged by Carnap himself (1939), it is not possible to account for the current scientific theories relying solely on the very narrow basis of verificationism. Consider, for example, the difference between the actual observational correlates $\{(T_1, l_1), (T_2, l_2) \ldots (T_n, l_n)\}$ and the empirical law $l = kT$, where $l$ is the observed length of a metal rod and $T$ is the temperature of the rod. Not only the law extrapolates the correlation for arbitrarily high or low temperatures, but it also postulates a continuous spectrum of temperatures which transcends the resolution of actual and currently possible measurements.

Historically, the fatal objection to verificationism stemmed from an inability to account for the so-called 'dispositional properties'. In reference to the above mentioned example, it has been argued, quite evidently with good reason, that an object has a temperature even if it is not measured by a thermometer and that it is possible to talk meaningfully about the temperature of an object independently of the experimental operation giving the measure of temperature. This strongly suggests that most t-terms cannot be introduced via explicit definitions compounding o-terms.

A subsequent version of verificationism, which we may call 'confirmationism', was later proposed in order to account for this difficulty (Carnap, 1936). According to this revised ac-

count, the conditions $(C \square O) \rightarrow T$ and $(C \square \Re O) \rightarrow \Re T$ tell us that $T$ applies to all things that satisfy the condition $C \square O$ and does not apply to any of the things which satisfy $\Re(C \square \Re O)$. Unlike the explicit definition $T \leftrightarrow (C \rightarrow O)$, the tandem of conditionals doesn't tell us what $T$ is, nor it reduces $T$ to a measurement of $T$, but merely correlates $T$ with the experimental conditional $C \rightarrow O$ (the measurement of its magnitude, for example), where the latter can be considered somewhere along the lines of an 'observational trace' of the theoretical term. To use a better suited example, talk about a beam of electrons ($T$) is justified by the presence of droplets ($O$) in a condensation chamber following a certain experimental procedure ($C$); conversely, absence of droplets ($\Re O$) given the same experimental procedure ($C$) renders talk about electrons illegitimate. Since the meaning of $T$ is not exhausted, confirmationism is compatible with semantic realism, that is, with the thesis that t-terms refer to ontological items beyond those justified by observational sentences (Feigl, 1950). Nevertheless, Carnap and, following him, Nagel (1950) took t-terms and sentences containing them to be primarily instrumental, that is, devoid of any referent outside the theory and pertaining exclusively to the internal functioning of the theory. A consensus among logical positivists was however achieved, namely that t-terms are no longer eliminable (Hempel, 1950; 1963).

Confirmationism is compatible with falsificationism since it allows $H \rightarrow E$ but not $E \rightarrow H$. However they do differ since, under falsificationism, it is not required to justify the introduction of t-terms, but only their eventual elimination following falsification. In other words, according to Popperian falsificationism, hypotheses are constrained exclusively at the level of the 'output' predictions, while their 'input' introduction remains completely free of restriction. The natural upshot of this freedom is that, in contrast with confirmationism, which requires that the introduction and subsequent use of t-terms remains contingent upon the satisfaction of certain

observational conditions, under falsificationism, t-terms can enjoy a massive proliferation hindered only by the limits of a scientist's ability to imagine new hypotheses. The falsificationist strategy is therefore to give a chance to every hypothesis, no matter how far-fetched, and hope that among the many contenders only few will survive the constant test of falsification.

Falsificationism and confirmationism compare as follows:

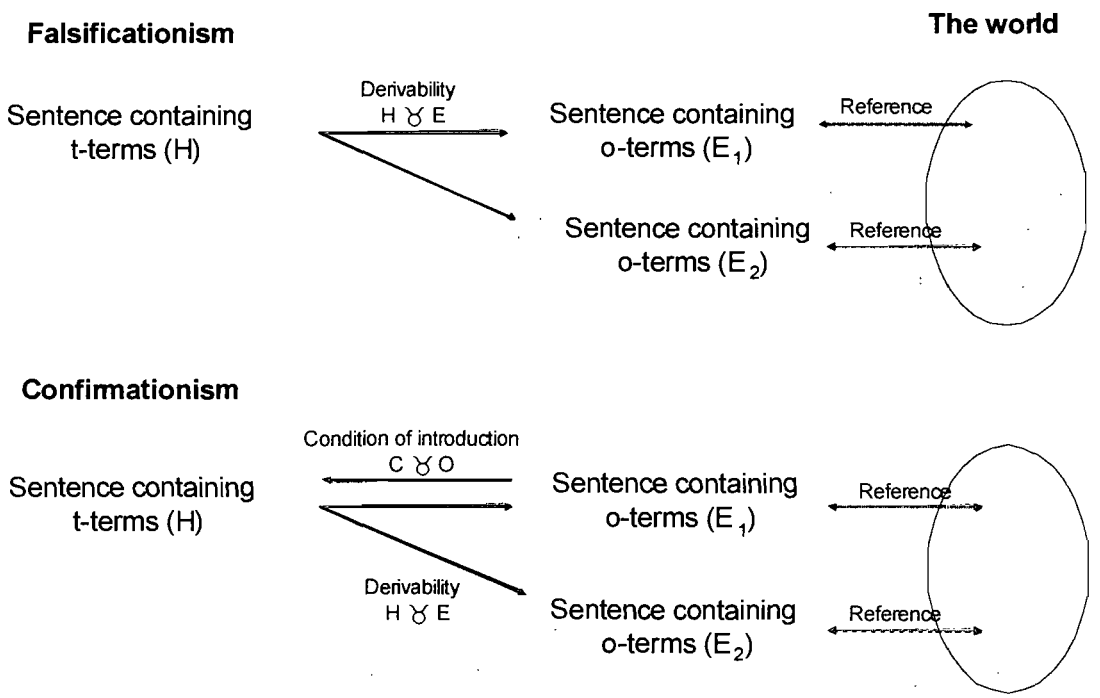



**Figure 2. Falsificationism and Confirmationism Compared**

According to falsificationism, if $E_1$ or $E_2$ or both are falsified, $H$ is falsified. There are no conditions constraining the introduction of $H$. In contrast, confirmationism allows the introduction of $H$, but only on condition that $C \rightarrow O$ is satisfied. Once $H$, which does not reduce to $E_1$, is introduced, $E_2$ can be derived. However, if $C \rightarrow \Re O$, then $H$ cannot be introduced and, as a consequence, it becomes impossible to derive $E_2$.[2]

[2] The use of Ramsey sentences (1929) may provide an increased level of flexibility. The basic idea is to convert t-terms, from constants, into variables. Braithwaite (1953) further proposed to introduce t-terms as properties about which higher level propositions are true and from which lower-level propositions concerning observables can be

### *1.4 Lower-Level ('Input') Empirical Constraints on Hypothesis Formation*

By combining confirmationism with falsificationism the basic structure of *HD* reasoning is preserved while taking into account the fact that, often times, lower-level constraints guide the formulation of higher-level hypotheses.

Since many theories, models and explanations in general are designed in such a way that they entail an already established set of empirical laws, it follows that many scientific hypotheses are from the very beginning empirically adequate in respect to some preliminary data. This posits a challenge to falsificationism. By failing to distinguish between the empirical constraints a theory satisfies by design and the further predictions which it must entail for the purposes of epistemological justification, one might erroneously conclude that the theory cannot be falsified. For example, in Figure 2, $E_1$ is empirical evidence incorporated in the hypothesis $H$, and as such bears no relevance to the issue of falsification and corroboration; in contrast, $E_2$ counts as new consequence entailed by $H$ and can be used in order to justify the survival or downfall of $H$. This revised version of *HD* ensures that most scientific hypotheses are empirically adequate in at least one respect.

---

derived. Thus, instead of enumerating all t-terms that fit a certain description, it is possible to create a class with an indefinite number of members which contains all possible values $x$ that fit the required description, some of which we can name, some of which we haven't thought about yet. For example, we may say that there is an $x$ such that $x$ causes droplets to form in a condensation chamber following a certain experimental procedure ($E_1$). $x$ may be an electron, or it may be a proton or a great deal of many other particles. This description fits many hypotheses. If $H$ is the hypothesis that $x$ is an electron, then $E_1$ (trivial) and $E_2$ (non-trivial) follow. If $E_2$ (say, the beam of particles is deflected in a certain way in the presence of a magnetic field) is falsified, then, according to falsificationism, we have to reject $H$. We are then free to posit another hypothesis, yet not just any hypothesis, but one from which $E_1$ can be derived.

# CHAPTER 2

# OPERATIONALLY DEFINED COREFERENCE

## *2.1 Operationalism and Reference*

Upgrading the *HD* method with some elements of confirmationism provides a second advantage: it fixes reference. Logical positivism allows theoretical concepts to be defined operationally, that is, to be equated or correlated to a set of laboratory procedures and protocols (Bridgman, 1927). Many authors are quick to note that Bridgman's operationalism was short lived. The approach certainly didn't live up to the philosophical expectations of verificationism, but to claim that it is dead is, to put it euphemistically, a bit of an overstatement. As Klee (1997 pp. 53-54) and many other authors (Stevens, 1963; Hardcastle, 1995) point out, operationalism is still very much alive in psychology and, I would add, in biological sciences as well.

Why is this so? Here is a quick example. In psychology, it is imperative that certain properties can be ascertained of a subject. For instance, one may claim that depressed people evaluate causal correlations between actions and their alleged effects more accurately than normal, un-

depressed people, who tend to overestimate their control over the desired effect. This surprising finding is known in the psychological literature as 'depressive realism' (Dobson, et al., 1989). 'Depressive realism' seems to support a higher level theory which establishes a link between depression and cognition, thus giving a further theoretical meaning to the term 'depression'. Nevertheless, talk about the cognitive implications of depression is pointless if it cannot be tied down to some previous description of depression. Most notably, for any studies in this field to be reliable, it is absolutely essential that the subjects tested are normal or depressive the same way. Thus, before starting their experiments, all researchers must administrate the same depression test to their subjects, or at very least tests that consistently yield highly similar results.

The test measures the symptoms of depression. This is not to say that depression is what a depression test measures, as Boring (1923) might have put it. If this were the case, depression would be analytically defined as 'what a depression test measures', and therefore unrevisable. As Boyd points out, this conclusion is incompatible with the fact that tests and lab protocols are constantly revised (Boyd, 1985); hence, a strict verificationist reading of operationalism is to be rejected. Confirmationism is more flexible since it merely ties down t-terms to lab protocols without reducing them to the latter (Carnap, 1936). On one hand, depression has many further implications, many of which remain quite puzzling, such as 'depressive realism'. On the other hand, failure to diagnose depression in a uniform manner undermines the common reference of various studies in this domain of investigation, since it would not be clear whether these studies refer to the same 'normal' and 'depressive' subjects. A 'soft' version of operationalism, compatible with confirmationism and devoid of any anti-metaphysical ambitions, solves the problem.

### *2.2 Descriptive and Causal Accounts of Reference*

The above use of psychological tests illustrates a fundamental aspect of scientific methodology in general. Any paper published in a respectable scientific journal contains a 'Methodology and Protocols' section. Metaphorically, it is a finger pointing to a set of observations: 'Follow this list of procedures and you shall obtain exactly the same observable results as I, the researcher who wrote this paper, did.' It is not question of interpreting data. A machine could do the operations and, if the laboratory protocols used are any good, they should invariably yield the same observations. For example, given cells issued from the same cell line or extracted the same way, stimulated in the same way with the same concentrations of the same chemical, harvested the same way, lysed the same way, centrifuged the same way, etc., the same bands on an electrophoresis gel are observed. Just as in the case of a depression test, the reproducibility of a set of observable outputs associated with a given chain of experimental procedures guarantees coreference.

The use of experimental protocols links to both the descriptive (Frege, 1893; Russell, 1905; 1919; Searle, 1983) and causal accounts of reference (Kripke, 1972; 1977; Devitt, 1990). Before any investigation and further hypothesising takes place, the referent is fixed by a laboratory protocol. For example, the 'normal' (control) and 'depressive' (test) subjects are literally produced the same way from one experiment to the next. It is not merely question of describing them, but an active filtering of the physical ingredients of the experiment takes place. The 'reference-fixing' is therefore primarily causal.

Note however that a distinction should be made between the causal links that constitute the experiment proper (i.e., the list of operations and their correlated observations) and the causal links allegedly underlying the 'natural' functioning of things, as posited by higher-level interpre-

tative hypotheses. There are no grounds to further assume, as Putnam (1991) might have, that the test picks up 'natural kinds' and therefore succeeds in capturing some essential structural difference between 'normal' and 'depressed' subjects. In the context of experimentation, the subjects are simply 'produced' the same way during a series of experiments. The depression test is reliable simply because a certain number of behaviours tend to occur in an all-or-nothing pattern, such that the more symptoms a subject exhibits, the more likely is that he or she exhibits the other symptoms as well. The correlation of the depression syndrome with the observational results of a depression test is a typical example of an 'experimental correlation' between empirical data and experimental procedures (operations). Whether deeper structural differences are associated with the 'depressed' and 'normal' subjects is a matter of further investigation. Even if 'depression' actually refers to such differences, it is impossible to know that a priori. The inductive association of these symptoms under the larger umbrella of a 'depression syndrome' is empirically accurate, yet this does not entail in any way that the symptoms necessarily share a common cause.

Although necessary, a purely causal account remains insufficient for the purposes of large-scale experimentation. In the actual scientific practice, a description is attached to the causal story, namely a description of the protocol – in this particular example, of the test and the way in which the test is administered –, as well as of its results. From the standpoint of standard scientific practice, it is not enough that co-reference is conserved by producing subjects the same way. In addition, experimenters must communicate their results to a larger scientific community and coordinate their experiments with those of different research groups. 'Reference-borrowing' relies therefore on a detailed description of the 'pointing tool', in this case, the depression test.

Attaching a 'birth certificate' to each 'birth act' solves a number of difficulties attached to purely causal accounts of reference. The central idea is that an experimental protocol is able to physically generate or select the relevant object of study; in this sense, the protocol is causally linked to an object to which it points. A description is however further needed on at least two accounts:

1) in order to establish that the protocol consistently points to objects having highly similar observable properties (i.e., it picks the same kind of objects); and
2) in order to ensure that the same protocol (i.e., the same 'pointing tool') is used every time.[3]

### *2.3 Reference Incommensurability*

Notoriously, Kuhn's paradigm account of scientific theories supports the thesis of semantic incommensurability (1970 pp. 102, 128, 149). But if the meaning of a term changes in various theories, how it can be ascertained that its referent is always the same? Presumably, under a new theoretical treatment, the same term may point to something else (Kuhn, 1976; Sharrock, et al., 2002). Kuhn's historicist approach joins Feyerabend's 'contextual theory of meaning', which claims that the meanings of scientific terms depend on the theories to which they belong (Feyerabend, 1962). We have just seen that, in the context of cognitive psychology, 'depression' is

[3] Note that the above account diverges from most hybrid descriptive-causal accounts available in the literature (Evans, 1973; 1982; Devitt, 1981; Devitt, et al., 1999; Devitt, 2004). The latter are concerned with the problem of reference change and, as a rule, propose that reference-fixing is both causal and descriptive while keeping reference-borrowing primarily causal. The aim here is to avoid problems of misrepresentation and false belief, such as those associated with purely causal accounts of a naturalistic or evolutionary sort. For example, fear is typically thought to be experienced in reference to dangerous objects, persons or situations, yet a considerable proportion of the population fears harmless snakes. According to evolutionary explanations, the fear experienced by the subject refers to venomous snakes that are no longer present in the subject's natural habitat. The behavioural response refers to nasty snakes which used terrorise some remote ancestors, but the subject takes it to refer to something else, namely a harmless snake crawling on the asphalt (Murray, 2004). A description (i.e., an internal representation) of that to which fear refers helps, if not in avoiding feeling fear, at least in realising that the fear refers to something else than what triggered it.

about evaluating causal relationships in a certain way; in parallel, according to neurological models, the same term 'depression' points to a serotonin-norepinephrine imbalance (Castren, 2005). To this date, there is no available explanatory model establishing a connection between the levels of serotonin and the ability of subjects to evaluate contingencies. The example suggests, first, that that it is impossible to reduce one theory to another; and second, that it is not at all clear what theories of depression refer to and, more importantly to the issue of incommensurability, if they refer to the same thing.

In relation to the realism – antirealism debate, Cartwright (1983 pp. 103-104) and Morrison (2000) argue that the presence of different models of the same phenomenon entails a commitment to incoherent ontologies. The claim is that it is impossible to rationally commit to incoherent ontologies, and therefore the multiplicity of models must be an indication that these models have, most probably, a purely instrumental or pragmatic value. For instance, Morrison observes that

> *"we use different representations for different purposes: the billiard-ball model is used for deriving the perfect-gas law, the weakly interacting attracting rigid sphere is used for the van der Waals equation and a model representing molecules as point centres of inverse-power repulsion is used for facilitating transport equations.* [Thus,] *an explanation of the behaviour of real gases (something the van der Waals law is designed to explain) requires many different laws and incompatible models."*
>
> (2000 p. 49)

### *2.4 Reference Continuity across Distinct Models and Theories*

Experimentally (operationally) defined reference can help solve the problem of reference incommensurability by dissociating it from the wider issue of semantic incommensurability

(Dretske, 1981; Psillos, 1999 pp. 293-300). Back to the depression example, I already noted the absence of a theoretical connection between cognitive psychology and neuropsychology: since it is not known how a neurotransmitter imbalance can affect contingency judgments (or vice versa), there is no theoretical connection between the two. The semantics of 'cognitive depression' seems therefore to be completely independent from the semantics of 'neurological depression'. Nevertheless, despite the blatant absence of a theoretical connection, the two usages of the term 'depression' definitively co-refer since the subjects used for testing both hypotheses are literally produced the same way. Without reading the 'Methodology and Protocols' section of scientific papers we could have never figured that out and continued to err in the higher spheres of incommensurable paradigms.[4]

In the case of Morrison's example, a theoretical connection is already present: both models belong to the larger class of mechanical-kinetic models. The problem raised by Morrison stems, in part, from an incomplete theoretical understanding of how and why a gas like $N_2$ behaves like an ideal gas at room temperature, yet tends to behave more and more like a van der Waals gas under high pressure and low temperature. Without further knowledge or assumptions about the atomic structure of nitrogen molecules, kinetic models fail to provide an explanation of this transition and assume the status of alternative theoretical descriptions contradicting each other. Nevertheless, with or without a unified explanation, the two models of the nitrogen gas co-refer in the sense that they are both models describing the behaviour of the same chunk of matter. Instead of rejecting both because their partial semantic incompatibility, co-reference ensures that we are equally justified in accepting both kinetic models.

[4] It is worth noting that Davidson (Davidson, 1984) challenges Kuhn's attempt to abolish the distinction between 'the given' and 'the interpreted' precisely in relation to the issue of reference.

Different models reveal different aspects of the phenomenon under investigation. Their conjunction offers a more complete, but not a unified account; this angle of attack is usually known as 'perspectival realism' (Giere, 1999; Rueger, 2005). Note however that the 'perspectivism' I have in mind is limited to cases where partially confirmed models are able to explain two phenomena, but not the transition of these phenomena from one another. In other words, I concentrate on situations where two phenomena are or can be in direct physical continuity, yet the explanatory models fail to reflect this continuity at a theoretical level. This does not include models issued from distinct theories, each postulating different sets of unobservables, yet aiming to explain the same phenomena, or again alternate mathematical formalisms equally successful in providing a theoretical treatment of the same body of empirical data.

The net result is that one can accept the truth of several models without having to worry about theoretical continuity. The 'billiard ball' model accounts for the behaviour of $N_2$ at room temperature; the 'sticky balls' model accounts for a different phenomenon, namely the behaviour of $N_2$ molecules at low temperatures combined with high pressure, and for the eventual liquefaction of nitrogen. In lack of a more complete understanding of the transition between the two phenomena, scientists settle down for a piecemeal understanding of one phenomenon at a time. Similarly, 'depressive realism' investigates an empirical correlation between the syndrome of depression and cognitive faculties, while neuropsychology investigates an empirical connection between the same syndrome of depression and serotonin levels. What is lacking is a unified understanding encompassing at the same time the empirical correlations of depression with their cognitive and neurological manifestations, not evidence that depression has a dual, cognitive and neurological correlates.

### *2.5 Cumulativity without Unification*

The above considerations suggest that scientific knowledge is open to the addition of new co-referential models, where this multitude of models does not necessarily collapse into a unified model. In one of his vulgarising books on physics, Asimov makes use of a very compelling example which illustrates this point:

> "*Imagine a cone constructed of some rigid material such as steel. If you hold such a cone point-upward, level with the eye, you will see its boundary to be triangle. Holding it in that orientation (point-up), you will be able to pass it through a closely-fitting triangular opening in a sheet of steel, but not through a circular opening of the same area. Next imagine the cone held point toward you at eye-level. Now you see its boundary to be that of a circle. In that orientation it will pass through a closely-fitting circular opening in a sheet of steel, but not through a triangular opening of the same area. If two observers familiar with two-dimensional plane geometry but not with three-dimensional solid geometry, were conducting such experiments, one may hotly insist that the cone was triangular since it could pass through a triangular hole that just fit; the other might insist, just as hotly, that it was a circle, since it could pass through a circular hole that just fit it. They might argue thus throughout all eternity and come to no conclusion. If the two observers were told that both were partly wrong and both partly right and that the object in question had <u>both</u> triangular* and *circular properties (based on two-dimensional experience) might be an outraged, 'How can an object be both a circle and a triangle?' However, it is not that a cone <u>is</u> a circle and a triangle, but that it has both circular and triangular cross sections, which means that some of its properties are like those of circles and some are like those of triangles.*"
>
> (1966 pp. 136-137, vol. 2)

This fictional scenario is meant to provide a metaphorical understanding of the wave-particle dual nature of electrons. It can however be used as schematic representation for a wide

variety of cases. Schaffner (1994) discusses a fascinating example from immunology. In the early 70s, on the general background of Burnet's 'clonal expansion' theory, three mechanisms responsible for generating genetic diversity were proposed: the germline hypothesis, the somatic mutation hypothesis and the recombination hypothesis (Podolsky, et al., 1997). The initial project aimed to establish which of them is true. It turned out that the mechanisms postulated by all three hypotheses are responsible to some extent for generating genetic diversity of mature B-lymphocytes.

Asimov's schematic example shows that in order to provide a higher-order theory capable of yielding a conceptual model satisfying both empirical constraints – the object behaving both like a circle and a triangle, as determined experimentally by making the object pass through circular and triangular openings –, the scientists inhabiting a fictional two-dimensional space have to postulate the existence of a third dimension and hypothesise that the object in question is a cone. Yet whether or not their theoretical resources enable them to formulate a unified tri-dimensional model, the fact of the matter is that the object is, experimentally speaking, both circular and triangular.

If a unified model is available, then they conceive no contradiction in asserting that the object behaves both ways. If it so happens that a unified model transcends their conceptual abilities, the scientists fail to conceive how the object can possibly behave both ways and are faced with a conceptual contradiction, yet they still have no other choice but to accept the fact that the object behaves nevertheless both ways and that the alternative bi-dimensional descriptions depicting it as circular and triangular are both true. Coreference maintains and empirically justifies a link between two descriptions independently of whether this link is further reflected at the level of a higher-level theoretical understanding.

Alternatively, assuming that the two-dimensional scientists inhabiting Asimov's fictional world can hypothesise a tri-dimensional cone-model and that they have the logical means to argue that this model is the only unified model conceivable, it still doesn't follow that the object is a cone. In order to show that a tri-dimensional model is true and not only empirically adequate, the scientists must find a way to manipulate the cone in a three-dimensional space, tilt it for example, to make it pass through a series of openings matching various combinations of conical sections. Then the third dimension would be made accessible to experience, confirming, or at very least corroborating the cone model. On the other hand, if it so happens that the scientists cannot manipulate the object in the third dimension, then the cone is bound to remain a theoretical concept, a genuine unobservable that explains, but to which no empirical reality is attached.[5]

Ultimately, Asimov's example indicates that it is not always the case that knowledge, reason and concepts succeed in unifying the empirical datum. In the real life, any scientist is also an empiricist and, as an empiricist, he or she knows when to bow in front reality and acknowledge the fact that sometimes it is empirical reality which keeps together our knowledge, our reason and our concepts. In the scientific practice, this 'keeping together' of knowledge by extra-theoretical factors amounts to experimentally (or operationally) defined coreference.

[5] This last case matches van Fraassen's constructive empiricist approach (van Fraassen, 1980; 1989): the circle and triangle two-dimensional descriptions can be embedded in the cone tri-dimensional model, where the cone is merely a higher unity postulated via a theoretical unobservable, while only the circle and triangle-like behaviours are experimentally assessable.

# CHAPTER 3

# THE OBSERVABLE – UNOBSERVABLE DISTINCTION IN THE ACTUAL EXPERIMENTAL PRACTICE

## *3.1 The Observable-Theoretical Distinction*

The above use of coreference as an argument against incommensurability is very much akin to Carnap's two-language approach (Carnap, 1956). On one hand, science relies on a lower-level observational language $L_O$ which is essentially uninterpreted and whose vocabulary $V_O$ designates observables; this is, for example, the depression test. On the other hand, there is also a higher-level theoretical language $L_T$ whose descriptive vocabulary $V_T$ consists of theoretical terms; the hypotheses linking depression to cognition and neurotransmitter imbalance belong to this second language. Correspondence rules $C$ connect the theoretical vocabulary $V_T$ with the observational vocabulary $V_O$; in this particular example, depression is defined both observationally and theoretically in such a way that it is impossible to talk about the cognitive or neurological implications of depression in the absence of a symptomatic description of depression.

Unfortunately, Carnap's approach relies on the now infamous observable-theoretical distinction. At some point in his illustrious philosophical career, Carnap wanted to elucidate the underlying logical structure common to all scientific theories (the so-called 'language of science'). In order to achieve this level of universality, he relied, among other things, on a fixed observational and theoretical vocabulary, as determined by a strict criterion of observability. Briefly, a predicate $P$ stands for an observable property of object $x$ if, under suitable circumstances and given an inter-subjective consensus, empirical observations suffice to distinguish between $Px$ and $\Re Px$; conversely, any predicate failing to satisfy this requirement is deemed to be theoretical (Carnap, 1936).

Carnap's criterion of observability encountered numerous critiques, most of which can be classified in two categories. One variety of objections points out that some t-terms are observable in some circumstances, just as some commonly observable properties may be unobservable if attributed to unobservable entities (Putnam, 1962; Maxwell, 1962). The aim here is to show that there is no sharp observable-theoretical distinction in an attempt to rescue realism about unobservables. The other class of objections aims to show that all observations are theory-laden and therefore there are no such things as observables (Duhem, 1906; Hanson, 1972; Kuhn, 1970). The deeper philosophical motivations underlying the latter class of objections are more diverse, ranging from holism to relativism and social constructionism.

### *3.2 Psychological Considerations*

Typically, the theory-ladenness of observation is formulated as a combination between a psychological thesis pertaining to the peculiarities of human perception and a thesis about the use of observation in science (Heidelberger, 2003). Hanson's Gestalt approach to categorisation and

object recognition is an example of the former, while Duhem's formulation of the theory-ladenness thesis illustrates the latter.

Hanson argues that "*all seeing is seeing as*", by which he means to say that sensory data is already interpreted in light of previous knowledge: "*Observation of x is shaped by previous knowledge of x*" (Hanson, 1972 p. 19). To illustrate the notion of 'top-down' processing[6], we are asked to consider the Necker cube (1972 pp. 8-9). There are three ways in which we can 'interpret' the drawing: we can see it as a bunch of lines intersecting on a plane, as a cube seen from above or as a cube seen from below. The example is meant to demonstrate that there is no uniquely favoured 'interpretation' of perceptual observations, where the term 'interpretation' refers loosely to perceptual and activities including biological conditioning, socially-imposed learning, representation, categorisation, recognition, inference, etc. On the basis of this quick insight, Hanson defends a weak form of perception relativism, that is, the view that what we perceive varies depending on our previous experience of reality.

Hanson's insight is certainly in agreement with popular psychological schools, including Gestalt psychology, Neuro-Linguistic Programming, etc., and it is backed up by some research in cognitive psychology. Nevertheless, if Hanson is right in pointing out that whether a subject will recognise a drawing as belonging to a certain category is a matter of learning, especially if given an incomplete amount of information, the real question is whether this learning is dependent on a pre-existing theoretical framework. This joins a more general analysis applying to the subsequent examples discussed by Hanson (the young-old woman optical illusion, the recognition of an X-ray tube, etc. (1972 pp. 11-16)).

[6] For a quick introduction to cognitive illusions and the distinction between 'bottom-up' and 'top-down' processing, see Richard (1997). For the social implications associated to 'top-down' processing, see Palmer (1999). For Helmholtz's initial insights, see Helmholtz (1866).

For instance, a physicist recognises an X-ray tube via a cognitive mechanism of categorisation while a tartar shaman fails to do so. Note however that both the shaman and the physicist agree whether they are looking at the same object (or drawing/picture of the object) independently of the category in which they place the object. Hanson is under the impression that such an agreement is impossible because the

> "*influence on observations rests in the language or notation used to express what we know, and without which there would be little we could recognize as knowledge*".
>
> (1972 p. 19)

This may very well be the case, yet, as any experimenter in the field of cognitive psychology knows, it is not required that the subjects describe the object in words; subjects can draw it, or simply choose among several objects the object depicted in the picture of an X-ray tube. Irrespective of what the shaman and the physicist think or say, their ability to pick the X-ray tube and its corresponding picture clearly indicates that they just plain 'see' the X-ray tube independently of what they say or think they 'see it as'.[7]

This indicates that the disagreement concerning the 'interpretation' of the drawing occurs at a further level of perceptual processing, usually pertaining to cognitive activities related to categorisation, while we all agree about a lower level processing whereby the visual impressions are produced and become accessible to our consciousness. Typically, psychologists argue that categorisation is important for the purposes of decision making. Several findings indicate that categorisation is a constantly updated guessing strategy whereby our brains try to reconstruct bigger pictures from a limited number of clues. This guessing activity is essential for survival since it triggers rapid decisions based on a limited amount of information, allows a rapid com-

[7] Unlike Hanson, it seems that Kuhn understood the problem and tried to address it by making a distinction between 'data' and 'stimulus' (Kuhn, et al., 1969). Unfortunately, in his later works he reverted back to a version of original views by promoting the notion of a theory-driven 'taxonomy'/categorisation (Kuhn, 1977 p. 310). The present critique of Hanson's arguments from psychology also applies to Kuhn's 'taxonomical' approach.

munication of the essential information in a given situation, etc. It does not follow from here that the perceiving of the category hinders our ability to perceive the original empirical data from which the category is inferred in the first place.

This said, Hanson rightly observes that the physicist and the shaman disagree about the category to which the object belongs because they have been socially trained in a very different way. The basic idea underlying the argument is supported empirically. Presumably, we can show the shaman a series of pictures, ask him if the object he sees in the pictures is an X-ray tube, and then tell him if he is right or wrong. Soon enough our shaman will become an expert in the newly created field of X-ray tube recognition. This clearly indicates that categorisation is the result of learning. In turn, learning is impossible without a feedback. But where does the feedback comes from? Certainly not our own minds. In the case of the shaman, the feedback is artificial, but none the less external. We can conclude therefore that categorisation, or what Hanson calls 'interpretation', is not transcendental, in the Kantian sense, but remains nevertheless socially-relative.

It seems therefore that our social upbringing, and in particular our education, can influence the way we perceive the world. Is however this socially programmed 'seeing as' immune to the influence of what we actually see? The answer seems to be "No". It can be easily shown that not all categories are socially-relative. For instance, a physicist using an X-ray tube knows what she can do with it whether or not she can consistently classify pictures of X-ray tubes as belonging to the 'X-ray tube' category. Unlike the shaman, who uses visual cues, the physicist relies on experimental properties in order to construct a category of objects which transform electricity in X-rays. Although limited to the personal experience of the physicist, the category is neither arbitrary, nor artificially imposed. In spite of all the social conditioning in the world, not any object

can be viewed as an X-ray tube: only those objects which actually convert electricity in electromagnetic radiation bear some similarity with an 'X-ray tube' exemplar or prototype.

A similar argument can be made for other categories: even if they are socially reinforced, but they remain open to the influence empirical experience. Categorisation is a constantly updated guessing strategy. Granted, by socially reinforcing certain categories, for instance via aggressive political ideologies or shameless marketing, subjects can be mislead in 'recognising' something which does not exist based on an empirically inadequate association of certain cues, present in a situation, with other properties, not present, but postulated by the category. Nevertheless, categories are constantly revised. Whatever learning mechanism was responsible for forming the category in the first place doesn't cease to function. Even assuming that social training continues ad infinitum and whoever is responsible for it doesn't run out of resources, which by itself is a physical impossibility, at least a conflict arises between the category formed from experience and the one imposed via social pressure.

Thus, even though some aspects of perception are relative to the empirical experience and the social background of the trained subject,

1) it does not follow that our access to empirical reality is exclusively mediated via a learning mechanism; and
2) it does not follow that learned perception, such as categorisation, is exclusively a matter of theoretical interpretation.

### *3.3 Higher-Level Theoretical Interpretations*

Duhem's conception of the theory-ladenness of observation stems from a very different kind of considerations. Duhem remarks that scientists seldom work with raw, theoretically un-

interpreted experimental data. Scientists don't just perform experiments, but try to understand what they are doing, such that, ideally at lest, a theoretical explanation is attached to every experimental protocol, technique, instrument or operation (1906 pp. 147, 159).

Duhem's argument is not that we have trouble observing, or just 'seeing' as Hanson might have put it, but rather that scientists invariably interpret their observations in light of a theory:

> "*In the first place,* [an experiment] *consists in the observation of certain facts; in order to make this observation it suffices for you to be attentive and alert enough with your senses. It is not necessary to know physics; the director of the laboratory may be less skilful in this matter of observation than the assistant. In the second place, it consists in the interpretation of the observed facts; in order to make this interpretation it does not suffice to have an alert attention and practiced eye; it is necessary to know the accepted theories and to know how to apply them, in short, to be a physicist.*"
>
> (Duhem, 1906 p. 145)

Evidently, Duhem does not deny the existence or possibility of strictly observational data. Rather, he dismisses them as a lesser form of knowledge proper to younger sciences:

> "*When many philosophers talk about experimental sciences, they think only of sciences still close to their origins, e.g., physiology or certain branches of chemistry where the experimenter reasons directly on the facts by a method which is only common sense brought to greater attentiveness but where mathematical theory has not yet introduced its symbolic representation.*"
>
> (1906 p. 180)

It seems however that the theory-ladenness thesis is accurate solely in respect to what we may call 'mature sciences' (the textbook version of science), it is reasonable to conclude that during the development of any science – physics included, since it too was a young, immature, predominantly experimental science at some point in the past – an initial body of experimental

data is gathered in the absence of a satisfactory theoretical explanation of the experimental methodology, while a full explanation of the 'how' of methodology becomes available later. To give a banal example, anyone can ride a bicycle without understanding how this is possible. Even more importantly, it is the experimental possibility of riding bikes which justifies the conservation of the angular momentum and not the conservation of angular momentum which justifies us riding bikes. Similarly, it is possible to separate proteins based on molecular weight without fully understanding how this feat is achieved. Biochemistry and molecular biology moved forward without waiting for a complete physical explanation of the techniques it uses.

Thus, while acknowledging that, as Duhem points out, a theoretical interpretation is typically attached to text-book descriptions of scientific experiments, we must also keep in mind that

1) a theory-free, observational/experimental description of a phenomenon is also possible; and
2) it is conceivable that experimental data posits an initial (or 'input') set of constraints on the formulation higher level theoretical interpretations, as opposed to only theory positing a constraint on the interpretation of empirical data.

### *3.4 Knowledge in the Absence of a Theoretical Interpretation: An Example from Microbiology*

The well known Gram staining technique consists in collecting a bacterial sample (i.e., white mounds growing in a Petri dish; for the sake of brevity, I will skip the experimental protocol responsible for producing the 'white mounds' starting from patient blood, sputum, etc.), smearing on a glass slide, treating the sample with crystal violet, wash & dry, treat with iodine, wash & dry, stain with safranin and fushin, wash & dry, and, finally, examine under a light microscope; for a complete protocol and explanation, see Ryan and Ray (Ryan, et al., 2004). Let's

say untreated samples yield little black rods. Certain treated samples yield thicker purple rods, some samples yield thicker dark blue rods, while other samples yield a mixture of purple and blue rods. Since it is possible to tell the difference between the mixing and non-mixing of the sample with various recognisable reagents, experimenters know whether they subjected a sample to *C* or not. The observational outputs are likewise distinguishable, as they involve the presence of purple and blue rods:

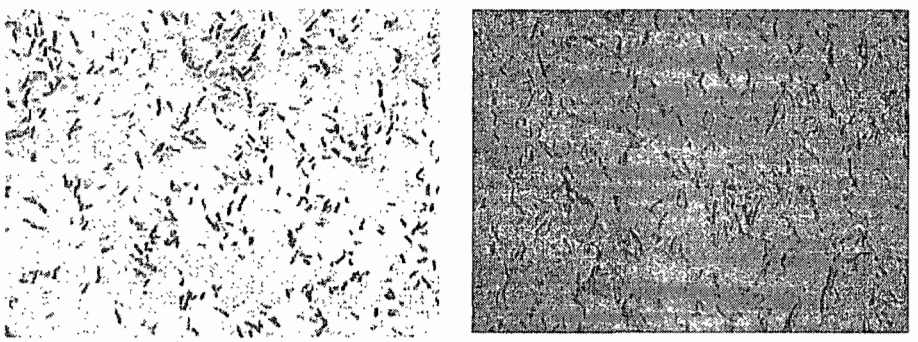
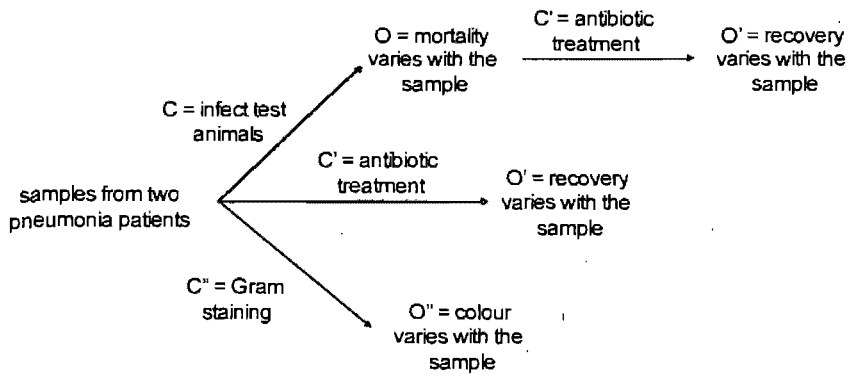



**Figure 3. Gram Staining**

Thus, perceptually indistinguishable input samples are correlated with different observational outputs when subjected to the same experimental protocol. No special theoretical knowledge is required in order to perform the staining. Also, no special theoretical language is required in order to describe the procedure or its observational outputs.

The reader might be curious to find out how this technique was developed without the benefit of a theoretical understanding. As a matter of fact, students who try differential staining for the first time often complain that it is really hard to tell the blue rods from the purple rods, especially if the sample is over-stained due to a sloppy technique. It is therefore conceivable that Gram might have missed the difference. Yet he didn't miss it. Why is that? The answer is very simple. The two bacterial samples he used have different degrees of pathogeneicity in respect to

their ability to cause pneumonia. After Gram developed his technique, it was also found that the bacteria in question differ in their susceptibility to known antibiotics.[8] Thus, the same samples, when subjected to different experimental treatments, are consistently associated with a variety of respectively different observational outputs, as summarised in the chart below:




Correlations of observational outputs: Gram - (red) – less virulent – susceptible to beta -lactam antibiotics

Gram+ (blue) – more virulent – resistant to beta -lactam antibiotics

**Figure 4. Correlations between Observational Outputs and Empirical Data**

Differential staining was perfected precisely so that anyone capable of performing a certain series of operations (including a machine devoid of any capacity of understanding) can rap-

[8] Interestingly, the above strategy relates to Hempel's solution to the 'raven paradox' (Hempel, 1945). In respect to enumerative induction, it has been observed that the fact that one particular raven is black confirms, to a small extent, the generalisation "*All raven are black*". The same inductive rule can be used for 'non-ravens': the fact that a non-black thing is a non-raven, say a white shoe, confirms, to a small extent, the generalisation "*All non-black things are non-ravens*". But "*All non-black things are non-ravens*" is logically equivalent to "*All raven are black*", entailing the paradoxical result that a white shoe confirms the generalisation "*All raven are black*". Hempel's solution to the paradox is to observe that the actual hypothesis extrapolated via induction is not "*All raven are black*", but rather the context-sensitive disjunctive "*Everything is either a black raven or it is not a raven*". In other words, the world is divided in three exclusive categories of objects: 'non-black ravens', 'black ravens' and 'non-ravens'; the first category falsifies the hypothesis while the last two confirm it.

At some point in the history of microbiology, Gram applied something very similar to Hempel's solution, but in reverse. If we would want to find out if there is a category of birds which are all black, we would start collecting all the birds and place them in two bins depending on whether they are black or non-black. If, at any point during the enumeration procedure, there is a certain type of birds, say birds of type 'raven', which fall all in the 'black' bin and none in the 'non-back' bin (as Hempel points out, we have to check both bins in order to establish this), then we have reason to inductively conclude that there is a category of birds which are all black. Similarly, in the differential staining example, it is not question of establishing a direct correlation between a certain colouring and virulence, but of establishing a correlation between a set of recurrent descriptors of the observational output. Thus, a correlation between 'blue staining' and enhanced virulence can be established if, within the class of all observable patterns obtainable following a certain laboratory protocol, only 'blue staining' correlates with allergy while non-'blue staining' patterns don't correlate with high virulence.

idly distinguish between more and less pathogenic pneumonia causing bacteria and administer the right treatment. No theoretical understanding as to why some bacteria stain differently from other bacteria is required. Not only the development of the staining technique was a trial and error process, but microbiologists continued to use this technique in order to administer the proper antibiotic treatment long before a satisfactory explanation of the staining process was proposed. In other words, for almost half a century, 'Gram positive' continued to mean 'blue' and 'Gram negative' continued to mean 'red', where 'blue' is bad for the patient while 'red' prompts to optimism.

### *3.5 Properties of Objects vs. Properties of Objects in Certain Experimental Contexts*

According to Maxwell (1962) and Hacking (1982), instrument-mediated observations allow for the 'indirect observation' of objects and properties of objects inaccessible to direct, perceptual observation. It seems therefore that what is unobservable in one context becomes observable in another and vice versa. In this sense, Maxwell likes to argue that seeing through a window, through prescription glasses, through low power binoculars, through a lunette, and finally through a telescope is always a matter of seeing. Epistemologically, all observations have the same status: they are all perceptual.

Although it is correct to assume that instrument-based observations extend the empirical knowledge of the world, I don't think this demolishes the observable – unobservable distinction. Maxwell argues for a smooth, unproblematic continuity between direct and instrument-mediated observations. Note however that it is not immediately clear that the observational properties of instrument-mediated observations are necessarily properties of the referent of the original naked

eye observations. It is not my intention to deny that they are. All I am saying is that there is an inference at work here and that this inference requires some additional justification.

The Gram staining example shows that observability is not about properties of objects, but of properties of objects given a certain experimental setup. The uninterpreted observational output of a 'differential staining & visual examination of samples under light microscope' experiment is such that we can distinguish colours. The same is not true if the same samples are observed through an electron microscope. By contrasting the telescope observation example with the Gram staining example, it becomes clear that Maxwell's argument is incomplete, since he would most probably like to ascertain that Jupiter is as depicted by the observational output of a telescope observation experiment, but deny that Gram-negative bacteria are red independently of the technique used to observe them.

Scientists are absolutely certain about the observability of the property $P$ (red or not-red); however, they not sure whether $P$ is a property of the object $x$ or of object $x$ under a certain experimental treatment. Thus, there is a subtle discontinuity between direct and instrument-mediated observations that must be bridged by a more or less sophisticated theoretical inference.[9]

[9] As bonus point, confining the observability criterion to the context of a particular experiment consisting of a series of operations $C$, $C'$, $C'''$, ... takes care of 'the redness of microscopic red particles' kind of paradoxes. Maxwell argues that the same property $P$ is observable in one context, but unobservable in another context. This claim yields conflicting reactions, for on one hand we are really eager to agree that 'red' is an observable property of objects, while on the other we are reluctant to claim that the 'redness' of microscopic particles is something we can observe. As pointed out earlier, the solution to the paradox is quite simple: we are absolutely certain about the observability of the property ($P$ or $\Re P$); what we are not sure about is whether $P$ is property of the object $x$ or of object $x$ under a certain experimental treatment.

### *3.6 Towards a More Flexible Approach to the Observable-Unobservable Distinction*

With the above comments and examples in mind, it becomes possible to establish a link between the observable – unobservable distinction and inferences about the sameness or difference of several objects of study given an identical experimental treatment.

The observational outputs of experimental operations are by definition perceptually distinguishable or indistinguishable. Experiments are designed in such a way so that they yield observational outputs which fit our natural ability to distinguish, differentiate and recognise. It is not required to adopt Carnap's binary mode of comparison; human beings can distinguish at the same time between two or more sizes, colours, shapes, sounds, etc., and, as a consequence, they can compare at the same time two or more properties of several objects.

The operations associated with experimental manipulation are likewise assumed to be distinguishable. As a general rule, the central requirement is that the objects under investigation receive an identical experimental treatment; accordingly, operations are conducted in such a way that variability is reduced to a minimum or at very least is kept under control. The repeatability of the technique and the consistency of its test and control observational outputs justify the reliability of the operation and offer internal criteria determining whether the operation was 'properly' executed (i.e., it does not diverge from the statistically relevant distribution of the positive and negative control outputs; this is usually known as 'calibration').

If perceptually identical objects yield distinguishable observational outputs under identical experimental contexts, then the two objects must be different and their difference is empirically justified: the difference must belong to the object itself rather than the experimental setup, which remains unchanged. In this case, the difference is said to be 'observable' since the two objects behave differently under identical conditions.

In contrast, if the observational outputs associated with the experimental manipulation of two perceptually indistinguishable objects of study are identical, it is still possible to claim that the two objects are different (or, more rarely, that the seemingly experimental treatment they receive is different or that both the objects and the experimental treatment are different), only this time the differences are unobserved (i.e., theoretical or hypothesised) rather than empirically justified. In this case, their distinguishability is deemed to be 'unobservable' and, as a general rule, further experiments are conducted until some distinction can be experimentally ascertained. Typical hypotheses include further theoretical glossing such as talk about different unobservable structures, mechanisms, properties, etc.

Finally, note that the converse situation whereby an unobserved identity of objects is hypothesised despite a consistently repeatable empirical distinguishability under identical experimental treatment is deemed to be logically inconsistent with the fact that the two objects have different experimental properties.

We are now in position to make a couple of important points about the observable-unobservable distinction:

1) Against Carnap's and Maxwell's conceptions alike, the distinction, as drawn in the experimental practice, doesn't necessarily have something to do with the properties of an object (e.g.: the redness of red cells, where 'red' is deemed to be observable no matter what), but rather with a comparison of objects based on a comparison of their experimental properties (the redness of the Gram stained samples should be distinguished from the redness of the bacteria, the former being an empirical property of the experimentally manipulated object while the latter is a theoretically hypothesised property of the object).

2) The distinction does not entail that every property belongs to either the observational part ($V_O$) or the theoretical part ($V_T$) of the scientific vocabulary; rather it is question here of perceptual distinguishability (e.g.: it is not question of deciding whether 'red', as a property of an object or of the observational output of an experiment, is observable, but only whether 'red' is distinguishable from other properties, such as 'blue')
3) The distinction marks the border between knowledge about the behaviour of various objects in various experimental setups (e.g.: empirical correlations between bacterial samples and various observational outputs) and hypotheses about the objects themselves (the bacteria are structurally different).
4) It asymmetrically justifies the jump from experimental knowledge to hypotheses about the structure of the studied objects: we are empirically justified to conclude that two objects (usually perceptually indistinguishable, such as the sputum samples from two patients) are different given their differential behaviour under identical experimental treatment; the same distinction becomes unobservable (or theoretical) the moment it hypothesises differences in the objects despite identical behaviour under identical experimental treatment.
5) The justification of the distinguishability of objects given their differential experimental behaviour relies on consistent repeatability; if two objects consistently display a differential behaviour in a variety of experimental setups despite the best efforts to provide a uniform experimental treatment, experimenters infer that the differences are due to the objects themselves; such low-level inductions yield predictions open to verification (e.g.: blue-stained samples predict antibiotic resistance).

6) The distinction does not require that the differences introduced by theoretical hypotheses remain forever unjustified empirically (i.e., unobservable), but it does state that they are not justified by and therefore transcend our current empirical and experimental knowledge.

### *3.7 Concluding Remarks*

The above characterisation refers explicitly to the early stages of experimentation, preceding theory formation, and may not apply to experiments designed to test models or theories (the so-called 'crucial experiments'). In line with this disclaimer, note also that this characterisation fits common patterns of experimentation in biology and psychology, which are relatively 'young' sciences. It seems however reasonable to conclude that early experiments in physics and chemistry obeyed a similar pattern. For instance, it can be easily shown that the initial experimental knowledge behind the functioning of devices such as ammeters, voltmeters ohmmeters and galvanometers relied on a web of empirical correlations between the various observational properties of an electric source rather than on a consistent theoretical interpretation. The same goes for the acid-base distinction and the classification of elements in chemistry.

Ultimately, the initial stages in experimentation amount to something along the lines "Keep doing something to two indistinguishable objects until you discover a difference in their behaviour". Thus, in an initially theory-free context, experimentation is nothing else than a systematic endeavour to uncover the potential (or 'thus far hidden') empirical properties of objects.

This change in the premises of the problem entails a divergence from Carnap's formulation of the observable/unobservable distinction. We are no longer concerned with providing a database of observable properties (the needle of a voltmeter, the height of the red marker in a

thermometer, etc.), but with the issue of perceptual distinguishability: perceptually indistinguishable samples, objects, properties, events, situations, etc. are differentiated on the basis of their perceptually distinguishable behaviour (different positions of the needle, different heights of the red marker, different band patterns, etc.) under identical experimental treatment. While the observability of properties changes with the epistemological status of the objects possessing them, perceptual distinguishability is dependent solely on the functioning of sense organs. In as much as these organs are identical for all the members of the species, perceptual distinguishability is also bound to remain constant for all human observers.

Once the description of objects in terms of observable properties is replaced by comparisons of objects in terms of experimentally-produced distinguishability, it becomes possible to retain a stable distinction between purely experimental knowledge and further theoretical interpretations. As discussed previously, this distinction is needed in order to make possible coreference and thus overcome the problem of reference incommensurability.

# CHAPTER 4

# THE THEORETICAL INTERPRETATION OF EXPERIMENTAL DATA

## *4.1 The Historical Development of Scientific Knowledge*

The two main ingredients of a scientific theory are experimental data and theoretical hypotheses. Regrettably, logical positivism overemphasised the former at the expense of the latter, while post-positivist philosophy of science tends to do just the opposite. Hoping to restore the balance, I showed in the previous section how the distinction between observables, or experimental data, and unobservables, or higher level theoretical hypotheses, is drawn in the experimental practice.

Once we begin to realise that it is possible to gain experimental knowledge preceding any theoretical interpretation, we can envisage the possibility that the former can impose constraints on the formulation of the latter. This order of determination is required in order to justify the use of a wide array of reasoning strategies, including inference to an explanation, devising empiri-

cally adequate theoretical explanations, and establishing a criterion for distinguishing between 'good' and 'bad' science.

As the reader may have noticed, science textbooks rarely mention the painful tribulations of experimenters and the timid correlations drawn without the guiding light of a suitable theoretical interpretation, just as they seldom mention the immense technological payoffs of these tribulations. Instead, they provide detailed explanations telling us why what scientists do or have been doing for a long time works. What matters for a typical science textbook is a concise, well rounded and consistent theoretical treatment subsuming a handful of key pieces of empirical data. What doesn't matter is how scientists acquired the knowledge they claim to have. Vulgarisation books, especially if bent on a historical perspective, might touch some details, yet not too many and certainly not too systematically, for a thoroughgoing description would soon bore the reader. Rather, a great deal of importance is given to obsolete interpretations initially proposed and eventually abandoned in favour of other, more adequate interpretations. This sequence of hypotheses adds a bit of suspense and excitement bringing science vulgarisation literature closer to their more successful competitors, the police investigation and espionage thrillers.

I make these remarks with a purpose. The material covered in this section is neither a technical elaboration of textbook material, nor a more detailed historiography of science, but something complementary to both: it presents a chunk of pure experimental knowledge, that is, it exposes that portion of knowledge which remains unchanged throughout the series of interpretations enumerated by the historians of science up to the current interpretation covered by contemporary science textbooks.

### *4.2 Mixing Direct Observations and Uninterpreted Experimental Data: An Example from Immunology*

The example I want to discuss comes from immunology and was originally introduced by Klee (1997 p. 34). As the reader may already know, antibodies (immunoglobulins) are responsible for humoral immunity. The antibodies are proteins found in blood and on mucosal surfaces, where they bind antigens such as allergens or proteins on the surface of bacteria and viruses, cross-linking them in order to form heavier, insoluble and hopefully biologically inert aggregates or marking them for digestion by macrophages and other cells involved in the defence against parasitic organisms. One of the key elements responsible for elucidating the functioning of humoral immunity pertains to knowledge about the structure of antibodies. In particular, it was found that each antibody comprises two binding ($F_{ab}$) domains, capable of preserving the antigen binding specificity of the whole antibody protein, while the rest of the antibody ($F_c$), which does not interact with the antigen, is responsible for the formation of precipitates or is recognised by immune cells. The schematic representation below may be of some help:

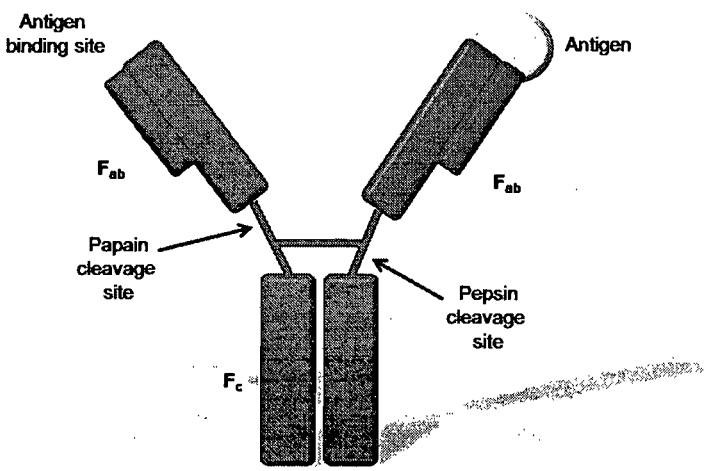



**Figure 5. Molecular Structure of Antibodies**

Suppose now that we go back to a moment in time where immunologists didn't know any of these structural details yet. By analogy with Mendel's 'long-shot' hypothesising of alleles and their segregation as an attempt to explain certain features of sexual reproduction, we can simplify matters and state that immunologists eventually hypothesised the existence of bio-molecules having a particular geometrical structure in order to explain the phenomenon of humoral immunity phenomena as evidenced by passive immunity, immunisation with inactive pathogen, etc.

The empirical constraints imposed onto Mendel's genetic theory consist of direct observations about the relative proportions of various phenotypic traits across several generations. The theory had to be empirically adequate in respect to this body of empirical data describing sexual reproduction in order to constitute a scientifically valid hypothesis about the phenomenon of sexual reproduction. Thus, in agreement with confirmationism, the introduction of theoretical terms is conditional upon entailing at least one empirical feature of the explained phenomenon.

However, in the case of immunology, there is a complication to be taken into account. The empirical data used by Mendel had the benefit of not needing any further interpretation. Shortly put, the statement "25% of the offspring are pea plants bearing yellow bridled fruits" means that 25% of the offspring are pea plants bearing yellow bridled fruits. The empirical constraints imposed onto the antibody-model of humoral immunity posit a further challenge, as they consist of experimental data in addition to direct sense observations. As discussed previously, the properties of observational outputs produced via the experimental treatment of objects cannot be automatically considered as equivalent to the directly perceived properties of the objects in question without running into problems. In order to convert the properties of observational outputs into properties of objects themselves a theoretical interpretation of the experimental techniques is required.

Does this mean that the empirical adequacy constraints imposed onto the antibody-model of humoral immunity are bound to be theory-laden? Not at all. For as long as experimenters don't make the mistake of confusing the properties of an observational output with properties of the studied objects, and understand that experimental data asserts properties something about the studied objects only in light of a further inference, they are free to mix observational and experimental data without having to endorse any theoretical interpretation of the experimental techniques used. Contrary to Duhem's expectations, scientists are not required to assume an understanding of the techniques used to investigate a phenomenon before they can hypothesise an explanation of that phenomenon. Scientists can also first attempt a holistic explanation of the sum total of direct observational data and experimental data, and then cross-reference this explanation with other explanations making use of data gathered via the same techniques in order to divide it into an explanation of humoral immunity and a theoretical interpretation of the experimental techniques.

In order to understand this approach, let us consider some typical experiments in immunology. As pointed out at the beginning of the chapter, it is not my intention to follow the historical development of immunology [for a thoroughly documented history of immunology see (Silverstein, 1989)]. Rather, I want to show by means of a simplified example – an example roughly mirroring the historical sequence of some key discoveries in the field of immunology – how an interpretation of experimental data emerges from the overall explanation of a body of empirical data consisting of both direct observations and experimental outputs.

First there is well defined body of directly accessible data. For instance, the initial antibody sample is, at the origin, serum extracted from the blood of an animal, say a rabbit, exposed to an antigen. The antigen is known, readily available and chemically characterised. Positive and

negative control sera have different biological properties. For instance, if the antigen is an allergen, then whenever injected to a test animal, it causes an observable allergic reaction. Yet if the same antigen is mixed with the serum derived from an animal already exposed to the antigen before being injected, no allergic reaction follows. The negative control sera don't have this property. The biological/immunological properties constitute an operational description of the initial antibody sample. In terms of low-level inferences, immunologists can further establish that something in the blood of the sensitised animal is responsible for 'neutralising' the effects of the antigen; they know it is not a cell, since cells have been filtered out and are no longer observable in the serum.

Next comes a relatively large body of experimental data further extending the network of already available observational data. A brief description and interpretation of each technique is given in parentheses for reference purposes. Note that just as in the case of Gram staining, the experimental techniques mentioned here were used before a satisfactory theoretical understanding of the physics and chemistry underlying their functioning was hypothesised; in fact, even if a general explanation is available today, the precise mechanism behind most techniques is still uncertain.

Here is some of the biochemical analysis data. If purified antibodies is digested with papain (an enzyme which cleaves proteins in several pieces; operation $C_1$ in the figure below) and subject the digested antibodies to gel electrophoresis ($C_2$; typically SDS-PAGE, that is, a technique whereby denatured or partially denatured proteins acquire a negative charge and are then separated by molecular weight while traversing a gel medium under the influence of an electric current; proteins having different molecular weights travel different distances through the gel; bands are visualised by staining the gel with a chemical dye), a certain pattern of bands is ob-

served on the gel (see $O_2$ below). The electrophoresis procedure ensures that the antibodies were successfully digested by papain since a certain smearing of the blobs observed for digested antibodies will not be observed when whole antibodies are used. If the relevant pattern of bands is observed, researchers can proceed and purify the rest of the papain treated mixture by means of some form of chromatography ($C_3$; a technique whereby mixtures are separated based on their molecular weight or binding affinity for the stationary medium which they traverse) in order to separate the various fragments and use them for further study. A certain observational output ($O_3$) is associated with $C_3$. The purified fractions can be subjected to a variety of chemical and immunological assays. For instance, during $C_4$ the various fractions purified in $C_3$ are mixed with antigen and subjected to a version of the electrophoresis operation described in step $C_2$ in order to yield an observational output $O_4$ (commonly, co-iommunoprecipitation followed by SDS-PAGE is used, but for the purposes of this example we can suppose that the antigen is a short DNA sequence, in which case EMSA would do just fine).

The figure below summarises these four steps of the molecular analysis, providing a schematic, but fairly descriptive rendering of the observable outputs, together with the standard interpretation of each technique used [more details on the theoretical understanding of the separation and purification techniques mentioned here can be found in (Tinoco, et al., 1995)]:

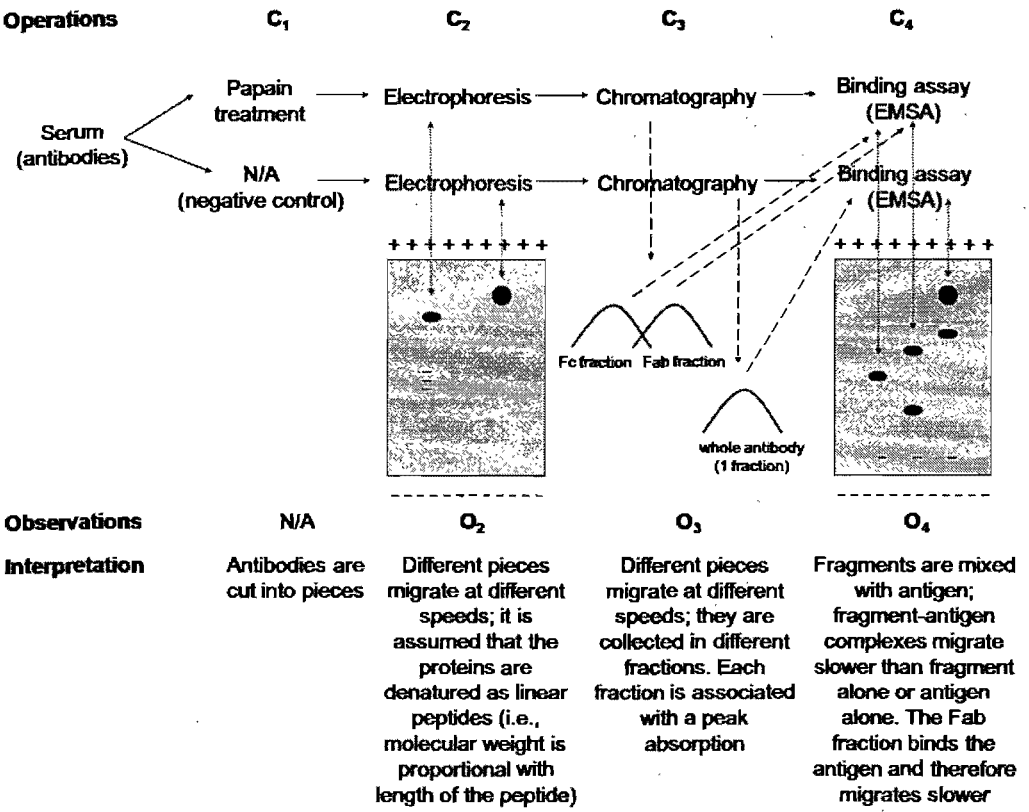



Figure 6. Example of Biochemical Analysis

The observational outputs $O_2$-$O_4$ further correlate with immunological assays. After connecting observational data with experimental data, researchers are free to return back to observational data. For example, they can test the chromatography fractions and see if it the latter have the same biological property as the initial serum sample. Usually, they don't. On the other hand, chromatography purified undigested antibody always does preserve the biological properties of the initial serum sample, indicating that the loss of function is not due to the chromatography procedure. Curiously enough, if the original antibody is digested with pepsin rather than papain while retaining the rest of the protocol, the four observational correlates described above are very similar, with the difference that the $F_{ab}$ fraction retains the antigen blocking activity and migrates at a slower speed during electrophoresis. The fractions collected on step $C_3$ can also be subjected to chemical analysis, which will further enlarge the web of correlations between operations, ob-

servational outputs, biological properties and chemical properties. More complicated experiments can likewise be conducted. For instance, if papain-treated antibodies are mixed with the antigen prior to adding the whole antibody or the pepsin-treated antibody, the animal still develops an allergy. And so on and so forth.

There is no immediately evident limit to the extent experimental knowledge can take. New experiments can add on the top of previous experiments yielding a huge database of uninterpreted or partially interpreted data. It is quite probable that the extent of actual experimental correlations roughly determines the initial boundaries of various sciences and domains of investigations. Higher-level hypotheses explain a phenomenon as described by networks of empirical data such that, prior to eventual inter-theoretical reductions and unifications, their empirical adequacy and explanatory power does not transcend the boundaries of the initial empirical description. In the context of this example, we can see how immunology overlaps with biochemistry, while it usually fails to invade the domain of psychology given the absence of stable correlations between psychological traits and immunological properties.

### *4.3 Lower-Level Inductive (Non-Interpretative) Inferences*

Experimental knowledge relies on a subtle web of low-level inductions linking operations and observations. Each inductive generalisation, usually amounting to the modest extension from one experiment to the next, counts as a mini-hypothesis. The beauty of experimentation is that there is no holistic semantics to worry about. The basic requirement is that $C$ should be perceptually distinguishable from $\Re C$ and $O$ from $\Re O$. Once this requirement is satisfied, each operation ($C$) and each observational term ($O$) is independent of other operations and observational terms. Thus, each inductive hypothesis ($C \rightarrow O$) is independent of similar hypotheses concerning other

correlations (in the example from immunology, $O_2$ does not follow as a matter of definition from $C_2$, just as $C_3 \rightarrow O_3$ does not follow a priori from $C_2 \rightarrow O_2$; rather, each material implication are the result of an induction generalising over a finite number of repetitions).

These 'baby step' hypotheses remain very close to the initial data and the strength of each correlative association is revised on an individual basis. Even though they are deemed non-threatening for empiricism and empirical justification, it is important to realise that they retain an inferential character whereby something substantial is posited, namely

1) the existence of an 'object' or 'substance-like' entity consisting of 'potential properties' existing independently of the experimental treatment which renders them actually observable; and
2) different objects given different observational behaviour under identical experimental treatment.

Carnap initially supported the idea that there is no object beyond its perceptual and experimental properties (Carnap, 1928). Nevertheless, a minimal object must be posited in order to account for the reproducibility of experiments. For instance, a Gram-positive result reliably predicts penicillin-susceptibility before the latter actually manifests. The reliability of such low-level inductive generalisations – generalisations which concentrate most of the technological and practical worth of experimental knowledge – indicates that whether or not we test for certain experimental properties, and therefore whether or not these properties are actually made empirically manifest by means of an experimental technique, they are 'potentially' there at least in the sense that the expectation to find them there is always confirmed. The old problem of dispositional properties resurfaces again, and it seems the only way out is to make place for an object, or

'substance' made of latent properties which become manifest only under specific experimental treatments.

The second kind of inferences relies on the premise that two observable phenomena cannot be identical at the level of their unobservable structure if they have different experimental properties. Although this principle does not hold true if applied to the entities postulated by probabilistic, statistic and stochastic models, it seems to be always taken for granted when dealing with observable phenomena. In other words, scientists don't hypothesise different underlying structures for objects that are perceptually and experimentally indistinguishable. To do so would be tantamount to recognising that some unobservables are by definition bound to remain forever unobservable no matter how far we push experimentation and empirical investigation.[10]

### *4.4 "Navigating By the Instruments"*

In the immunology example discussed previously, if the papain digestion ($C_1$) – electrophoresis protocol ($C_2$) yields a certain observed pattern of bands on the gel ($O_2$), this further correlates with a certain number of purified fractions ($O_3$) subsequently obtained following chromatography ($C_3$); conversely, failure to obtain $O_2$ following $C_{1\text{-}2}$ correlates with a failure to obtain $O_3$ following $C_3$. The same applies to the $C_4$-$O_4$ tandem in respect to the previous $C_3$-$O_3$ tandem. The whole chain must be satisfied if researchers are eventually to obtain the $F_{ab}$.

Thus, Carnap's tandem of conditionals ($(C \square O) \rightarrow T$ and $(C \square \Re O) \rightarrow \Re T$) usually maps onto what experimenters call the 'positive' ($C \rightarrow O$) and the 'negative' ($C \rightarrow \Re O$) controls. The idea is that if $C \rightarrow O$ is not satisfied (e.g., a specific pattern of bands is not observed), chroma-

[10] This said, it might be worth noting that the objects and their differences are not immutable givens, but constantly revised hypotheses updated as required by the addition of new correlations and the eventual demise of older ones.

tography separation will not yield the same fractions. Most importantly, separation will not yield a $F_{ab}$ fraction which preserves the biding specificity of the original antibodies and which contains twice as many individual proteins than the original sample of antibodies. Thus, it makes sense to talk about $F_{ab}$ and their properties in reference to the positive control, but not in reference to the negative control.

This analysis precludes a possible misconception. It might be argued that experimental protocols are complicated procedures pointing to or picking up some arbitrary category such as 'three-legged cats with green eyes', or, in this particular case, 'papain treated sera described by certain electrophoresis patterns and chromatography fractions'. The '$F_{ab}$ production protocol' differs from 'three-legged cat with green-eye' criterial filter in at least one important aspect: it compares pairs of 'positive'/'test' and 'negative' control descriptions. The same serum sample containing purified antibodies is divided in two halves. Half is digested with papain, the other half is not; then, both are subjected to the same series of operations including electrophoresis, chromatography and binding analysis. This establishes that the observational differences between the positive and negative controls correlate with the presence or absence of a particular operation, in this case, the mixing of serum with papain. In particular, this means that researchers cannot physically obtain the $F_{ab}$ fraction following chromatography and study it afterwards without having previously subjected the sample to papain digestion. Conversely, researchers can gather a whole database of, say, gel pictures ($O_2$ in the figure above) whereby they note that digested samples of purified antibodies or other proteins display a similar 'smearing' pattern. This induction allows them to infer that, given the observed pattern, the sample might have been subjected to papain (or some other enzyme) digestion. There is no such knowledge of experimental causes associated with mere criterial filters.

True enough, from a strictly operationalist point of view, most of the molecular analysis serves the sole purpose of producing 'observational traces' correlating band patterns with other pieces of experimental knowledge concerning various allergy and immune reaction assays. This is especially true about steps $C_2$ and $C_4$, which rely on physical techniques of separation during which the samples are trapped in the gel and cannot be used for further biological or chemical assays. Without further knowledge or assumptions, interpreting the observational outputs of these operations remains impossible. Nevertheless, these outputs systematically correlate with different operations such as 'papain-treatment' and 'pepsin-treatment', thus allowing the experimenter to literally navigate by the instruments in order to achieve the desired results in the absence of any systematic interpretation of the experimental techniques.

### *4.5 Higher-Level Interpretative Hypotheses*

The positivists and logical positivists were so impressed with the potentially unlimited expansion of purely experimental knowledge and its ability to generate practical applications independently of any overarching theoretical interpretation that they begun flirting with the idea that, set aside the requirement of low-level inductions, scientists might not need theoretical hypotheses after all. The project failed on two accounts. First, it disagrees with the actual scientific practice. And second, the moment the positivists tried to translate theoretical interpretations in terms of experimental knowledge, they were suddenly left with an unaccounted excess of information. Carnap's tribulations from meaning verificationism to confirmationism illustrate the problem quite convincingly.

When it comes to the interpretation of experimental data, two approaches can be envisioned. The first one is a Duhemian piecemeal interpretation whereby scientists propose an inter-

pretation for each operation *C* in relation to observation *O*, and, given this interpretation, try to infer something about the phenomena under study. The alternative is to propose an interpretation consistent with an explanatory hypothesis about the phenomenon under study. Under the first approach, the interpretation of experimental data precedes and is independent of the further hypotheses about the nature of the phenomena under study. Under the second approach, the theoretical interpretation is a corollary of the explanatory hypothesis (i.e., if this hypothesis is true, then this particular interpretation of the experimental data must also be true).

In reference to the example from immunology, it is quite obvious by now that there is a gap between the experimental knowledge exposed so far and knowledge about antibodies being Y-shaped proteins. In order to breach the gap, scientists had to abandon the safety of low-level inductions and conjecture much bolder hypotheses. The Duhemian approach amounts to the hypothesising of a mechanism for each biochemical analysis technique, and assuming this interpretation is correct, infer that the original sample is composed of whole macromolecules while subsequent operations are concerned with parts of these macromolecules. Ironically, Duhem advocated confirmation holism while arguing that scientists must provide a piece-by-piece interpretation of experimental data before hypothesising an explanation of a phenomenon. Note also that the nature of the interpretation determines, or at very least constrains, the nature of the explanatory hypotheses.

The alternative approach is largely ignored in philosophical circles, so I will spend more time elaborating it. The idea is to start by hypothesising an explanation and, given this explanation, infer an interpretation consistent with it. Let us suppose that antibodies bind specific antigens and that papain and pepsin treatment cuts the antibodies in several pieces. It becomes now possible to further hypothesise that different pieces migrate at different speeds during electropho-

resis and chromatography. An electrical current and, respectively, gravitational pull provide the motive force, while a porous material retards bulkier pieces more than, smaller and lighter pieces. The bottom smearing observed in positive control lane of $O_2$ is due to the production of small fragments. Complete digestion must lead to the complete disappearance of the corresponding band in the negative control lane (i.e., the whole, undigested antibody). Incomplete digestion would leave it there along with the smaller bands, and sharper bands corresponding to fragments of the antibody. This can be verified by varying the time of exposure of the sample to papain or pepsin. Additional bands appearing in both the positive and negative controls must be impurities. Similarly, the bottom bands in $O_4$ must be the free, unbound excess antigen, the middle bands must be the free $F_{ab}$, $F_c$ or whole antibody, while the top bands must be the antigen bound to $F_{ab}$ or the whole antibody ($F_c$ does not bind the antigen, hence the presence of only two, instead of three bands). But wait a minute! If the bottom line is the antigen and the top one is the whole antibody, then it would be interesting to add three extra control lines along with the three test lanes, one to run the original, undigested antibody, one for the digested antibody and one for the antigen. Lo and behold, we have a match! The bottom line is indeed the antigen, the top one is antibody, while the middle pieces correspond to the digestion pattern. Let us now add three more control lanes containing the three fractions collected in $C_3$ and corresponding to the three peaks observed in $O_3$. This is indeed our lucky day! $O_2$, $O_3$ and $O_4$ all fit perfectly our interpretation. Even if we don't understand yet very well how electrophoresis works, we can already infer that the bands migrating at the same speed (all the aligned bands) are or point to the same fragments or proteins.[11]

[11] The interpretation extends to the biological assays as well. If $F_{ab}$ obtained following papain digestion blocks allergies, we may suppose it binds the antigen in such a way that the latter cannot bind something else and cause the allergy. We can further hypothesise that the difference between $F_{ab}$ produced by papain digestion and whole antibody or $F_{ab}$ produced by pepsin digestion is due to a different cleavage of the initial antibody. On one hand, this explains

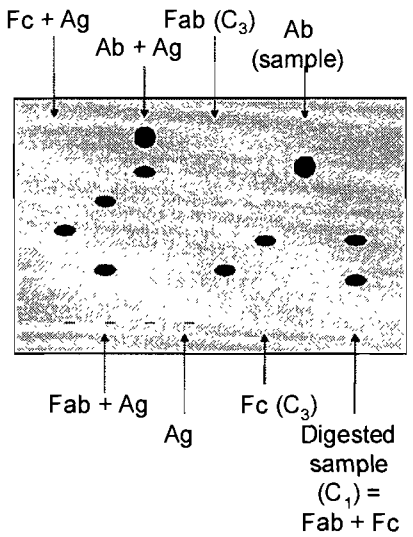



**Figure 7. Reading Electrophoresis Gels**

Historically, the development of the molecular explanation of humoral immunity relied on a mixture between the two approaches. For instance, in the case of electrophoresis, some aspects of the interpretation – most notably the idea that electrostatic forces act on charged macromolecules – were borrowed from physics in order to provide a partial interpretation of the technique. Other aspects – namely those pertaining to the separation properties of the gel, ultimately enabling researchers to 'read' the gel – pertain to an interpretation consistent with a hypothetical explanation of humoral immunity and other biological phenomena. What is important to realise is that experimental data does not need to be interpreted before an explanation of a phenomenon under study can be proposed. Uninterpreted data can also constrain the formulation of higher-level explanatory hypotheses.

why pepsin-$F_{ab}$ migrates slower than papain-$F_{ab}$. On the other, this is compatible with the hypothesis that whole and pepsin-digested antibodies link antigens in heavier complexes which precipitate or simply prohibit the antigen to reach its target due to its suddenly acquired mass and bulkiness.

# CHAPTER 5

# INFERENCES TO SYSTEMATIC EXPLANATIONS

## *5.1 Holistic Semantics and Inference to the Best Explanation*

Many realists like to argue that the fitting of all observations, experimental operations and overall empirical correlations between the various elements of the experiment amounts to an 'inference to the best explanation' [*IBE*; (Maxwell, 1962; Smart, 1963)]. The idea behind *IBE* is very simple: it would be very hard to find another explanatory hypothesis that fits experimental data as well as the current one. This links to a fundamental intuition implicit to the scientific practice: available empirical data constrains, or posits an 'input-constraint' onto the formulation of theories. A more modest, and more honest, way to put it is to state that not all explanations allow for systematic interpretations of the available experimental knowledge, especially if cross-referenced against larger and larger domains of investigation. In practice, some explanations allow for interpretations that can be uniformly applied over a given body of experimental data, while others cannot; the former are 'better' than the latter.

The proponents of *IBE* remain somewhat vague as to what may ultimately count as the 'best' explanation. For example, the above sketched interpretation is a partial one. Some details are left out. Unless we specify that the 'binding' and 'digestion' are chemical processes, we might as well posit little 'animacules' embracing, fighting or eating each other. For the sake of the argument, nothing prohibits these 'animacules' to become electrostatically charged and thus be propelled against their will through electrophoresis gels. More so, even if we specify that these terms should be understood chemically, meaning that both the 'binding' and the 'digestion' are functions of the chemical structure of the antibody, antigen and the protease, we still didn't specify what kind of chemical interactions are in place (covalent, ionic, H-bonds, etc.). *IBE* captures the fundamental intuition that each additional correlation in the web of experimental knowledge posits an additional constraint onto higher level interpretations, thus limiting the number of viable hypotheses we might propose. However, it does not tell us what counts as a complete explanation or what should be the optimal dimension of the body of experimental knowledge under interpretation.

The *IBE* approach evolved in parallel with two related arguments for realism: the 'no miracle argument', also dubbed the 'cosmic coincidence' or the 'convergence' argument [*CA*; (Putnam, 1975; Boyd, 1984)], and the 'experimental success' argument [*ES*; (Hacking, 1983)]. Without entering the details, these arguments state that when a consistent interpretation can be preserved not only for several experiments, but over whole domains of investigation, such as physics, chemistry and biology, *IBE* is strengthened to such an extent that it seems to become inconceivable that a single interpretation fits such a gigantic amount of empirical data without capturing something about the underlying structure of empirical reality. In the case of the example from immunology, the proposed interpretation is congruent with other pieces of knowledge

concerning molecular genetics, biology, microbiology, physical chemistry, etc. The implications are huge and it takes somebody trained in the field to fully appreciate them.

Despite its grandiose airs, the realist arguments try to reproduce at a holistic scale what low-level inductions achieve for individual empirical correlations. From this perspective, we can at least conclude that the strategy behind the argument is legitimate. It remains however to be seen how exactly such a large-scale inference would look like in practice and how aberrant interpretations are discarded. I will come back to this very important issue in the last chapters of the book.

### *5.2 Experimental Constraints and Justification*

Operationalism has a role to play in the overall practice of science and I can hardly see how experimentation could dispense of it. This said, I also think that this role is limited to the preservation of reference. Once a particular state, phenomenon or property can be reliably identified or produced, bare-bone experimental knowledge is embedded into a higher-level interpretation. Although it may be possible to push experimental knowledge further in the absence of any significant interpretative work and obtain knowledge that can be of immense practical interest, as a general rule, once a certain body of experimental data is acquired, scientists do not proceed any further without some guiding higher-level hypothesis.

Ideally, a scientifically adequate interpretation of experimental techniques should be 'uniform'. It should not ascribe different mechanisms or explanations to experimentally indistinguishable procedures or to empirically indistinguishable observations. Nevertheless, there is no strict interdiction to attribute different mechanisms to the same technique used in different circumstances. For instance, there is no formal interdiction to hypothesise, say, that the whole anti-

body does not migrate through a gel according to the same mechanism as the $F_{ab}$ fragments. Such a situation may arise in cases in which an interpretation fits some data, but not all the data associated with an experiment. Then, in lack of a better option or because of the success of the interpretation in relation to other experiments, researchers may decide to keep the interpretation and try to uncover some subtle experimental differences.

Note however that to systematically allow theoretical distinctions in the presence of contrary empirical data and non-conditional to an eventual empirical justification is tantamount to settling down for the view that an interpretation has nothing to do with empirical reality, in which case we should seriously ask ourselves what exactly the interpretation in question interprets and to what extent it makes sense to pretend that we deal with a theoretical interpretation rather than a dogma. As a former experimental scientist, it is my firm conviction that this kind of reasoning is healthy and that it captures the good intention behind empiricist eliminativism. Undeniably, this good intention often miscarried, yet, initially at least, it was definitively there. Even a convinced empiricist such as Mach understood perfectly well the necessity of introducing hypotheses:

> *Observation only leads, in the first place, to the conjecturing of laws of motions, which, in their special simplicity and accuracy are presupposed as hypotheses in order to try whether the behaviour of bodies can be logically derived from these hypotheses. Only if these hypotheses have shown themselves to hold good in many simple and complicated cases, do we agree to keep them.*
>
> (1893 p. 306)

We can see from this quote that the ultimate goal of empiricism is not to prohibit the formulation of hypotheses introducing excess semantic content, but rather to ensure that all hypotheses sat-

isfy the constraints imposed by actual experimental knowledge and therefore are at least empirically adequate in respect to it.

Empiricists disagree about how much excess content should be introduced and whether this excess content refers or not. Mach, Carnap and most positivists want to keep the excess to a minimum, that is, keep hypotheses as close to experimental data as possible. Usually, they also doubt that the extra t-terms introduced by a hypothesis refer and tend to ascribe them a purely instrumental role (i.e., they facilitate comprehension and computation, but may have nothing to do with reality). In contrast, van Fraassen (1980; 1989) is ready to allow the introduction of hypotheses as heavy and complicated as we want on sole condition that empirical adequacy is preserved. He also abstains from denying semantic realism or promoting instrumentalism; instead, he adopts an agnostic point of view according to which t-terms may refer and be true in a Tarski-style correspondence theory of truth usually adopted by semantic realists, yet, in the absence of proper justification of our beliefs, we cannot rationally commit to the reality of these terms and act as if they refer. On the realist side, but in an empiricist spirit, Feigl and Hempel would want to keep hypothesising to a minimum in order to facilitate empirical justification. In contrast, contemporary realists are not afraid to introduce more substantive hypotheses and rely on *IBE* and the *CA-ES* arguments as means of justification transcending the limits of classic, piecemeal empirical justification.

### *5.3 Conclusions*

Ultimately, it seems that talk about better and worse inferences is a function of empirical adequacy: an explanation allowing for a single interpretation uniformly consistent with all the experimental data is better than an interpretation consistent with only some of the data, which in

turn is better than an interpretation systematically inconsistent with most of the experimental data. To what extent a better explanation justifies realism remains however open to debate.

For the time being, I will retire the debate by highlighting a number of more modest points:

1) There is such a thing as a purely experimental knowledge devoid of any significant theoretical interpretation and relying solely on low-level inductive correlations between observations. Some aspects of this kind of knowledge are adequately captured by Carnap's confirmationism and a version of Brigdman's operationalism.
2) The main virtues of experimental knowledge are its ability to preserve reference to observables and to uncover necessary causes responsible for differential observational outputs.
3) More substantial hypotheses are needed in order to provide an interpretation capable of unifying several experiments and, beyond that, several domains of investigation such that knowledge from one domain (for instance, the physical chemistry needed to understand electrophoresis) can be transferred in another domain (such as immunology). Standard, textbook exposition of experiments relies on interpretative hypotheses which transcend strictly experimental knowledge.
4) Experimental knowledge ties down higher-level hypotheses to observable data, forcing them to refer at least partially to phenomenal reality at all times, and also posits empirical constraints on what may count as an acceptable hypothesis. It is not clear however if experimental knowledge suffices to force a unique higher-level interpretative hypothesis, as postulated by *IBE* and how this unique interpretative hypothesis would entail realism as postulated by the *CA-ES* arguments.

# PART II

# THEORETICAL CONSTRAINTS ON HYPOTHESIS FORMATION

# CHAPTER 6

# THEORETICAL EXPLANATIONS

## *6.1 Theoretical Contexts*

In the actual scientific practice, hypotheses are tied down to experimental data. At the same time, they are also 'tied up', so to speak, to more general patterns of explanation and interpretation. This 'top-down' determination of hypotheses is tightly linked to the theory-model distinction in science, as well as with issues in explanation and reductionism.

The intuition at work here is the following: in the immunology example, there is not immediate stringency forcing the 'chemical' interpretation over the 'animalcule' one when judging the situation from the standpoint of the experimental data alone; still, if we keep in mind that a chemical interpretation is consistent with experimental data issued from a whole host of other experiments in various other domains of investigation, there is a net import of information to be considered. The integration of a particular set of experimental data into a larger theoretical con-

text can be approached along the lines of a reduction, or that of a conjunction. In the case of reduction, the data is entailed by the larger theoretical context as one consequence among many other, hopefully confirmable, consequences. Alternatively, the experimental data may entail new predictions in conjunction with a given theoretical context.

### *6.2 The Deductive-Nomological Account of Explanation*

Since many explanations establish a link between an empirical description of a phenomenon and a set of more general laws or principles, the deductive-nomological (*DN*) account of scientific explanation proposed by Hempel and Oppenheim (Hempel, et al., 1965) is of immediate interest. The guiding idea behind the *DN* account is that a set of premises consisting of laws plus statements specifying initial conditions logically imply an empirical description of the phenomenon to be explained. The reductive character of the explanation comes from the fact that the description contained in the conclusion must be deduced from, rather than merely consistent with the propositions stated as premises; this marks a difference between deductive and semantic approaches to modelling/explanation. The second requirement, pertaining to the 'nomological' component of the account, is that the necessary premises must include at least one law of nature. Explanation is thus tightly linked to the notion of 'nomic expectability', that is, to the notion that a particular prediction or empirical description is to be deductively expected given a certain set of laws (pp. 247-248).

### *6.3 Explanation vs. Justified Explanation*

The account faces several difficulties [reviewed in (Salmon, 1989; Woodward, 2003)], mainly in relation to its 'nomological' component. It has been often pointed out that there are no

satisfactory criteria defining the essential characteristics of what counts as a 'law of nature'. Initially, the 'nomological' requirement was introduced in order to mark a distinction between accidentally universal statements and genuine laws of nature. Hempel contrasts "*All members of the Greensbury School Board for 1964 are bald*" with "*All gases expand when heated under constant pressure*". The ideal gas law explains the behaviour of some gases, while the fact that a person is a member of the Greensbury School Board does not explain why that person is bald. Given this common intuition, Hempel argues that the explanatory value of the later proposition comes from the fact that it is a law of nature, while the former is merely an accidentally true statement, holding true in respect to a very specific domain of application.

Unfortunately, the notion that laws of nature are exceptionless, absolutely universal regularities is highly problematic. Most generalisations in biology and psychology, as well as most empirical laws in physics are not exactly exceptionless. We don't have to search very far for counterexamples: the law of ideal gases, which Hempel introduces as a paradigmatic example of a law of nature, is not a general statement true of all gases in all situations.

Presumably, Hempel insists on tying down explanation to 'nomic expectability' in an attempt to ensure that the explanation is not merely conceptually possible, but also justified because it is the corollary of a universal proposition. But what if it is impossible to decide a priori whether a proposition is indeed a 'law of nature'? In a broad positivist tradition, the key requirement is that the premise-propositions should describe empirical reality. However, since induction from particulars is always imperfect, it is often difficult to distinguish between accidental regularities, confined to particular cases, and genuine 'laws of nature' applying to a whole class of phenomena. For example, if our knowledge resumes to a single proposition, "All members of the Greensbury School Board are bald", then, when facing a bald person, the only explanation

we can think of is "He must be bald because he is a member of the Greensbury School Board". The explanation is conceptually possible given our current knowledge of the world. This does not mean however that the explanation has to be true or empirically adequate. It is merely a hypothesis we formulate. As any non-bald person would soon find, joining the Greensbury School Board would not result in her or him becoming bald, thus falsifying the 'because' underlying the explanation, as well as the universality of the initial premise on which the explanation is built.

In light of these considerations, I think it is important to distinguish between conceptually possible explanations given a certain set of theories and background beliefs and true, partially confirmed, approximately true or probably true explanations. The *DN* account applies readily to the former, but fails to provide an adequate characterisation of the latter.

### *6.4 The Causal Connections Underlying Scientific Explanations*

Scriven (1962) pushes the argument further and argues that it is hardly the case that all scientifically valid explanations rely on universal laws of nature. His conclusion is based on cases of singular causal events (e.g., "*The impact of my knee on the desk caused the tipping over of the inkwell*") which we commonly take as having some explanatory value.

In response to this objection, Hempel observes – correctly in my opinion – that singular events, even if genuinely causal in nature, have no explanatory power. Only reproducible causal events (causal regularities) have explanatory power (Hempel, 1965 p. 360). Irrespective of first impressions, there is always a causal regularity setting the difference between causal explanations and mere chronological narratives whereby events simply follow each other (e.g., we ex-

plain the tipping of the inkwell based on our previous knowledge that kicking desks results in objects on it tipping over).[12]

The exchange between Hempel and Scriven is highly informative. It strongly suggests that the deductive, and therefore purely logical, connection between premises and conclusions must be doubled by an instance of causation – actual or hypothesised – in the physical world. For example, the ideal gas law explains the behaviour of gases because it is possible to obtain an increase in volume (the empirical description standing on the side of the conclusion) by heating air at constant pressure (the premises from which the conclusion is deduced). In contrast, it is impossible to make a person bald by making him or her join some school board.

Scriven argues that explanations reveal causal connections. Hempel retorts that we cannot gain knowledge of causal connections unless the connection in question assumes the form of a readily reproducible regularity. Hempel's answer is not entirely satisfactory. Although it seems reasonable to conclude that regularity plays a role in the justification of knowledge about causation, at least in the more selective context of scientific explanations, it is still not immediately evident that regularity alone suffices to guarantee a causal connection.

For instance, it seems that universal-like statements can be explanatory independently of any reference to a causal connection if they reflect part-whole relationships. This is exemplified in the following diagram:

[12] Kitcher's distinction between the 'ideal' or 'full context' underlying an explanation and the 'non-ideal' or 'incomplete' formulation of the explanation in a given instance supports a similar kind of argument (Kitcher, 1989).

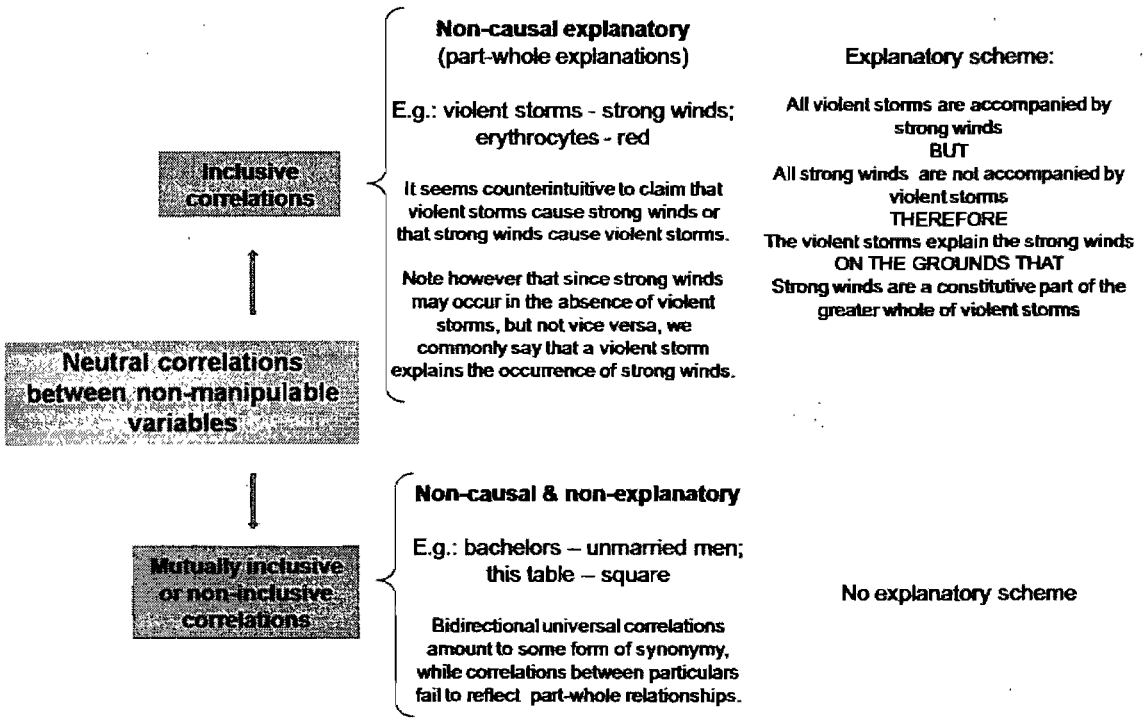



Figure 8. Non-causal explanations associated with universal laws or statements

Note however that, although popular in the everyday explanatory folklore, this kind of explanations is rather superficial and remains unsatisfactory from a scientific point of view.

Several counterexamples and study cases strongly suggest that the main shortcoming of the *DN* account stems precisely from its failure to take into consideration the causal connections associated with typical scientific explanations. Most notably, the *DN* account doesn't pay sufficient attention to the causal asymmetries involved in such explanations. This is illustrated by the notorious flagpole example. A flagpole is stabilised in a vertical position by 50 meters wire attached at one end to the top of the flagpole and at the other to the ground 40 meters away from the flagpole. The flagpole itself has 30 meters. These propositions fit the deductive scheme provided by the Pythagorean Theorem any way we want. We can deduce that the flagpole has 30 meters from the premises stating that the wire has 50 meters and it attached 40 meters away from

the flagpole, just as we can deduce that the wire had 50 meters or that it is attached 40 meters away from the flagpole from the remaining two propositions. Although deductively sound, Bromberger (1966) complains that the explanation merely reflects a geometrical proportion without really telling us why the flagpole has this specific height.

It might be retorted here that what stands on the side of the premises is the Pythagorean theorem, while the estimated height of the flagpole and the length of the wire stabilising it are in fact the conclusions of the deduction (i.e., if the Pythagorean Theorem holds true about the chunk of space containing the flagpole and its wire, then there is a fixed ratio relationship between the length of the flagpole and the length of the wire). Then, the claim would be that the Pythagorean metric explains the actual measurements asserted above, while the actual empirical measurements confirm, rather than explain, the hypothesis of a Pythagorean metric.

Nevertheless, once the geometrical relationship is complemented with physical laws, such as the laws of optics, the explanation becomes inherently asymmetric, while the *DN* explanation preserves the same symmetric structure. For example, in the case of the 'shadow of the flagpole' version of the above example (a 30 meters flagpole casts a 40 meters shadow at a particular time of the day), we would say that the length of the flagpole explains the length of the shadow and not vice versa despite the fact that we can calculate the height of the flagpole given the length of its shadow just as easily as we can calculate the length of the shadow given the height of the flagpole.

To make things worse, a version of the 'bald member of the school board' example also indicates that failure to take into account information about causal events renders some *DN*-style explanations irrelevant or even absurd. Consider the following "explanation":

> *"All males who take birth control pills regularly fail to get pregnant. John Johns is a male who has been taking birth control pills regularly.* [Therefore this explains why] *John Jones fails to get pregnant."*
>
> (Salmon, 1971 p. 34)

The deductive part is valid and one of the premises is universal non-accidentally, yet it is clear that the 'explanation' doesn't explain anything because, pills or no pills, it makes not sense to talk about male pregnancy in the first place.

### *6.5 Statistical Relevance*

It seems that the only way around cases of explanatory irrelevancies is to amend Hempel's 'regularity approach' to causality. Salmon's (1971) 'statistical-relevance' (*SR*) account of explanation aims to provide the required corrections without transcending into the realm of metaphysics. According to the *SR* account, an explanation is a 'body of information' relevant to an explanandum. Typically, the 'body of information' in question sums up the experimentally manipulable factors that correlate with an increase or decrease in the probability of a certain event (e.g., smoking in relation to lung cancer).

In Salmon's probabilistic formulation, given a population $A$, an attribute $C$ is statistically relevant to another attribute $B$ if and only if $P(B|A.C) \neq P(B|A)$. For example, given a certain population of men or women, the correlation between the use of the pill and the absence pregnancy is equally well supported for men and women. However, the probability of pregnancy among men = probability of pregnancy among men who take the pill = 0. In contrast, the probability of pregnancy among women is greater than the probability of pregnancy among women who take the pill. We can see therefore that the correlation between the use of the pill and the

absence pregnancy is 'statistically relevant' for women, but not for men. Shortly put, a correlation is 'non-accidental' or 'statistically relevant' if and only if it doesn't describe equally well a suitable control group.

The *SR* account is meant to be a fully functional alternative to the *DN* account: an explanation is no longer an argument conducted on the premise of universal laws – and certainly not on the premise of a theory or fundamental hypothesis –, but the expression of a 'statistically relevant' correlation. Nevertheless, it is interesting to note that the concept of 'statistical relevance' is compatible with the general structure of *DN* explanations.

According to Hempel, "explanation = deducibility from more general statements + further condition *X* defining an essential characteristic of the statements to be found on the side of the premises". Hempel argues that some of the premises must be laws of nature and defines condition *X* along the lines of 'non-accidental regularity'. Problems arise when Hempel further tries to establish a link between 'non-accidentality' and universality over some sufficiently large domain of application. Most empirical laws hold true only in respect to a limited and often incompletely specified domain of application. This strongly suggests that the 'non-accidental' character of the regularities they express has nothing to do with absolute universality over a well defined domain of application. Now, given this difficulty, we can reasonably argue that laws of nature are 'non-accidental' because they reflect a set of correlations between the occurrence (or probability of occurrence) of an event *E* and the presence or absence of some factor *F*.

### *6.6 Experimental Manipulation*

Ideally, Salmon's notion of 'statistical relevance' provides an epistemic definition of causality and, at the same time, a measure of the difference between 'naturally-occurring' instances

of $E$ given the presence or absence of $F$. The underlying premise here is that we decide whether a phenomenon is causal or not based on passive observations alone. Alternatively, cognitive psychologists tend to further reinterpret 'statistical relevance' along the lines of an artificial algorithm emulating the unconscious psychological processes behind our everyday causal judgments.

Nonetheless, in many cases the causal relevance of $F$ in the occurrence of $E$ is established via active experimentation. In the context of a typical scientific study, a correlation between $F$ and $E$ is locally evidenced in the context of a specific experimental setup in which researchers have direct control over $F$ (say, they can freely add to and subtract $F$ from the experimental setup). This means that researchers make a distinction between manipulable and non-manipulable experimental variables prior to and irrespective of the statistical methods employed to calculate the difference between test and control groups. This posits an interesting problem since, once this distinction is made, it is not at all clear whether 'statistical relevance' is synonymous to causality or if it is merely a measure of manipulability.

We can see therefore that, in some cases, the *SR* account overlaps with 'manipulationist' accounts of explanation (von Wright, 1971; Woodward, 2003). Manipulationist accounts are particularly efficient in accounting for the ability of everyday explanations to provide answers to 'why-questions'. For example, a falling barometer is a reliable indicator of an imminent storm, yet nobody claims that the barometer explains the occurrence of the storm (Bromberger, 1966). Intuitively, it seems that we distinguish between "the storm explains (or causes) the barometer to fall" and "the falling barometer explains (or causes) the storm" on the grounds that a falling barometer is a manipulable experimental variable while the storm is not directly manipulable. We observe that imminent storms reliably correlate with falling barometers, yet we also know that artificially placing the barometer under a vacuum pump fails to yield the same correlation. We

conclude that since we cannot produce storms by making barometers register lower pressures, it must be the storm that 'causes' the barometer to fall. The same goes for the 'shadow of the flagpole' example: we know that the height of a flagpole correlates with the length of its shadow, yet we also take into account the fact that one can always change the height of the flagpole, while it is absolutely impossible to directly change the dimensions of a shadow; we say therefore that the height of the flagpole explains the length of its shadow and not vice versa.

A more systematic classification of the relationship between experimental correlations and causal judgements having some explanatory value is provided in the figure below:

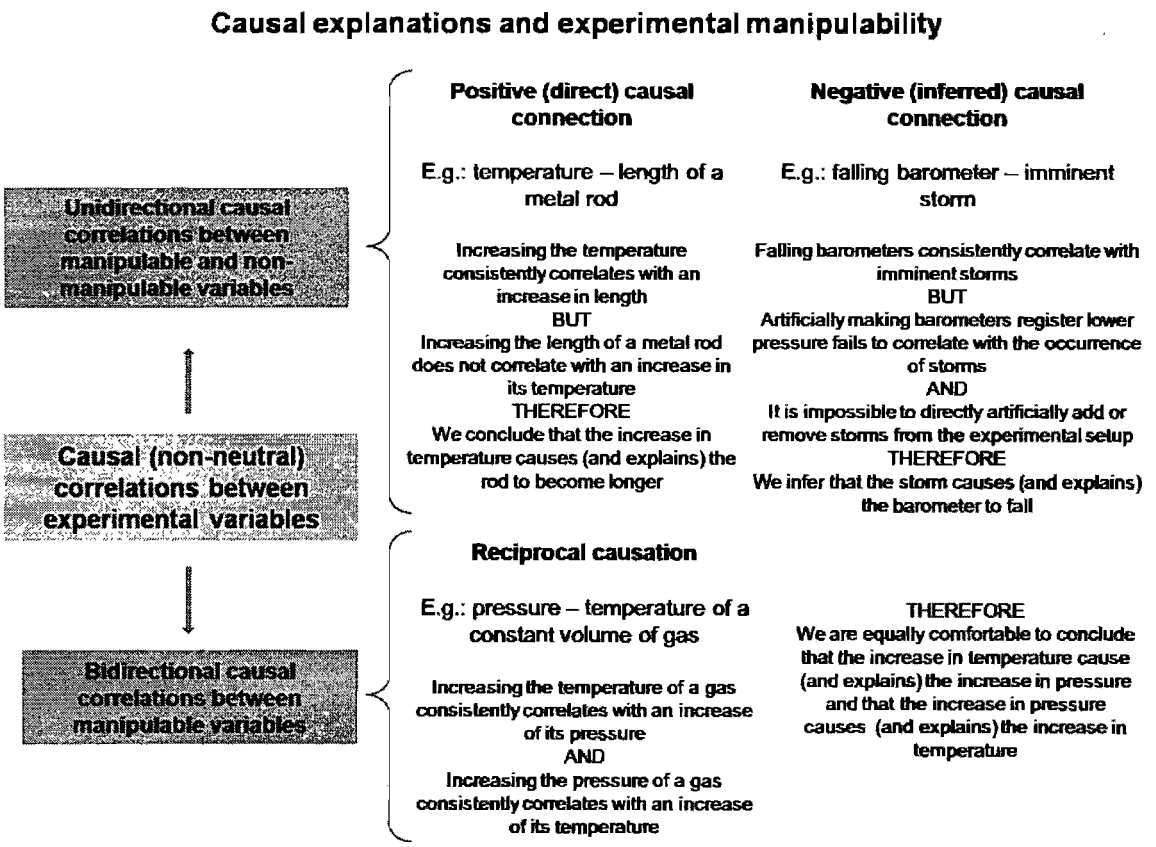



Figure 9. Causal explanations associated with experimental manipulation

### *6.7 Cognitive and Realist Interpretations of Experimental Manipulation*

Statistical approaches to causation, as well as manipulationist accounts are empirical and epistemic. Causality does not constitute a primitive notion, but is merely a body of empirical data satisfying certain conditions. It is in sense that Salmon, just as Hempel before him, as well as most contemporary researchers in the field of cognitive psychology, tend to talk about 'causality (or contingency) judgments' rather than causality proper.

In as much as it is possible to establish 'statistically relevant' correlations based on passive observations alone, it is also possible to reduce causality to judgments about statistically relevant regularities. However, in most cases 'statistical relevance' is established on the prior grounds of active experimentation involving the artificial manipulation of certain aspects of the experimental setup. In cases of active experimentation, causality judgments rely on the epistemologically primitive notion of 'experimental manipulability' and cannot easily dispense of it. For instance, without the distinction between manipulable and non-manipulable variables, it becomes very hard, if not impossible to account for the fact that the symmetric correlation between the pressure and temperature of a constant volume of gas grounds a pair of symmetric explanations whereby an increase of temperature explains an increase in pressure, just as an increase in pressure explains an increase in temperature, while the equally symmetric correlation between a falling barometer and the occurrence of a storm grounds an explanation functioning in one direction only, namely the storm explains the falling barometer while the falling barometer fails to explain the occurrence of the storm. The problem is not at much one of providing the right statistical formula capable of emulating most of our causal judgments; 'statistical relevance' is defined the same way in *SR* and manipulationist contexts. Rather, it seems that judgments about 'statistical relevance' are a necessary, but insufficient condition for concluding causality.

In addition to a strictly empirical understanding of the notion of experimental manipulability, some authors choose to attach further interpretations to this notion. At one end of the spectrum, Waskan (2006 pp. 225-253) proposes a cognitive interpretation of manipulability, dubbed the 'model model', according to which "*explanations for events and physical regularities are constituted by intrinsic cognitive models of the mechanisms that produce them*" (p. 225). The 'mechanisms' to which Waskan alludes pertain to strategies or sequences of events whereby certain effects are produced, in real life or in our imagination (pp. 227-228).

This said, it is worth observing that these intuitive mechanisms may or may not coincide with the mechanisms postulated by physical explanations. Presumably, there is an excellent overlap in the case of classical mechanical explanations or vulgarised versions of geometrical explanations in biochemistry and molecular biology; in contrast, quantum mechanics explanations and chemical explanations are bound to be much less intuitive. Also, it is likewise important to keep in mind that what we can imagine doing via intuitive mechanisms and what we can do in the actual experimental practice does not perfectly coincide either. For example, it is possible to send an electron through two slits at the same time, yet our imagination has a hard time imagining how this feat is achieved. Granted, imagination plays a huge role in devising explanations; examples such as the rock-on-a-string model for planetary motion or Einstein's thought experiments amply demonstrate this. Nevertheless, it seems reasonable to assume that we imagine intuitive mechanisms in accordance to an implicit set of fundamental rules. When these rules are rendered explicit then it becomes possible to construct models outside our imagination; most notably, it becomes possible to take into account new rules which our natural capacity to imagine may or may not accommodate.

At the other end of the spectrum, Baskar (2008) proposes a realist, non-epistemic version of the experimental/manipulationist approach:

> *"To ascribe a law, one needs a theory. For it is only if it is backed up by a theory, containing a model or conception of a putative causal or explanatory 'link' that a law can be distinguished from a purely accidental concomitance.* [...] *it must be assumed, if experimental activity is to be rendered intelligible, that natural mechanisms endure and act outside the conditions that enable us to identify them that the applicability of known laws in open systems, i.e. systems where no constant conjunctions of events prevail, can be sustained. This has the corollary that a constant conjunction of events cannot be necessary for the assumption of the efficacy of a law."*
>
> (2008 pp. 12-13)[13]

Baskar's approach stems from a criticism of Hempel's *DN* account in relation to its inability to mark a sharp distinction between genuine laws and accidental/local regularities. His solution to the problem is to view laws as hypothesising causal mechanisms underlying experimental manipulability rather than being the result of passive observations of naturally-occurring regularities. Note however that

1) the same approach applies just as well to the hypothesising of a causal connection attached to an unspecified 'black-box' mechanism; and
2) the notion of regularity is sill needed for the justification of the causal connection or mechanism.

For example, a true statement of the kind "All members of the Greensbury School Board are bold" can be easily transformed into a hypothesis about a causal connection of the form "If $x$ becomes a member of the Greensbury School Board, $x$ becomes bald". Since, as far as our experi-

[13] Baskar illustrates his approach via examples from physics and chemistry (2008 pp. 163-169).

ence goes, nobody ever became bald by joining some school board, the causal connection is falsified and, as a consequence, we judge this statement to be an accidental regularity rather than a genuine law of nature. On one hand, Baskar's approach rightly identifies the statement as being an accidental regularity on the grounds that it has no deeper causal connection underlying it. On the other hand however, his claims to realism are not entirely justified. First, it doesn't matter which mechanism is hypothesised, meaning that the realism attached to this interpretation is mainly about the existence of a causal connection and not about a particular explanatory mechanism and its associated ontology. And second, the falsification of the statement is justified by an experimental regularity whereby we repeatedly fail to produce baldness by making somebody joining a school board. This indicates that the distinction between laws holding true over certain domains of investigation and accidental regularities hinges more on the fact that the former are experimentally tested regularities, while the later are passively observed regularities.

### *6.8 Causal Mechanisms*

In his more recent work, Salmon abandons epistemic approaches in favour of realistic ones (Salmon, 1984). The rationale behind this radical change in approach hinges on the observation that although many common explanations are indeed nothing else than 'statistically relevant' correlations, when it comes to scientific explanations, scientists do not say that smoking explains cancer, but only that smoking correlates with lung cancer while non-smokers are far less affected by this type of cancer. What explains lung cancer is a series of cumulative mutations caused by various carcinogens present in the tobacco smoke. The core idea at work here is that statistical relevance only provides a stock of 'black-box' correlations, while a more complete

scientific explanation requires the 'black-box' is eventually filled by a causal mechanism and/or explained as a theoretical consequence.

Very briefly, the key notions of the 'causal mechanism' (*CM*) account can be summarised as follows: A causal process is a physical process characterized by the ability to transmit a mark in a continuous way; in contrast, non-causal processes fail to transmit marks. For example, if we mark a beam of light in a certain way (polarity, wavelength, etc.), the mark is transmitted as light propagates through space-time, reflects, refracts, etc.; in contrast, if we try to mark the shadow of a ball (say, we modifying its shape by adding a second source of light), the mark doesn't transmit as the object and its shadow move in space unless we keep modifying it at every moment in time. We say that two causal processes coming in close spatial-temporal proximity causally interact if only if they leave on each other a mark which would have not been present in the absence of their interaction. The paradigmatic example is that of two cars colliding: each car leaves a mark, namely a deformation, on the other car.

The main difficulty with the *CM* account pertains to an internal tension between a desire to keep the account as close as possible to directly accessible empirical features and the fact that most mechanisms and fundamental interactions underlying scientific explanations are framed in terms of theoretical unobservables. One way to formulate this objection is to argue that the ability of a process to transmit a mark may or may not be the causally relevant feature captured by a scientific explanation (Hitchcock, 1995). For example, although it is entirely true that a moving car is able to transmit a ripple when colliding with another car, we commonly explain the final motions of the two cars by appealing the notion of a momentum-energy transfer and not to the ability of a car to deform another car.

The *CM* account faces a dilemma: either we confine ourselves to the more superficial level of 'experimental manipulability' in order to effectively contain the explanation in the realm of the purely empirical, or, if we seek a 'deeper' or more 'complete' explanation – or to put it bluntly, a more satisfactory explanation from a scientific point of view –, then we must take the risk of introducing theoretical terms and thus plunge straight at the heart of the realism-antirealism debates raging in contemporary philosophy of science. To put it in a somewhat simplistic formulation, a 'causal mechanism' is to 'experimental manipulability' what a mechanism of perception is to perception. Just like perception, 'experimental manipulability' is a self-justified empirical given; and just as a mechanism of perception provides an explanation of perception, but does not automatically justify itself as being obviously true, a 'causal mechanism' explains the primitive notion of 'experimental manipulability', but fails to justify itself empirically.

For example, a primarily empirical explanation of the behaviours of gases posits a link between volume, temperature and pressure and postulates that a certain manipulation of temperature results in a certain change in pressure. The explanation rests on the fact that the temperature of a gas can be manipulated experimentally and that its manipulation consistently correlates with certain changes in pressure that would have not occurred, or rather that do not occur spontaneously in the particular experimental setup under investigation. In other words, the correlation is explanatory because

i) it mirrors experimental manipulability and

ii) it is statically relevant.

The empirical correlation does not tell us what are the mechanisms or causal interactions underlying the manipulation of temperature and pressure, nor does it tell us what is the mechanism or causal process linking temperature and pressure.

In contrast, from the standpoint of kinetic models, pressure and temperature are two macroscopic manifestations of the same microscopic processes consisting of colliding air molecules. We can see therefore that kinetic models explain the co-variation of temperature and pressure as being the result of a common cause or mechanism. Note however that a *CM*-style explanation is theoretical and hypothetical. It does not reflect available means to experimentally manipulate the phenomenon and its associated statistical relevance cannot be calculated.

One possible solution to this dilemma is to maintain a connection between the empirical notion of 'experimental manipulability' and the deeper theoretical notion of 'causal mechanism'. For example, by interpreting the laws of dynamics along the lines of experimental regularities holding true of everyday macroscopic phenomena (the fundamental hypothesis) and by further postulating microscopic entities subjected to the same laws (a model-specific hypothesis) in order to derive an empirical law describing the behaviour of gases (the targeted conclusion of the deductive model), three things are achieved

i) an empirical description from the laws of dynamics is derived from the principles of mechanics (the *DN* aspect of the explanation)

ii) a causal mechanism in terms of fundamental physical forces and interactions is postulated (the *CM* aspect), and

iii) further means of experimental control over the explained phenomenon are hypothesised: assuming that we can push around molecules the same way we push macroscopic objects, it is possible to directly manipulate the temperature, pressure and vol-

ume of a gas, as well as the relationships between these macroscopic properties (the mechanism is translated in terms of hypothetical manipulability).[14]

Although the manipulationist interpretation of this particular explanation seems bound to remain hypothetical, most *CM*-style explanations in molecular biology, including Salmon's paradigmatic 'lung cancer' example, are systematically interpreted in terms of experimental manipulation. A typical explanation doesn't just state that mutation $M$ in protein $P$ causes over-proliferation of a certain kind of cells, but also that, assuming we can directly create or remove $M$, we can directly control the occurrence and non-occurrence of the cancer. The concept of 'gene therapy' is based precisely on this overtly manipulationist interpretation of the causal mechanism. From a philosophical standpoint, this also ensures that the 'deeper' metaphysical causal connection can always be translated in the empirical language of 'surface' experimental manipulation open, at least in principle, to empirical verification.

### *6.9 Conceptual Unification*

Following a different lead, Kitcher (1981; 1989) proposes an unificationist account of explanation. The account hinges on the observation that some scientific propositions can be derived from other scientific propositions while the reciprocal is impossible. The latter seem to 'naturally' fit the place of premises, while the former seem to stand 'naturally' in the place of conclusions, hence the suggestion that laws of nature are 'universal' and 'necessary' (i.e., universal and necessary in a deductive scheme relating them to some other laws).

The observation is certainly valid, and the unificationist account tries to make further use of it by claiming that explanation is tantamount to the highest explanatory power, that is, to the

[14] For a more detailed example, see Woodward's 'manipulationist' explanation of motion on an inclined plane (2003 pp. 11-12).

possibility to derive the maximum of consequences from the smallest set of premises. If the derivation in question amounts to a deduction[15], this conception brings us back to Duhem's original insight, according to which a theory is

> *"a system of mathematical propositions, deduced from a small number of principles, which aim to represent as simply, as completely, and as exactly as possible a set of experimental laws."*
>
> (Duhem, 1906 p. 7)

The notable difference is that while Duhem considers explanation and causation to be distinct from deducibility, Kitcher conflates them by arguing that we decide what causes (and explains) what based on the systematization of our beliefs (1989 p. 477).

In some respects, Kitcher's unificationist account captures commonly accepted assumptions about what counts as a scientific explanation, yet in some other respects it is also extremely counterintuitive. On the positive side, it puts back on the table the requirement for generality often neglected by causal accounts. Kitcher's account captures the widespread notion that a scientific explanation should be attached to some general explanatory pattern, such as a theory. Manipulability and causality are certainly important, but many successful scientific explanations, especially in physics, also tend to achieve a greater unity of scientific knowledge by deriving a wealth of predictions from a relatively small set of fundamental laws. Whether unity is an essential feature of explanation might be debatable, but the fact is that many scientific explanations explicitly aim and often succeed in unifying scientific knowledge.

On the negative side, Kitcher's quasi-Kantian approach to causality is at odds with both the actual scientific practice. Intuitively, just because we can derive empirical laws from more general propositions, does not automatically entail that we are granted an increased experimental

[15] Alternatively, Kitcher also envisages multiple physical interpretations of the same set of "*schematic propositions*". For example, he attributes the explanatory power of Mendelian genetics to its ability to apply the same concept of, say, 'dominance', to a wealth of different phenotypes.

control over the explained phenomenon. At any rate, Kitcher's unificationist account fails to make a clear distinction between a theoretically possible explanation and a true, or at very least a partially corroborated one. In addition, Kitcher's solution to the traditional objections against the *DN* account is particularly unintuitive. Kitcher solves the problem of causal asymmetry illustrated in the 'shadow of the flagpole' example by arguing that adding extra instructions about how to calculate the dimensions of objects starting from information about the size of their shadows to the currently available stock of explanatory propositions does not enable us to derive more consequences. Since it is already possible to derive at least the same number of consequences without relying on these additional instructions, they don't have any explanatory value (1989 p. 485). A similar remark applies to various cases of non-explanatory correlations, such as the correlation between the use of the contraceptive pill and the absence of male pregnancy.

The unificationist approach is rather demanding. As Woodward remarks, in order to conclude that the length of the shadow does not explain the height of the flagpole casting it, we must prove that it is possible to derive more consequences from a smaller number of premises and using shorter derivations by ignoring instructions telling us how to infer the dimensions of objects from the dimensions of their shadows. It is not at all clear how anyone could prove this with full mathematical rigour, nor is it in any way evident that this kind of proofs is really required in order to make causal judgements (2003 p. 369).

Equally alarming, the unificationist account denies any explanatory value to older theories having a lesser domain of application than their newer rivals. Again, this is highly counterintuitive. For example, physicists usually consider that the special theory of relativity is explanatory relevant in some particular situations, while the general theory of relativity provides a more universally applicable pattern of explanation (Woodward, 2003 pp. 367-368).

It has also been pointed out that many theories are time-symmetric. For example, given a set of initial conditions, classical mechanics allows us to derive the past, as well as future trajectory of an object (Barnes, 1992). If we accept Kitcher's unificationist account, we have to conclude that the theory explains equally well past space-time positions, as well as future space-time positions open to verification, conclusion that doesn't sit well with the generally accepted notion that, typically, explanations are causally asymmetric.

Finally, I would like to add that another concern relates to the introduction of additional background assumptions, such as model specific hypotheses. If further assumptions need to be added on the side of the premises in order to achieve a derivation under the more modest scope of the *DN* account, it seems reasonable to expect that they will remain likewise irreducible under the wider scope of the unificationist account and therefore accumulate at the top of the logical hierarchy. If this turns out to be the case, then the store of explanatory propositions may proliferate out of control and thus endanger the very notion of unification.

Ultimately, I think that, just like the original *DN* account, Kitcher's unificationist account applies more readily to conceptually possible explanations, but fails to provide an adequate characterisation of what counts as a justified explanation. As noticed previously, it seems that some propositions are more suited to stand on the side of the premises (or *explicans*) than others; conversely, some propositions are more suited to stand on the side of the conclusions (or *explicandum*). This further suggests that some models are possible, while other models are impossible for purely logical reasons and that a logically impossible model cannot be shown to be true or successful for the very simple reason that it cannot be hypothesised in the first place. Thus, it seems correct to conclude that empirical justification is asymmetrically dependent on conceptual possi-

bility. Nevertheless, it can hardly be argued that a conceptually possible model is necessarily or probably true.

By dissociating the issue of justification from the possibility of providing an explanation, it becomes possible to define a more flexible link between explanatory models and theories. Given the total store of explanations, the propositions standing on the side of the conclusion (the *explicandum*) remain what they are, a description of the phenomenon to be explained. In contrast, the propositions standing on the side of the premises (the *explicans*) might eventually form a theory, where a theory is a set of recurring premises common to several models. This approach takes into account Kitcher's requirement that many scientific explanations are grounded by a set of propositions that can be applied in wide variety of situations, but does not restrict explanation to 'the most stringent derivation from the smallest set of premises'.

### *6.10 The Complementarity and Partial Overlap of the Proposed Accounts of Explanation*

Despite their respective critiques, each of the five accounts of explanation presented in the previous sections captures some important characteristics of the modelling practice in science. What further renders even more difficult a choice is the fact that many accounts overlap with each other. I already showed how the statistical account overlaps a manipulationist one.

Given the prevalence of the notion of 'mechanism' and the absence of well-defined laws in newly emerging sciences such as molecular biology (Wimsatt, 1972; Machamer, et al., 2000), some authors argue that intuitive, mechanism-centered approaches to explanation are radically different from deductive, law-based approaches (Waskan, 2006 p. 236). Nevertheless, many empirical laws and correlations reflect a 'black-box' kind of knowledge about how certain actions yield certain effects in the absence of any further understanding of the alleged mechanism under-

lying these laws and correlations. For example, it is possible to actively cause cancer without understanding the underlying mechanism; whether or not we further understand that smoking causes mutations and cumulative mutations result in uncontrolled proliferation, to say that smoking explains cancer is still a practically useful explanation. More so, some laws are deemed fundamental and therefore devoid of any underling mechanism. For example, the first law of dynamics states that applying a force result in acceleration of the bodies onto which the force is applied; in a classical context, there is no underlying mechanism and no further explanation of the mode of action of contact forces. This suggests that the difference between fundamental laws and more complex mechanisms hinges on the possibility of decomposing a 'black-box' empirical correlation into a specific sequence of more primitive correlations.

If this kind of thinking is correct, then there is a considerable overlap between the *DN* and the manipulationist accounts. The *DN* account frames explanations in terms of fundamental laws and 'black-box' empirical correlations while a mechanism-centered account assumes that 'black-box' empirical correlations reduce to combinations of more fundamental modes of actions. For example, the widespread quest for causal mechanisms in the 19$^{th}$ century is at the same time an attempt to derive empirical laws from mechanical theories. Is a kinetic model of air in a rigid cylinder a *DN*-style explanation or a *CM* explanation of the behaviour of air? Presumably, it depends on whether we adopt an instrumentalist or a realist point of view in regard to the unobservables postulated by kinetic explanations.

Even more striking, mechanisms always function in accordance with a set of rules. For example, Mendel's genetic explanation hypothesises a mechanism for heredity, and, at the same time, it is framed as a deductive consequence of a set of 'laws'. On one hand, Mendel hypothesised the existence of 'genetic elements', that is, of particle-like entities, that are transmitted via

semen or pollen, that mix with their female counterparts and that later on segregate in a specific pattern before being transmitted to the offspring. The subsequent development of molecular biology stemmed essentially from the study of the physical making and the mode of action of these 'elements'. On the other hand, the behaviour of the 'genetic particles', most notably their distribution along several generations, is described by probabilistic 'laws'. These 'laws' 'dictate' how the mechanism functions, that is, how the 'genetic elements' segregate following fertilisation. In the context of classical genetics, the mechanism remains essentially a hypothesis, while the 'laws' are essentially descriptive, and therefore readily open to empirical verification; thus, the 'laws' and the deductive character of the explanation is more important and better justified than its causal-mechanistic aspect. During the development of molecular biology, the tables turned, the mechanism becoming more and more important, while the 'laws' of classical genetics degenerated into the obsolete. Note however that molecular mechanisms too 'obey' laws. They obey the 'laws' of chemistry. The RNA polymerase binds DNA in accordance with the 'laws' of chemistry. This is not to say that the polymerase-DNA interaction is a deductive consequence of a set of premises; rather, this means that the interaction is dependent on the concentration of the substrate, their affinity, turnover rate, etc., as described by the 'laws' of chemistry. Taking into account these factors yields predictions about the behaviour of the interaction and, ultimately, about the functioning of the mechanism.

The above clearly indicates that *DN* and *CM* accounts focus on complementary aspects of the same explanation. Likewise, *CM* and manipulationist accounts also overlap. For one thing, in the context of classical mechanics, experimental manipulability is not easily distinguishable from causal mechanisms consisting of collisions and contact forces. In fact, we can easily argue that classical mechanical models are attempts to extend the notion of manipulability, as experienced

in the context of average-size macroscopic objects, to phenomena which are not empirically described in terms of forces, motions, collisions, etc. In addition, and as discussed on a previous occasion, the mechanisms postulated by molecular biology are almost always interpreted in a manipulationist sense, as statements about how the behaviour of a phenomenon can be altered by altering its underlying mechanism.

Finally, scientists don't think of an explanation that fails to subsume a phenomenon under the larger umbrella of a more general theory as being entirely satisfactory. As a matter of fact, all mechanisms are mechanisms of a certain kind. The Mendelian mechanism for heredity is a probabilistic one about mixing particles and then redistributing them. Molecular mechanisms are chemical; they function in virtue of chemical interactions. Physiological mechanisms are classical mechanisms relying on the notion of contact force and the laws 'dictating' its behaviour. And so on. It is impossible to talk of mechanisms simpliciter. A set of rules stating some fundamental ways of action must also be specified. On the other hand, a technologically sterile explanation is also deemed to be profoundly unsatisfactory. Although there are exceptions, typically, scientists think of their explanatory models as being at the same time a matter of deriving already established empirical descriptions and new predictions from more general principles, of uncovering the causal mechanisms underlying a phenomenon and of gaining experimental control ultimately responsible for developing technological applications. This applies just as well to physics as to newer sciences such as molecular biology and empirical psychology.

## CHAPTER 7

# MODELS RELATED TO THEORETICAL EXPLANATIONS

### *7.1 Confirmable Explanatory Models*

After paying due consideration to the various accounts of explanation, I will now concentrate specifically on a version of the *DN* account which, in my opinion, is able to provide a suitable philosophical framework for most investigations concerning the theoretical constraints on hypothesis formation.

Before pursuing any further, note also that the sense in which I use the term 'model' in this section should not be confused with the model theory in mathematics [i.e., as a structures that makes the sentences of a theory true, where a theory is a set of sentences in a formal language (Bell, et al., 1977)]. It is also distinct from the so-called 'models of data' (e.g., fitting the curve, statistical analysis, etc.) introduced by Suppes (1962). Finally, despite some commonalities, I also draw a distinction between theoretical and material models. Set aside models used for

didactic purposes, material models are particular experimental setups, usually reproducing in the lab naturally-occurring phenomena (Hesse, 1966) and that can have an explanatory value in virtue of the similarity they bear with the phenomena they mimic (Giere, 1988). In contrast, theoretical models explicitly aim to subsume a particular phenomenon under the larger umbrella of a more general pattern of explanation. My current investigation concerns exclusively the latter.[16]

Although highly constrictive, the framing of models as more or less rigorous inferences of conclusions from a set of premises is inevitable. As a former molecular oncologist, set aside experimental models and scale models used for didactic purposes, the models I worked with aimed to explain various biological functions and phenomena by appealing to biochemical mechanisms [this aspect of contemporary research in biology was already noted by Wimsatt (1976) and is being currently investigated by several philosophers of biology (Darden, 2006)]. Nevertheless, even though typical thinking in terms of molecular mechanisms does not involve any formal apparatus and rigorous mathematical derivations, inferences must be made. For instance, in order to explain cell proliferation, I may postulate a mechanism whereby protein *X* binds protein *Y* at some point *c* during the cell cycle. In order for my model to explain the proliferation of cells, I must be able to infer that, given my knowledge of molecular biology, if a certain sequence of events happens as described by the putative mechanism, cell proliferation should occur. Thus, the model plus a number of background assumptions serve as necessary and sufficient basis for deriving a number of conclusions. In parallel, for the sake of empirical confirmation, I must also be able to infer some easily verifiable prediction. Thus, given my knowledge of biochemistry and physical chemistry, if *X* binds *Y* at point *c*, I might infer that extracts from cells in *c* subjected electrophoresis must reveal bulkier *XY* complexes not present in control extracts from cells not in *c*.

[16] For a discussion of the various understandings of the term 'model' see Frigg (2006).

Note that I am not conflating models, explanations and confirmation. I simply say that they intersect: many models explain and, in order to have any empirical relevance, they must allow for the inference of predictions. Furthermore, I cannot help noticing that many inferences at play in scientific reasoning are often times straightforward deductions embedded in a realist interpretation under which explanation and prediction overlap.

According to clear-cut formulations of the deductive-nomological account of explanation (*DN*) and hypothetico-deductive method of confirmation (*HD*), the inferences at play in explanation and confirmation are distinct logical or mathematical deductions. Strictly speaking, the requirement for deduction embedded in classical HD/DN accounts is compatible with instrumentalism and does not require that the model explains or represents anything, but simply that it entails some target propositions. Nevertheless, many deductive models are physically interpreted in such a way that deductive bits of scientific reasoning amount to explanations of empirical phenomena. Stated in general terms, if a deductive model entails the target predictions standing on the side of the conclusion, it is hypothesised that the sum total of the propositions used to derive the conclusions in question hold true of the target phenomenon. No such hypotheses are formulated if it is assumed from the very beginning that the model has a purely instrumental use.

Sometimes, hypothesising that certain propositions hold true of a phenomenon postulates new means of experimental control over the phenomenon in question. This is the case, for example, of Newton's gravitational model for planetary motion, which predicts deductively that a body having a sufficient initial speed may be launched into orbit around Earth. The prediction is correct, meaning that Newton's mechanical explanation provides a 'recipe' for physically constructing 'planet-like' motions (i.e., the laws of mechanics are 'laws of nature', as opposed to mathematical propositions used to generate predictions). In other words, realistically interpreted,

Newton's model also specifies the 'mechanism' responsible for physically generating planetary motions.

In other cases, the propositions standing on the side of the premises describe the properties some entities must have in order to explain the target phenomenon. In an attempt to explain genetic inheritance, Mendel hypothesised the existence of physical entities transmitted from parents to offspring and which are responsible for determining the phenotypes of the studied organisms. These entities are transmitted from one generation to the next and determine the phenotype in a peculiar manner, as described by 'Mendel's laws'. Once again, it is worth mentioning that Mendel's genetic explanation constitutes a hypothesis only in reference to the existence of physical particles obeying 'Mendel's laws'. If these laws are not intended to describe the behaviour of physical particles, they retain their instrumental value as algorithms for predicting the target phenotypic distributions, but fail to constitute confirmable or falsifiable hypotheses about external reality. Mendel's explanation gained empirical relevance the moment it was discovered that chromosomes segregate during meiosis in agreement with 'Mendel's first law': there are physical entities playing a role in the determination of the phenotype and these entities are transmitted from parents to offspring as postulated by 'Mendel's first law'.

For Newton and his followers, the gravitational model was at the same time a deduction of conclusions from premises and an explanation of how planetary motion was physically generated as dictated by 'laws of nature'. Likewise, in the minds of most geneticists Mendel's model was and still is a rigorous deduction of predictions from a set of premise-propositions and, at the same time, a partially specified explanatory mechanism whereby chromosomes are paired and segregated during meiosis. I conclude therefore that even if scientists often reason in terms of mechanisms responsible for physically generating certain phenomena, this does mean that mod-

elling in science has nothing to do with deductively deriving propositions form other propositions for the purposes of explanation and confirmation. After all, mechanisms function according to a set of rules specific to the kind of mechanism we are dealing with and without which it would be impossible to talk about the functioning of the mechanism in abstracto, that is, in the absence of a physical, actually functioning mechanism.

### *7.2 Fundamental Hypotheses and the Model-Theory Distinction in Science*

As discussed above, I believe that several key aspects of the practice of modelling in science are captured by a combination between *DN* accounts of explanation (Hempel, et al., 1965; Kitcher, 1989; Salmon, 1989), and some version of the *HD* method [as originally described by Newton and Descartes, or again by Popper's falsificationism (1959; 1965)], which further transforms the premises present in deductive accounts into hypotheses about the phenomenon targeted by the model. The guiding idea is that a set of premises consisting of general laws, model-specific assumption plus statements specifying initial conditions logically or mathematically imply a description of some aspect of the phenomenon under study, where the implication relationship is further given a realist interpretation explaining how the phenomenon is generated and, by the same token, predicting novel means of experimental control.

I already quoted Morrison in reference to the semantic incompatibility problem (Morrison, 2000 p. 49). With that occasion, I noted that the exemplified models belong to the same family of explanations: the guiding idea underlying all kinetic models of gases and heat processes is that these processes reduce to particles interacting mechanically. Compressibility, rarefaction, liquefaction, diffusion, osmosis, Brownian motion, divisibility, heat, chemical experi-

ments, etc., all corroborate to some extent the fundamental hypothesis according to which a gas is composed of particles in motion, or at very least some other version of atomism.

In principle, there is some freedom as to how one might use this information in order to formulate higher-level hypotheses concerning the structure of the phenomenon and the behaviour of the alleged particles of which it is made. However, the preferred fundamental hypothesis states that whatever atomic explanation we may choose in order to model (and explain) a given phenomenon will not contradict the fundamental laws of classical mechanics, such as the law of inertia applied in reference to the 'mechanical atoms' a gas is presumably made of. In other words, all the models considered are models of classical mechanics. This is not to say that the laws and principles of classical mechanics hold true a priori (i.e., by convention) in respect to gas and heat phenomena. Rather, the fundamental hypothesis common to all these models aims to extend the domain of application of classical mechanics to new phenomena; whether the extension is justified or not remains a matter of empirical investigation.

Thus, it can be easily argued that the general propositions of classical mechanics, such as the three laws of dynamics, describe certain particular experimental setups and phenomena like rolling balls, colliding balls, pushing carts or one observer pushing another observer. As fundamental principles of the mechanical theory, these empirical laws are raised at the rank of general principles which all mechanical models must satisfy if they are to be models of classical mechanics. In as much as mechanical models are also meant to describe or assert something about a phenomenon, one of the fundamental hypothesis common to all mechanical models for various phenomena is that the three laws of dynamics hold true of the phenomenon, for instance in the sense that they describe the experimental control one can gain over the phenomenon (von Wright, 1971; Woodward, 2003).

Formulating these correlations as empirical laws such as $F = ma$ relies on further low-level inductions extending the correlation for any force, mass and acceleration magnitudes, where 'any' refers to the arbitrarily assigned domain of real number values (a domain larger than that of actual observations, where the latter is only a subset of rational numbers). Typically, it is only required that the mathematical representation of empirical correlations such as the second law of motion remains empirically adequate in respect to a range of possible measurements. No claims of truth are attached to such representations and the mathematical apparatus behind this representation is not said to have a reality outside the theory; quite on the contrary, the latter is more or less explicitly assumed to be purely instrumental. Since the induction at work here is low-level, that is, no new entities, properties or relationships are posited, mathematical representation of empirical laws is usually not a matter of dispute between realists and antirealists. Note however that, as discussed in the previous chapters, the mathematical representation of empirical knowledge undermines strict verificationism and operationalism. It also undermines strict falsificationism, since a minimal form of induction must be accepted in order to subsume empirical data under a convenient mathematical format.

As fundamental principles of a theory aiming at universality, the laws of classical mechanics are raised to the rank of general propositions which all mechanical models must satisfy if they are to be models of classical mechanics. Conversely, in as much as mechanical models are also meant to describe or assert something about a phenomenon, one of the fundamental hypothesis common to all mechanical models for various phenomena is that the three laws of dynamics, together with a law of addition of scalar and vectorial physical magnitudes, hold true of the phenomenon. Assuming a manipulationist interpretation of the truth relationship, we can further claim that these laws are true in respect to the modelled phenomenon in the sense that they

describe the experimental control one can achieve over the phenomenon (i.e., the desired motion can be achieved as hypothesised by these laws).

### *7.3 Target Empirical Laws*

Many scientific models are designed from the very beginning not only having in mind a set of theoretical propositions which they must satisfy if they are to be models of a given theory, but also with the specific aim to allow the derivation of an already empirically established law starting from the general propositions of the theory. A simple example is Newton's gravitational model for planetary motion. More complex examples are provided by kinetic models for gases or heat phenomena (Morrison, 2000). In both cases, the strategy is the same. On the side of the premises stand the three laws of dynamics. On the side of the conclusion stands a set of empirical laws, such as Kepler's laws or Boyle's law; once again, these laws constitute approximations of actual measurements summarising a great amount of data under a simple mathematical formulation. This indicates that scientific models are designed from the very beginning in such a way that

1) they must satisfy a set of theoretical propositions if they are to be models of a given theory (the explicans); and
2) they allow the derivation of already empirically established laws describing the phenomenon to be modelled (the explicandum).

In the above mentioned examples, classical mechanics constitutes a higher-level theoretical constraint on the formulation of hypotheses concerning the behaviour of gases and heat exchange phenomena. We can see right away that 'theoretical' does not necessarily mean 'conceptual', 'logical' or 'mathematical' since the laws of mechanics are already interpreted, hold true in

respect to some phenomena and experimental setups and are, at their origin, empirical propositions. Instead, it is question here of unifying several domains of investigation in such a way that the laws proper to one domain are extended to a new domain.

### *7.4 Model-Specific Hypotheses*

As Putnam, Lakatos and others pointed out on numerous occasions (Duhem, 1906; Quine, 1951; Lakatos, 1970; Putnam, 1991), it is not always possible to derive the desired empirical law from the laws of the theory alone.[17] In many cases, modelling becomes a constructive problem solving strategy. Assuming that the fundamental hypothesis holds true of the modelled phenomenon, we have to ask ourselves what other assumptions are required in order to derive the target empirical law. At this point, additional, model-specific hypotheses usually need to be introduced, such as hypotheses concerning the shape, size and interactions between gas particles thus yielding models for ideal gases, van der Waals gases, etc. From the conjunction of the two sets of hypotheses, the general ones belonging to the theory and the particular ones belonging to each individual model, predictions, such as positions, correlations between observables or laws, again involving observables and observable relationships between observables, are derived.

Assuming that the atoms are quasi-infinitesimal, non-interacting, perfectly elastic little balls, then Boyle's law should hold true of them, as postulated by the ideal-gas model. Or again, assuming that the atoms are weakly interacting spheres of non-negligible radius, then the laws hypothesised by the van der Waals model should hold true. In as much as the model-predicted laws match or approximate empirically established laws, it is inferred that if the principles of

[17] Of particular interest is Kitcher's solution to the problem of the 'underdetermination of theories by empirical evidence' (Quine, 1975; Laudan, 1996). Once the distinction between a fundamental hypothesis ('core theory') and model-specific hypotheses ('background assumptions', 'auxiliary hypotheses', etc.) is made, it becomes possible to argue that even if two theories are empirically equivalent, when supplemented with distinct model-specific hypotheses, they yield distinct predictions about observable phenomena (Kitcher, 1982).

classical mechanics hold true about the studied gas and for the given temperature and pressure conditions under which the empirical laws holds true, then the gas in question is empirically indistinguishable from or empirically equivalent to a collection of point-like particles in random, elastic collision, or again indistinguishable from and equivalent to a collection of 'sticky balls', etc.

### *7.5 Some Remarks Concerning Common Post-Positivist Concerns*

I don't think that the above treatment of the modelling practice will impress anyone as being radically novel and osé. Nevertheless I have to point out that it has the advantage of preserving the simplicity of traditional deductive accounts, while being flexible enough to accommodate some common post-positivist concerns. Most notably, acknowledging the necessity of model-specific hypotheses is tantamount to acknowledging at least a partial autonomy of models in respect to the theory. In reaction to the positivist, or syntactic view, which reduce models to physical interpretations of axiomatised theories (Campbell, 1920; Nagel, 1961), adepts of the semantic view claim that theories are collections of related models (van Fraassen, 1980; Giere, 1988). A more recent tendency is to argue out that it is reasonable to see models as essentially independent of theories. At the very extreme of this tendency, Cartwright discusses examples of models commonly used in engineering and experimental physics, yet not attached to any theory in particular (Cartwright, 1983; Cartwright, et al., 1995) and argues that theories do not provide efficient templates for constructing models (Cartwright, 1999). Following a different lead, Darden and Maull (Darden, 2006 pp. 128, 130-132) argue that the traditional positivist view according to which theories consists of a set of axioms used in the derivation of prediction-propositions should be replaced by the more flexible notion of 'field', where a field comprises a number of

propositions consists of a problem, presumably relevant data and associated experimental techniques used to acquire this data, and explanatory strategies revealing how the relevant data can be used towards providing a solution to the problem. The authors further argue that in order to provide the solution for any given problem, researchers may use data, techniques and explanatory strategies belonging to distinct fields of investigation, thus generating 'interfield' theories relying on an extensive body of background knowledge, assumptions and explanatory strategies (2006 pp. 132-143).

The 'fundamental vs. model-specific hypotheses' approach which I advocate links models to theories and treats them as attempts to extend the domain of application of pre-existing explanations – usually amounting to theories, that is, to general, and ideally universal patterns of explanations – to new phenomena and domains of investigation. On one hand, my approach retains some similarity with the positivist view, in the sense that different models represent attempts to apply the same theory to different phenomena, thus preserving an important role for theories in the development of models. On the other hand, since the internal resources of a theory must often be upgraded via additional hypotheses external to the theory in order to explain new phenomena, models should be regarded as partially independent constructions. This remark joins Morrison's (1999) claim that models are autonomous from theories in respect to their construction and to their role in the overall economy of scientific knowledge, as well as Redhead's (1980) observation that, often times, theories are incompletely specified and require additional, model-specific constraints before they can be applied to various situations.

My account is also partially compatible with Darden's 'field' approach. Since each model amounts to an attempt to extend the domain of application of existing explanatory strategies to new phenomena and since, as a general rule, such extensions require the introduction of model-

specific hypotheses external to the theory, the formulation of models must rely on some form of 'out-of-field' thinking. Even more striking, typical fields of investigation in biology (such as cytology and Mendelian genetics) can be characterised as being bodies of knowledge about a certain target phenomenon (the 'bits and pieces' living organisms are made of and sexual reproduction) obtained via the use of a specific experimental technique (microscopy & associated staining techniques and, respectively, artificial breeding) and eventually subsumed under the umbrella of a higher-level theoretical or explanatory hypothesis supported by the available data and entailing novel predictions about the target phenomenon (the cell is the smallest functional unit of life, which is a theoretical generalisation of the fact that all the studied organisms were made of cells and, in the case of genetics, Mendel's allele explanation). It seems therefore that, minimally, a field is a body of data acquired via certain technique of investigation; what I call a model is nothing else than an 'intrafield' explanation. In as much as it is possible to extend such an 'intrafield' explanation to a new domain of investigation consisting of seemingly similar or radically different phenomena without dropping any of its constituent hypotheses – although this may not exclude the addition of new, model-specific hypotheses – the model is officially upgraded to the rank of theory. Furthermore, the above understanding of the model-theory distinction in science is compatible with Darden's suggestion that new theories are developed by importing and adapting explanatory strategies already present in other fields of investigation (2006 pp. 150-151). Alternatively, nothing prohibits the generation of parallel 'interfield' explanations which may or may not be consistent with 'intrafield' explanations in respect to entailed predictions, necessary assumptions or postulated entities. In as much as 'intrafield' explanations are consistent with the 'interfield' explanation, theories can be delimited as a set of recurring hypotheses common to both 'intra-' and 'inter-field' explanations. Obviously, in neither case the fundamental hypothe-

ses associated with the theory are necessarily axioms; quite on the contrary, they typically do not provide sufficient theoretical resources for the derivation of all the propositions entailed by the various models attached to the theory. Nevertheless, they do provide an initial body of premises, that is, the incomplete draft of a deductive argument which, once complemented by model-specific hypotheses, allows for the derivation of target empirical laws responsible for the empirical adequacy and the explanatory value of the model, as well as novel predictions eventually submitted to empirical testing.

# CHAPTER 8

# THE CHALLENGES OF CONFIRMATION HOLISM

## *8.1 Confirmation Holism*

Since model-specific hypotheses are required in order to derive the target empirical law from the fundamental hypothesis, confirmation (or corroboration) becomes holistic. But if confirmation is holistic, then it is not clear what is targeted by confirmation, the fundamental hypothesis or the specific hypotheses introduced by the model (Putnam, 1991).

Given this ambiguity, Lakatos (1970) further argues that a theory can always 'digest' its counterevidence by producing more complex models, correcting inadequacies by adding more and more assumptions, until one model is finally verified. Lakatos's argument goes as follows: Imagine two true propositions. One is the fundamental hypothesis, which we take to be true a priori because we want to save the theory; the other, is an empirically established law, correlation, co-variation, etc, which is true in virtue of its empirical status. Now, given these two propo-

sitions, where the first one belongs among the premises and the second is the conclusion, we ask ourselves what other propositions are required in order to bridge premises and conclusion. It seems that the game is rigged. In as much as we can devise a suitable set of bridging auxiliary assumptions, we are bound to win. In extreme cases we can choose to defend a theory by postulating certain unobservable or difficult to confirm facts via auxiliary assumptions. Thus, Lakatos concludes, whether we keep or abandon the theory has nothing to do with confirmation and falsification, but rather with our determination to save the theory. Typically, we begin by defending a theory, but, if in the long run the hypothesised background assumptions are not verified, new theories are developed and the old one is eventually abandoned [for a discussion of the specific example of Newton's theory of gravitation, see Lakatos (1970 pp. 125, 133-8)].

Ultimately, this conclusion reiterates by means of particular examples the general argument for verification holism originally formulated by Duhem:

> "*The physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses.*"
>
> (1906 p. 187)
>
> "*The only thing the experiment teaches us is that, among all the propositions used to predict the phenomenon and to verify that it has not been produced, there is at least one error; but where the error lies is just what the experiment does not tell us.*"
>
> (1906 p. 185)

Picking up on this central idea, Quine frames a more radical form of holism: not only we don't know where the error lies, but, in principle, we could save any proposition within a theory by making enough adjustments elsewhere in the theory. He famously illustrates some of the most extreme consequences of holism by means of a rather compelling metaphor:

"*the totality of our so-called knowledge* [...] *is a man-made fabric which impinges on experience only on the edges.* [...] *A conflict with experience at the periphery occasions readjustments in the interior of the field.* [...] *But the total field is so undetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to re-evaluate in the light of any single contrary experience.* [...] *Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system.*"

(1951 pp. 39-40)

### *8.2 Newton's Gravitational Model of Planetary Motion*

In order to better understand the problem, and a possible solution, it is becomes profitable to introduce first a simple example that illustrates it. The example I have in mind is Newton's gravitational model of planetary motion.

The fundamental hypothesis underlying Newton's explanation states that any model of planetary motion should be a mechanical one. In conformity with at least the first two laws of dynamics, this means that the model must yield the unique spatial and temporal distribution of forces responsible for deflecting what would have otherwise been a uniform rectilinear motion of planets into the closed paths observed by astronomers. This unique distribution of forces is common to all mechanical models for planetary motion and will not change from one model to another. Any model that fails to comply with this fundamental requirement cannot be a model of classical mechanics.

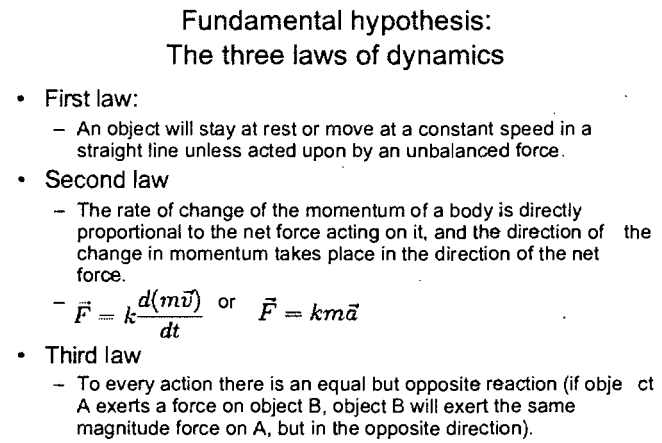



**Figure 10. The Three Laws of Dynamics**

Note that from the standpoint of the theory, it doesn't matter how the distribution of forces is obtained. Each model will offer a different solution. One model may postulate attractive forces acting at a distance, while alternative models may appeal local contact forces, such as friction, jet propulsion, angels pushing the planets, etc.

The target empirical laws responsible for the empirical adequacy of the model are Kepler's laws. Assuming that the Copernican interpretation stands true, these laws subsume in a suitable mathematical form a great number of astronomical observations about the space-time positions of planets. Any model, mechanical or otherwise, that fails to approximate these laws cannot be a true model of planetary motion because it would fail to be empirically adequate in the first place. The figure below provides a quick review of the three laws which made Kepler famous:

Empirical adequacy constraints:
Kepler's laws

- **The first law** : The orbit of every planet is an ellipse with the sun at one of the foci.
- **The second law**: A line joining a planet and the sun sweeps out equal areas during equal intervals of time.
- **The third law** : The squares of the orbital periods ($P$) of planets are directly proportional to the cubes of the semi-major axis ($a$) of the orbits. ($P^2 \propto a^3$)

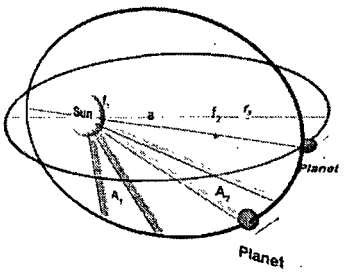



Figure 11. Kepler's Laws

Aside these two points of commonality between all empirically adequate mechanical models of planetary motion, each individual model adds its own, model-specific hypotheses. As holists like to point out, in many cases the theory fails to provide all the propositions required for deriving predictions. We have already seen that in the case of kinetic models additional propositions are supplied as model-specific hypotheses. Given the presence of these extra hypotheses, Duhem (1906), Quine (1951), Lakatos (1970), Putnam (1991) and many others argue that a scientific a model can never be derived from the theory alone, but a number of 'background assumptions' must be taken into account.

In a now classical analysis of Newton's gravitational explanation of planetary motion, Putnam makes the following case:

> "*What do we do, then, when we apply this theory to an astronomical situation? Typically we make certain simplifying assumptions. For example, if we are deducing the orbit of the earth we might assume as a first approximation:*
>
> *(I) No bodies exist except the sun and the earth.*
>
> *(II) The sun and the earth exist in a hard vacuum.*

*(III) The sun and the earth are subject to no forces except mutually induced gravitational forces.*

*From the conjunction of the theory of universal gravitation (U.G.) and these auxiliary statements (A.S.) we can, indeed, deduce certain predictions – e.g., Kepler's laws.* [...] *But it is important to note that these predictions do not come from the theory alone, but from the conjunction of the theory with A.S.*"

(1991 p. 124)

We retain therefore a number of idealisations, to be discussed later on, and a core model-specific hypothesis, namely Newton's inverse square for gravitational forces. This law is further required to bridge the fundamental hypothesis (the premise) and Kepler's laws (the conclusion).

The figure bellow summarises the alleged path of reasoning which led Newton to postulate the inverse square law as a necessary model-specific hypothesis required for the derivation of Kepler's laws from classical mechanics:

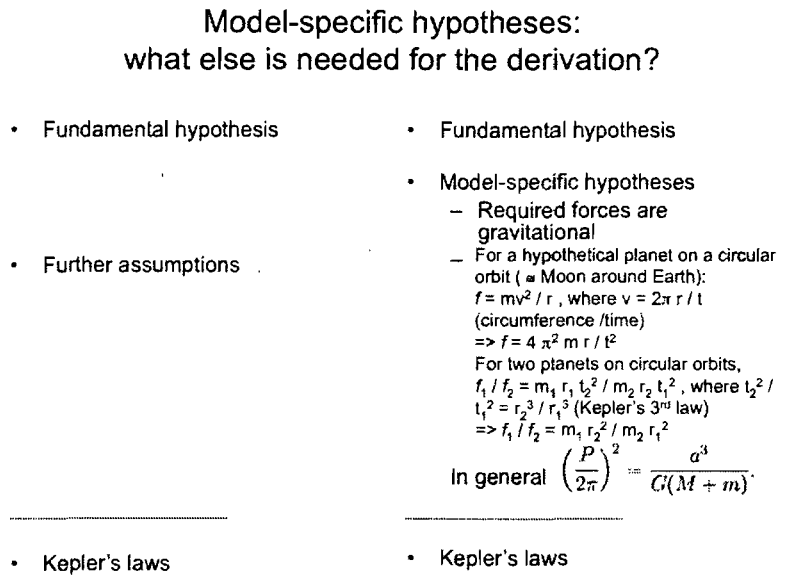



**Figure 12. Auxiliary Assumptions**

Due to its length, I will not reproduce the derivation in full mathematical detail. The reader can easily find several proofs in the literature, as well as simplified versions in introductory textbooks. For Newton's initial insights, see his *Principia*, Proposition XI (Newton, 1687). For a derivation of Newton's inverse square law from Kepler's laws, see Michels (1973). For a derivation of Kepler's laws from the inverse square law, see Hyman (1993). The figure below provides a quick overview of the main steps of the derivation:

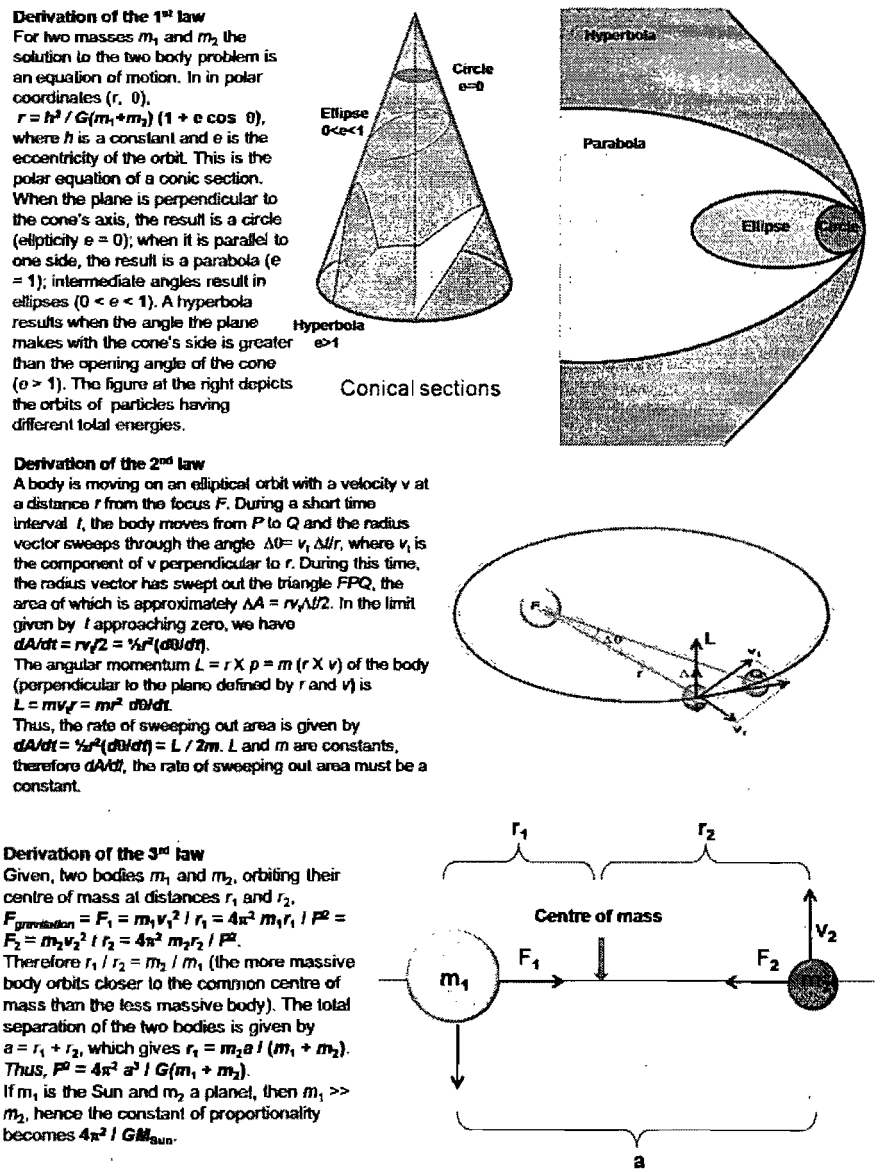



Figure 13. Mathematical Derivation of Kepler's Laws

### *8.3 Alternate Mechanical Models of Planetary Motion*

It is worth noting that different model-specific hypotheses bridge premises and conclusions for non-gravitational models. For instance, Putnam suggests alternate models in which a friction medium or a different kind of force is responsible for the acceleration of planets:

> "*When the predictions about the orbit of Uranus that were made on the basis of the theory of universal gravitation and the assumption that the known planets were all there turned out to be wrong, Leverrier in France and Adams in England simultaneously predicted that there must be another planet. In fact, this planet was discovered – it was Neptune. Had this modification on the A.S. not been successful, still others might have been tried – e.g., postulating a medium through which the planets are moving, instead of a hard vacuum or postulating significant non gravitational forces.*"
>
> (1991 p. 125)

Likewise, a different bridging law is attached to the friction model telling us how friction forces vary along the path of a planet in order to yield the observed orbit, thus revealing something about the hypothesised medium responsible for friction. Or, assuming that the planets move by themselves, yet a different law will state that there are no forces acting anywhere except on the planets such that, if launched into outer space, a particle will simply continue its eternal, straight-line inertial motion, undisturbed by anything else than an eventual collision with some other body. And so on.

### *8.4 Direct Confirmation of Model-Specific Hypotheses*

As Putnam argues in the above quote, in the event of falsification we can choose to abandon the fundamental hypothesis or choose to save the fundamental hypothesis by conjoining it with an alternate model-specific hypothesis. The latter option is illustrated by the alternative

'friction-medium' and 'self-propulsion' models of planetary motion. It follows from here that the falsification of a mechanical model does not entail the falsification of the fundamental hypothesis. This is indeed a very significant result since it constitutes a counterexample to Popper's claim that it is easier to falsify theories than justify them. In this case, it seems that the falsification of classical mechanics would require the falsification of all possible mechanical models of planetary motion. Nevertheless, an argument against Popper's anti-inductionism is not what retains Putnam's attention. As stated in the introduction of the chapter, Putnam is more concerned by the fact that it is not clear what is targeted by confirmation, the fundamental hypothesis or the specific hypotheses introduced by the model.

Is there a solution to this problem? I think there is, if not a general solution, then at least particular, case by case solution. The confirmation of any mechanical model of planetary motion can be achieved on two distinct levels. The first kind of confirmation is inferential and relies on modest extrapolations; the second is experimental and allows the direct testing of some elements of the model.

The first level of confirmation is achieved by associating the phenomena responsible for the distribution of forces with the relevant observations. For instance, if the force distribution is achieved via gravitational forces acting at a distance, then neighbouring massive bodies must be present; alternatively, if the forces are frictional, then an significant interplanetary medium must be present; if gaseous jets propel the planets, then the jets or volcanoes must be observed; and so on. At this level, confirmation relies on the extrapolation of low-level correlations. For example, in order for the discovery of Neptune to count as a confirmation of Newton's gravitational model, we must assume that since all objects on earth have a mass and since all masses attract each other on earth (we can consider here Lord Cavendish's experiments), the same holds true

for celestial objects as well. The risk associated with these extrapolations is comparable to those associated with any low-level induction.

The second level of confirmation is open to active experimentation. Since all models must yield an identical space-time distribution of forces along the observed planetary motions, it is pointless to argue that they must be true because they explain the motion of planets. New predictions must also be confirmed.[18] Most notably, a true model must also predict the behaviour of a test particle or body outside any known planetary orbit. For example, it must predict the distribution of forces responsible for shaping the motion of an artificial satellite, of a comet penetrating the Solar System or of a self-propelled spacecraft. In as much as empirical measurements fit the predicted trajectory predicted by a model, that particular model is corroborated while the other models are falsified.

The second level of confirmation targets directly and individually the model-specific laws giving the distribution of forces required to obtain the observed planetary motions independently of any correlation with further observations of corroborating phenomena such as the discovery of a new planet capable of exerting gravitational forces or the presence of star dust responsible for friction. For example, given the perturbations in the orbit of Uranus, even if Neptune were not discovered – let's say it reflects very little of the solar radiation it receives –, if a comet penetrating the Solar System or if a manmade spacecraft exiting it are suddenly deflected from their path in a way that correlates with the hypothesised gravitational forces responsible for the perturbations of the orbit of Uranus, but not with Putnam's hypothesised presence of a medium responsible for creating friction forces, then Newton's gravitational model of planetary mo-

[18] Initially, Lakatos argued that a prediction is 'novel' if the predicted phenomenon was never observed prior to the prediction (Lakatos, 1970). Later on, he deemed a prediction to be 'novel' if it is not among the problems or phenomena which the theory to which the prediction belongs initially aimed to solve or account for (Lakatos, et al., 1976). My distinction between the empirically adequate conclusions aimed at by the deductive structure of a model and the further predictions following from it matches the latter definition.

tion receives further confirmation while Putnam's friction model fails to do so (presumably, the two models can be clearly distinguished since Sun exerts gravitational forces throughout the Solar System while a medium can exert frictional forces only locally; more so, gravitational acceleration, and therefore resulting motion, is the same for all masses, while the same is not true for friction). We may continue to doubt that there really is a planet responsible for the deflection, yet we will have no other choice than to agree that, with or without a planet to cause it, something empirically indistinguishable from a gravitational field of forces is really present out there.

### *8.5 Direct vs. Holistic Confirmation*

Although a theory does not favour any of its models, direct confirmation of model-specific hypotheses corroborates some models while eliminating other models, which are bound to remain mere theoretical alternatives. In contrast to model-specific hypotheses, the fundamental hypothesis – which amounts to the extension of the theory to a new phenomenon – cannot be confirmed independently of some mechanical model of planetary motion. The latter is confirmed holistically, along with the model of which it is a constitutive part.

The figure below enumerates the various ways a model might receive empirical justification:

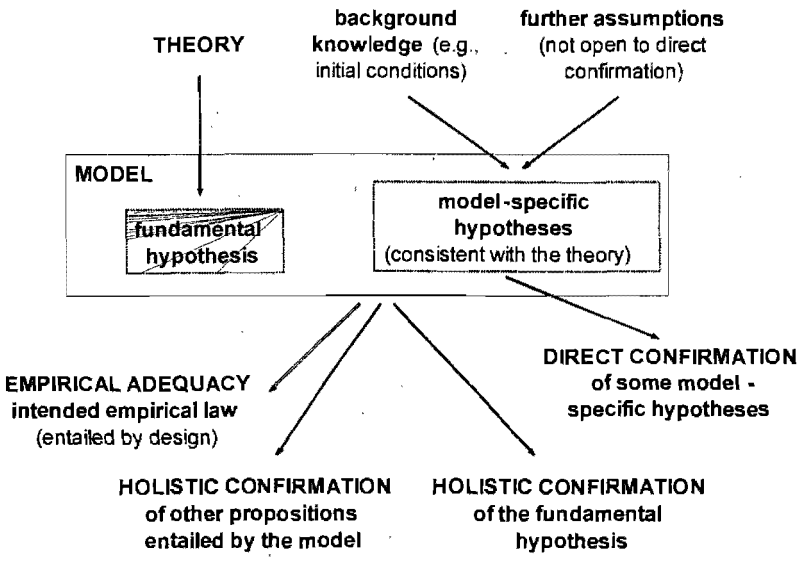



Figure 14. Direct and Holistic Confirmation

### *8.6 Falsification Conditions*

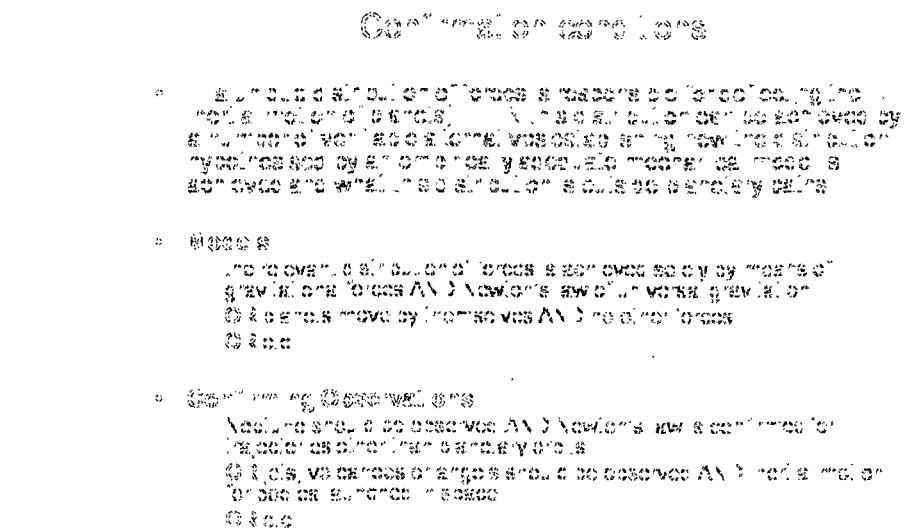
It is also possible to establish falsification conditions. Putnam and Lakatos rightly point out that the falsification of a model of a theory should be distinguished from the falsification of the theory itself (the falsification of the fundamental hypothesis). However, they erroneously conclude that a theory cannot be falsified. Their argument is based on the presupposition that, in order to falsify a theory, we must show that all its models are false; this cannot be easily done, as it is not clear if someone could be even in position to enumerate all the models of the theory in a first place.

Nevertheless, in this case, and presumably in other cases as well, falsification conditions can be established without having to eliminate one by one all the models of the theory. For example, if the inverse square law is verified on an individual basis, then it would take classical mechanics to be fundamentally wrong about planetary motion in order to obtain a true conjunction of the principles of classical mechanics as established by local experiments, Newton's in-

verse law, the right distribution of massive objects and the presence of frictional forces, as demonstrated by the existence of a significantly dense inter-planetary medium. If such a conjunction of observations were ever shown to be true empirically, it would mean that the law of addition of forces holds true on Earth, but not in Heavens. Then the fundamental hypothesis would be explicitly and unambiguously falsified, classical mechanics shown to be false about planetary motion and, as an immediate and unavoidable consequence, all mechanical models for planetary motion falsified.

### *8.7 Dissolving the Holist Conundrum*

In summary, the following confirmation and falsification strategies apply to mechanical models of planetary motion:



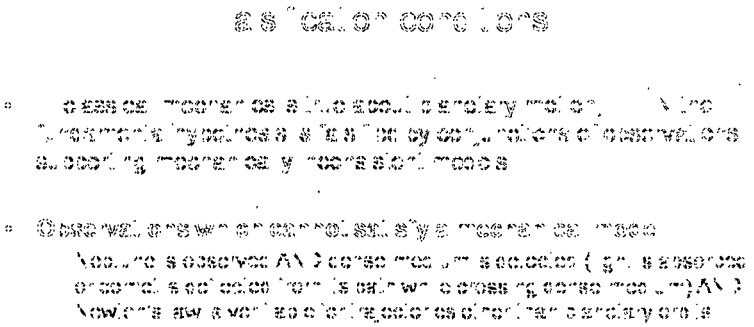


**Figure 15. Confirmation and Falsification Conditions**

There are several points of divergence between the above outlined path of scientific reasoning and the Putnam-Lakatos version of it. First, there is a fundamental hypothesis common to all mechanical models not stated by Putnam and only tangentially alluded to by Lakatos. This is unfortunate, because this essential premise indicates that all explanations of planetary motions, including Newton's model and Putnam's suggestions of alternate explanations, must be statements about forces. The holist argument relies on the implicit assumption that the distance separating the premise of the fundamental hypothesis from the desired conclusion could be bridged, in principle at least, by any kind of propositions, introducing as many unobservables and purely theoretical explanatory principles we wish. This is not so. The propositions mediating the link must be statements about forces.

This realisation leads to a second point. Statements about forces are not purely arbitrary, but link to our previous knowledge of forces. The assumption here is that the same kind of mechanical forces with which we might be concerned in our daily life shape the path of the planets. Among other things, our current knowledge about forces includes laws stating how forces cause changes in motion and a collection of correlations between different forces and phenomena (e.g., gravitational forces are associated with the presence of massive bodies, friction forces with media for motion, etc.). This initial knowledge about what causes and what kind of phenomena are

associated with a given kind of forces is extrapolated to planetary motion. The extrapolation is not as gratuitous as it may seem. It states that the motion of a body in the Solar System can be controlled in order to travel to a desired destination precisely because it is a question of countering naturally occurring forces with artificially produced ones. This assumption contains the seeds of potential technological applications whose eventual success will directly contribute towards the confirmation of the relevant model.

Third, Putnam assumes that the three 'background assumptions' listed in his version of the analysis are independent of Newton's inverse square law, meaning that they could be replaced by a different set of assumptions. It can be shown that they are not independent and therefore they cannot be replaced by any other set of assumptions. Assumption (II) "*The sun and the earth exist in a hard vacuum*" and (III) "*The sun and the earth are subject to no forces except mutually induced gravitational forces*" are inseparable from the gravitational model. Within the limits of empirical observation, planetary motion is eternally repetitive and seems to conserve itself ad infinitum. One way to construct this kind of motion starting from the three laws of dynamics is to hypothesise that the planets undergo an inertial motion constantly deflective by a force acting perpendicularly on the direction of the motion, hence the 'rock on a string' model for planetary motion proposed by Newton. This precludes forces acting in the direction of motion. In particular, this precludes friction forces and therefore the existence of a medium such as air.

The fourth point of divergence pertains to the fact that not any set of propositions about forces is consistent with classical mechanics. The inverse square law is consistent with classical mechanics only if we assume that planetary motion is shaped by gravitational forces alone. Classical mechanics is deterministic. If gravitational forces suffice to obtain the distribution of forces

required to preserve empirical adequacy, then other kinds of forces can be present only if we either modify the inverse square law or, if this law is confirmed on an individual basis, by modifying classical mechanics. Thus, Newton's inverse square law is consistent with classical mechanics only if certain other constraints are satisfied. This clearly indicates that the theory posits a constraint on the sets of propositions which may count as its models.

The confirmation of Newton's law constitutes a fifth point of divergence. Putnam explicitly assumes that Newton's law for universal gravitation cannot be confirmed directly and/or independently. This is not so. The inverse square law can be empirically verified for objects travelling on trajectories other than planetary orbits. It is certainly true that the distribution of forces postulated by the fundamental hypothesis is set in such a way as to preserve empirical adequacy; however, the empirical adequacy in question concerns solely already observed planetary orbits used as an empirical constraint on the formulation of the model. Each model-specific law further hypothesises a particular distribution of forces outside planetary orbits; this distribution can be used to predict trajectories for bodies like comets, artificial satellites and spacecrafts.

This leads us to a sixth and most important point. If the inverse square law is confirmed, then it would take classical mechanics to be fundamentally wrong about planetary motion in order to have a true (confirmed in actual experience) conjunction of the principles of classical mechanics as established by local experiments (the three laws of dynamics), Newton's inverse law, the right distribution of massive objects and the presence of frictional forces. Thus, despite Putnam's and Lakatos's claims, there is a point when the theory cannot be saved by appending new explanatory clauses and its most fundamental principles must be revised.

### *8.8 Concluding Remarks*

Although significantly more detailed than Putnam's, the above analysis is still incomplete. For instance, when postulating the existence of gravitational forces, Newton assumes an unequivocal assignment of the active and reactive forces; this assumption was eventually challenged by Einstein's equivalence principle. Nevertheless, my aim is not to support Newton's views on gravitation, but to defend the hypothetico-deductive (*HD*) method against the objections raised by Putnam and Lakatos. The above analysis suffices to show that confirmation holism does not necessarily entail that theories cannot be falsified, or worse, validate a jump from Popper's epistemological (or allegedly 'naïve') falsificationism to Lakatos's methodological falsificationism, which in the end seems to suggest a conventionalist epistemology fundamentally incompatible with the requirements of empirical science.

Since the criteria of confirmation are different from the criteria of falsification – in the above example the latter being more demanding than the former – it is reasonable to conclude that a single crucial experiment cannot decide between the truth and falsity of a theory in respect to a given phenomenon. In this sense, Lakatos's point against "*instant rationality*" is well taken. On the other hand however, it is erroneous to conclude that scientists are forever stuck in a vicious circle of a reasoning they can escape only by making an arbitrary choice to either pursue the defence of a theory at all costs or abandon it in favour of some other theoretical option. There are confirmation and falsification criteria, some of which are open for testing at the time the theory and its relevant models are proposed. Knowing that at least some of these criteria are satisfied provides an empirically-justified impetus to continue to work on the premise of a so far plausible theory or to abandon a so far implausible one.

Quine (1975) himself admits that the holistic scenario whereby a theory can be saved no matter what is just a possibility, not something necessarily true of science in general. Based on the initial argument by Quine, Newton-Smith reminds us that

> "*this holistic assumption that* [...] *any aspect of a theory can be maintained by making suitable adjustments elsewhere, is question begging in the context of discussions of underdetermination. For, unless we already assume underdetermination, there is no reason to think that* [...] *scientists can pull off this trick. What we find in practice is that sooner or later one avenue becomes blocked.*"
>
> (2000 p. 535)

# CHAPTER 9

# MODELS AS IDEALISED REPRESENTATIONS

## *9.1 Introduction*

Different philosophical accounts of representation tend to apply better to particular kinds of models. Scale models and diagrammatic representations, especially those used for didactic purposes, are literally meant to function as illustrations or simplified reproductions of the target phenomena. It is tacitly assumed by teachers and students alike that inferences about their respective targets should be drawn in light of some perceptual analogy. In contrast, instrumental or instrumentally-used models such as computer simulations and mathematical models are used primarily as means of surrogate reasoning. Such models are not thought to resemble target phenomena beyond their similarity at the level of inferred prediction-propositions.

In this chapter, I argue that there is also a third, very common kind of models that serve primarily an explanatory purpose largely responsible for their technological and experimental

relevance: I have in mind common scientific explanations framed in a hypothetico-deductive format. Neither inferential, nor resemblance accounts do full justice to this kind of models. Instead of passively describing the phenomenon, as postulated by resemblance accounts, or merely allow for surrogate reasoning, as postulated by inferential account, such models assess the influence of a number of factors that have a direct or indirect incidence on its manifestation. In this sense, they provide a 'dynamic representation' best captured by experimental and manipulationist accounts of explanation.

### *9.2 The Resemblance and Inferential Accounts of Scientific Representation*

If hypotheses are meant to assert something about the target phenomena, it seems reasonably to assume that the model amounts to some kind of representation of the said phenomena. In the remaining of the paper I will compare the above approach to modelling with rival accounts in respect to the specific issue of representation.

Semantic approaches to the modelling practice, such as van Fraassen's isomorphism account (van Fraassen, 1980; 2002) and Giere's similarity account (Giere, 1988; 2004), explicitly endow models with direct representational power whereby model and target phenomenon resemble each other.[19] Unlike linguistic denotation, which is a matter of arbitrary stipulation, similarity

[19] Using van Fraassen's seven-point geometry example (van Fraassen, 1989 pp. 218-220), Giere summarises the difference between positivist, or syntactic (Campbell, 1920), accounts and semantic accounts as follows:

> "*On the classical view* [...] *a theory is (i) a set of uninterpreted axioms in a specified formal language plus (ii) a set of correspondence rules that provide a partial empirical interpretation in terms of observable entities and processes. A theory is true if and only if the interpreted axioms are all true.* [...] *A semantic approach requires looking at the axioms* [...] *a little differently.* [...] *Rather than regarding them as free-standing statements, consider them to be part of a theoretical definition, a definition of seven-point geometry. The definition could be formulated as follows: Any set of points and lines constitutes a seven-point geometry if and only if A1, A2, and A3* [A1, A2 and A3 being the axioms of seven-point geometry]. *Since a definition makes no claims about anything and is not even a candidate for truth or falsity, one can hardly identify a theory with a definition. But claims to the effect that various things satisfy the definition may be true or false of the world. Call these claims theoretical hy-*

& isomorphism are thought to provide more or less accurate 'mirror images' of the target phenomena. And, in contrast to inferential approaches to the modelling practice in science (Hughes, 1997; Suárez, 2004), similarity & isomorphism allow for a direct mode of representation whereby the representation relationship concerns solely the model and its target.[20]

Stripped from a technical description of the notion of resemblance – such as isomorphism and similarity, both subjected to heavy attack on the grounds that they fail to display the logical properties of representation –, the core intuition behind them is fairly easy to grasp:

> "*Models are relata of analogy relations; that is, a model is an analogue.* [...] *For example, DNA models built of painted balls and metal struts are positively analogous to DNA molecules in spatial structure and connectedness, but negatively analogous in size, material, shape, and colour of the constituents, etc. These models have a neutral analogy with molecules insofar as their further detailed properties are used to explore as yet unknown features of genetic materials. The dividing line between these three sorts of analogy will of course shift as research goes forward – the better the model, the more of the neutral analogy will eventually be accepted as positive, whereas a poor model will become more and more negatively analogous. Models served* [...] *to introduce unobservable entities and processes by analogy with familiar observable entities and processes, thus providing pictures of the explanatory entities held to underlie phenomena.* [...] *Realists held that successful models are*

---

*potheses. So we may say that, on the semantic approach, a theory consists of (i) a theoretical definition plus (ii) a number of theoretical hypotheses.* " (Giere, 2000 p. 519)

Thus, we may say the model is 'true' in respect to the theory, although the 'truth' in question here has nothing to do with empirical (or correspondence) truth, but with the satisfaction of a set of propositions. Empirical truth is replaced by the notion of isomorphism (van Fraassen, 1980; 1989), whereby a derived consequence of the theory is indistinguishable from an empirical law, or again by the notion of similarity (Hesse, 1966; Giere, 1988; 2004; Teller, 2001).

[20] As a side corollary, it further follows from here that representation exists independently of the scientist's judgements and intentions, a conclusion deemed problematic by some authors (Suárez, 2003; Frigg, 2006). It seems however that neither Giere, nor van Fraassen explicitly favours a naturalised account of representation. On the contrary, they seem to argue that similarity, isomorphism and other forms of resemblance are common means of representation rather than sufficient conditions for representation. Thus given a certain phenomenon a scientist aims to study, he or she may propose a suitable model in virtue of some resemblance relationship (Suárez, 2003 p. 230; Chakravartty, forthcoming).

*positive analogues of the real world; positivists denied the reality of the theoretical entities referred to, and regarded models as working pictures to be dispensed with in accepted theories, having at best a formal analogy with the world.* "

(Hesse, 1966 pp. 299-300)

The guiding idea is highly intuitive and seems to be adequate in a number of cases. To use Hesse's own example, although the pictures of nucleic acids one can find in biochemistry textbooks are idealised and highly simplified schematic representations, this doesn't mean that they are to actual DNA or RNA molecules what a rose might be in respect to love or the symbol of a number in respect to the mathematical number. Such diagrams are neither denotational, nor metaphorical. The scientists who wrote the textbooks and the students who learn from them take these pictures to literally mirror geometrical aspects of macromolecules, as opposed to arbitrary interpretations of a formal language or as intuitive aids for grasping some underlying complex concept. In short, the diagrams resemble their targets and it is in this precise sense that they are thought to represent.

At the other end of the spectrum, instrumentally-interpreted models are suitable illustrations of the inferential approach to representation. Elaborating on Hertz's views on scientific theories, Hughes argues that a representation consists of three parts: the denotation of phenomena and their properties by means of variables and other devices such as equations, mathematical functions, diagrams, etc; the demonstration of the "*dynamical consequences*" of the representation, such as the derivation of predictions; and the physical interpretation of these consequences in terms of empirical phenomena and properties (Hughes, 1997).

Historically, some of Hertz's insights on scientific representation were further developed by Mach (1893), and later on by logical positivists (Menger, 1979). One of the core realisations that shaped the development of logical positivism was the Humean notion that many empirical

laws are mere correlations. As correlations, they do not point to any specific underlying mechanism, causal connection or fundamental mode of action (in general, they don't favour any particular physical, realist or ontological interpretation of the correlation). Rather, they represent surface relationships between observables that could be realised by a variety of unspecified mechanisms/causes/modes of action. In fact, Mach further insisted that scientists shouldn't assume the existence of something underlying these correlations in the first place.

In many simple cases, observables are denoted by variables, while their correlation over a more or less well defined chunk of space-time is represented by linear mathematical functions; more complex formalisms introduce differential and partial differential equations (the derivation of Kepler from Newton being an example of the latter). The variables are interpreted as properties of a physical system, while their various relationships remain generic structures, as it doesn't matter what the relationship is from a physical point of view (van Fraassen, 1997 p. 516). Back to Hertz's insight, we can see how, at the most basic level, a deductive-style model can consist of several correlations, each embedded in a suitable mathematical formulation, further combined in order to derive new correlations, which are ultimately related back to empirical reality via the physical interpretation of their variables. Thus, to use Hertz's terminology, a model is an instrumental surrogate whereby "*intellectually necessary consequences*" represent the "*naturally necessary consequences*" (Hertz, 1899 pp. 9-10; Preston, 2008).

Borrowing from the literature on representation in art, Suárez proposes a more general approach meant to circumvent altogether the requirement for structural resemblance between an empirical correlation and its mathematical formulation. According to Suárez, the primary function of scientific representation is to allow for "*surrogate reasoning and inference*" such that

> "*A represent B only if (i) the representational force of A points towards B, and (ii) A allows competent and informed agents to draw specific inferences regarding B.*"

(Suárez, 2004 p. 73)

Condition (i), amounting to denotation, is meant to take care of a number of problems plaguing structuralist accounts of representation, most notably objections pertaining to the logical properties of representation (non-reflexivity, asymmetry and non-transitivity). Condition (ii) aims to provide a general enough condition for representation, capable of handling cases in which the model and its target do not share a common structure (Suárez, 2003 pp. 230-237). Since both conditions presuppose an inferring agent, Suárez argues that misrepresentation occurs when an uninformed agent intends the model in reference to the wrong target or incorrectly draws inferences from the model about the target (Suárez, 2003 p. 238). The account is therefore apt to handle the logical properties of representation and allow for misrepresentation, two feats notoriously difficult to achieve by resemblance accounts.

### *9.3 Strengths and Weaknesses of the Inferential Account*

If Suárez is right, it follows that a model can serve the purpose of representation independently of a poor or absent resemblance with the target; in other words resemblance is not necessary (the non-necessity argument), which further entails that some models can represent without resembling their targets (the argument from variety). From an inferential point of view, a rock spinning on a string doesn't necessarily represent the motion of Moon around Earth in virtue of some structural similarity or identity, but rather in virtue of the fact that inferences about the Moon revolving around Earth are drawn based on a study of a rock-spinning-on-a-string experimental setup (or what is sometimes called a 'material model'). That the accuracy of some of these inferences is, or could be justified by a structural resemblance is merely a peculiarity of the study case.

The potential irrelevance of isomorphism, similarity or any other kind of resemblance to the issue of representation becomes more obvious in cases where the resemblance between the experimental setup and the target is either assumed for the purpose of generating a preliminary experimental setup or constitutes the hypothesis under test. For example, the *in vitro* HIV infection of immortalised T-cell lines is commonly used as a substitute system for studying the *in vivo* human infection of HIV. The inferences from the cell line infection to the *in vivo* infection are hypothetical, that is, it is not clear a priori if the *in vitro* model accurately models the *in vivo* HIV infection. Granted, the cell line system is selected in virtue of some initial similarity (T-cell lines are immortalised T-cells collected from lymphoma patients, and therefore easier to work with; HIV remains HIV in both cases, although a less virulent strain is usually used), but whether the same mechanisms are responsible for the infection in both cases is merely a conjecture. In this case, an *in vitro* experimental setup is intended as a physical substitute to the actual HIV infection, but the ability of the experimental setup to generate correct inferences about the target remains conditional on the confirmation of the well founded of the inference/extrapolation from *in vitro*/cell-lines to *in vivo*/primary cells. This suggests that whether the experimental setup accurately resembles the target – by whatever means, including structural resemblance – remains secondary to the fact that it is used by scientists as a substitute for the target phenomenon. More so, the inferences about *in vivo* HIV infection based on the study of *in vitro* systems may turn out to be false, in which case, instead of representing, the experimental setup actually misrepresents its target, thus ruining the possibility of isomorphism, and presumably that of strong similarity as well.

The above considerations strongly indicate that Suárez is right in claiming that representation has a lot to do with pragmatics and cannot reduce to a factual relationship between the rep-

resentation and its target. Nevertheless, despite the net progress, I don't the think the question is settled. Even if it is true that, in general, resemblance is not necessary to representation, and even though, in particular, some scientific models may not resemble their targets, many models simply cannot dispense of resemblance. Too much is assumed on very general grounds and too many cases supporting some form of resemblance are ignored.

Suárez insists that a distinction should be made between the conditions for representation and those for accurate representation. Prima facie, this doesn't constitute a problem, but an asset. After all, an account of representation should be flexible enough to allow for misrepresentation. On the other hand, this requirement entails a rather puzzling consequence: if a representation doesn't have to be accurate in order to function as a representation (Suárez, 2003 p. 226), then the inferences it grants don't have to be accurate, reliable or truthful either. It follows from here that given sufficient will, 'representational force' or intent, anything can be used as a representation of anything else.

I assume Suárez is confident that representation will not collapse into denotation because of condition (ii). Still, it is not all that clear what prohibits aberrant inferences. In arts, and the realm of the social in general, suffices to declare the intention, that is, claim that *A* represents *B*, and 'competent and informed agents' will do the job of inferring something about *B* based on the properties of *A*. The underlying premise is that within any given socio-cultural circle agents abide to similar standards for what counts as acceptable inference and therefore come to a common understanding of how and in what sense *A* represents *B*. This accounts for the fact that what is used as a representation usually succeeds in functioning as a representation. As for misrepresentation, it is bound to occur when individual agents ignore or are unaware of the declared in-

tent of the representation, or again when their inferences diverge from the standards specific to their social and cultural milieu.

It seems therefore that, given a declared intent and a set of tacit social rules defining standards of acceptable inference, representation is ensured to be for the most part functional and reliable, while allowing for an occasional misfiring. This strongly indicates that, in addition to the two conditions highlighted by Suárez, there must also be a third condition for representation, namely a relative socio-cultural context fixing acceptable inference. Consider that given the phonetic structure of the English writing system, it is tacitly assumed that we should not infer something about roses based on the physical shape of the letters forming the word 'rose'; this is not the case in other cultural contexts, where equally competent and informed agents are allowed to draw such inferences.

If this conclusion is correct, it follows as a special case that scientific representation must depend on a third condition fixed by the standards of acceptable scientific practice. Ironically, it seems to me that acceptable scientific practice requires that representation 'mirrors' at least one key aspect of the target phenomenon. For instance, if an epithelial cell-line is used instead of a T-cell line, researchers will legitimately infer that HIV is not an infectious agent after all, conclusion inconsistent with the very purpose of the experimental setup, which is to study HIV infection. The epithelial cell-line system is meant to represent HIV infection, but fails to do so, not because mistargeting or incorrect inference, which are human errors, but because the representation fails to objectively capture, mirror, depict, in short, resemble its target in respect to some fundamental properties. As discussed above, if the mechanisms responsible for infection and cell death are not the same, an *in vitro* infection of immortalised T-cells will provide an inaccurate representation of the *in vivo* HIV infection. In contrast, an *in vitro* infection of immortalised

epithelial cells is not, scientifically speaking, a model, accurate or inaccurate, of *in vivo* HIV infection for the very simple reason that HIV fails to infect and multiply in epithelial cells.

This conclusion is in agreement with my account of explanatory models, as illustrated in the analysis of Newton's gravitational model of planetary motion. It can, and it has been argued that the model amounts to a theoretical representation of Mars revolving around Sun, Moon around Earth and, under some idealisation conditions, of the Solar System as a whole (Giere, 1988). Alternatively, according to inferential accounts, the model represents planetary motions because it allows for surrogate reasoning. Both approaches capture some aspects of the actual scientific practice. In agreement with inferential accounts, the model allows for the inference regarding the existence of gravitational forces, 'perturbations' in the orbits of planets, 'wobbling' of their spinning axes, tidal effects, etc. A number of conclusions about planetary motion, some true, some false, are inferred from the model, leaving plenty of space for misrepresentation. Nevertheless, it is also the case that Newton specifically used a set of propositions that entails Kepler's laws, assumed to describe the orbit of Mars. As discussed previously, Newton explicitly constructed his model in such a way that it allows for at least one empirically accurate, or what he took to be an empirically adequate inference, namely the Keplerian description of Mars's orbit around Sun.

Something very similar must apply to experimental setups ('material models' used for purposes other than didactic), like cell-line systems. Just because a cell-line experimental setup is used as a substitute study system for *in vivo* HIV infection and allows for reasonable inferences, doesn't mean that the former is a model of the latter. The system used must be able to physically sustain HIV replication if it is to model and represent in any scientifically useful way the *in vivo*

HIV infection. Failure to comply with this minimal requirement for resemblance is a sure recipe for disaster.

### *9.4 To What Extent Models Resemble their Targets?*

It seems therefore that scientific representation is not just any kind of representation. Or again, it might be that the conditions spelled out by Suárez only pick up what is used as a representation in most domains of human activity, and not necessarily what successfully functions as a representation in the context of the scientific practice. Whatever the problem is, it has something to do with the fact that any scientific model, even a purely instrumental one, is empirically adequate in at least one respect, and, in that respect, it inevitably 'mirrors' its target.[21]

As a general rule, logical positivists assumed that theories are used instrumentally towards the derivation of ultra-specific predictions referring to phenomena occurring at specific points in space-time. In the actual scientific practice, the target predictions usually refer to regularities; furthermore, it is not raw descriptions of the phenomena themselves which are targeted, but rather models of data, such as statically relevant correlations matching a linear function within the limits of experimental error (Suppes, 1962). This is the case of Newton's gravitational model, targeting Kepler's laws, kinetic models of gases, targeting the ideal gas law or some corrected variation of it, or again Mendel's genetic explanation, whose target is the phenotypic composition of a given offspring population mathematically expressed as proportions, frequencies or probabilities. Since both the model and its target are structures, it is possible, at least in

[21] French (1999) argues that even if insufficient for representation, 'resemblance' – a notion which can be conceived of in terms of isomorphism or partial isomorphism – is always present, not only in science, but also in arts. This suggests that 'resemblance' plays a role in all kinds of representation.

principle, to have an isomorphism or homomorphism between the conclusion proposition of deductive model and a model of data.

More ambitious projects propose a further morphism – embedding (van Fraassen, 1980), homomorphism (Lloyd, 1988), partial isomorphism (French, et al., 1999) – between the explanatory model as a whole and a model of data describing the target phenomenon. Most of the objections raised against the isomorphism/similarity accounts of scientific representation have in mind this latter, more ambitious conception.

Developing an argument by Cartwright (1999 p. 39), Frigg points out that "*structure S does not apply unless some more concrete description of the target system applies as well*":

> "*Take transitive relation, for instance. There are many transitive relations: taller than, older than, hotter than, heavier than, stronger than, more expensive than, more recent than, etc.* [...] *transitive relation is true of a relation only if either greater than, or older than, or ... is true of it as well.* [...] *There simply is no such thing in the physical world as a relation that is nothing but transitive.*" (2006 pp. 45-46)

In addition, Frigg further argues, a model doesn't have a structure, but rather many structures which are not necessarily isomorphic. For example,

> "[t]*he methane molecule ($CH_4$) consists of four hydrogen atoms forming a regular tetrahedron and a carbon located in its middle.* [...] *What is the structure of a tetrahedron?* [...] *A natural choice seems to regard the corners (vertices) as the objects and the lines that connect the vertices (the edges) as the relations. As a result we obtain the structure $T_V$ which consists of a four-object domain {A, B, C, D} and the relation L ($L_{xy}$ = 'x is connected to y by a line'), which has the extension {(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)}. However, this is neither the only possible nor the only natural choice. Why not consider the lines as the objects and the vertices as the relations?* [...] *Following this idea we obtain the structure $T_S$ with a domain*

*consisting of the six edges {a, b, c, d, e f} and the relation I ($I_{xy}$ = 'x and y intersect'), which has the extension {(a, b), (a, c), (a, d) (a, f), (b, c), (b, d), (b, e), (c, e), (c, f), (d, f), (d, e)}.*"

(2006 p. 47)

The introduction of 'models of data' provides half of the solution to the problems pointed out by Frigg. To make things as simple as possible, we can conceive of Kepler's laws as being a set of space-time positions, and argue that there is a one-to-one mapping of the elements of this set and those of a set derived via Newton's gravitational model (or again, argue for the 'embedding' of an array consisting of actual observations within the structure of Newton's predicted orbits or Kepler's geometrical description of planetary orbits). The other half of the solution stems from the fact that both the prediction and the target empirical law are already interpreted, this being the case even for predictions issued from a purely deductive/instrumental use of models. Newton's gravitational model yields predictions or conclusions about motions and Kepler's laws are also about motions; Mendel's explanation yields predictions or conclusions about phenotypic distributions from one generation to the next; etc. In such cases, there is a structural identity or embedding, as well as a uniform physical interpretation specifying what counts as an 'element' and what counts as a 'relationship between elements'. As discussed previously, many predictions and target empirical laws take the form of correlations whereby the variables involved are interpreted, while the relationships between them are not; this ensures that 'physically interpreted elements' and 'structural relationships' do not switch roles in the manner described in Frigg's tetrahedron example.

What is less clear is how the same line of reasoning applies to the model as a whole. Following Suppes' instantiation approach to modelling, it can be said both the rock-on-a-string and Moon-revolving-around-Earth are instances of the same theoretical model involving a centripetal force perpendicular on speed determining a rotational motion. A variety of forces can play the

role of the centripetal force (tension in the string, gravitational force, electrostatic force, etc.), such that several kinds of phenomena can fit the same 'force perpendicular on velocity determining motion according to the laws of dynamics' structure. Note however that this particular model is incomplete, or idealised, in the sense that it makes abstraction of the reactive forces which, according to the laws of dynamics, must affect, however slightly, the overall behaviour of the system, external influences, deformation, etc. I believe that, as a general rule, models are bound to be idealised or at very least make abstraction of some characteristics of their target phenomena, such that no phenomenon is capable of perfectly instantiating some theoretical construct and no theoretical construct can perfectly describe a given phenomenon; the reasons for this will become clearer in a moment.

At this point two options are available. One, take into account the idealisations and focus on the conditions under which some factors tend to be null or have null effects on the overall manifestation of the phenomenon, as it is usually done in college textbooks and as described in my own hypothetico-deductive framing of the idealisation practice (see below); or two, claim that the model bears an overall approximate resemblance to its target, where the resemblance in question amounts to a weaker alternative to isomorphism.

Both approaches have strong and weak points. The positive side of the hypothetico-deductive approach I favour is that a model, even if idealised, remains unambiguously attached to its parent theory, and hypothesises that each empirical instantiation satisfies the theoretical model in all theoretically-relevant respects. However, in as much as idealisations are framed as hypotheses about the target phenomenon, the approach has the disadvantage of forcing a realist interpretation about entities which may not be observable for theoretical reasons. In the context of the example discussed in the previous section, the model may postulate the existence of negli-

gible, and therefore extremely hard or even impossible to measure, reactive forces. However, this is an issue pertaining to confirmation and transcends the scope of the present paper.[22]

The alternative, 'weak resemblance' approach has the advantage of dropping a too rigid realist interpretation, but achieves this by introducing the relatively vague notion of 'similarity'. Most notably, Giere (1988; 2004) proposes that a theoretical model is an abstract and/or ideal object similar to a target phenomenon. The similarity relationship intervenes at two distinct levels. First, when it comes to matching conclusions/predictions of the model and target empirical laws/models of data, Giere argues that instead of talking about the omnipresent 'matching within the limits of experimental error' or 'statistical relevance', we should rather talk about degrees of similarity. Giere justifies his proposal as follows:

> "*The margins of error rarely appear in the descriptions or calculations until one gets to the point of comparing theoretical predictions with actual measurements. This practice strongly supports interpreting the original equations, without explicit margins of error, as referring not to actual things but to abstract models of which they are true by definition. When it comes time to compare the abstract model with reality, the deltas may then be understood as speci-*

[22] My solution to the problem would be to argue for partial confirmation, whereby some elements of the theoretical model match empirical descriptions of the target phenomenon, while other elements remain hypothetical. The latter are entailed by the model, realistically interpreted, and might be indirectly confirmed when several models combine in order to jointly yield confirmed predictions. This approach is extensively used in molecular explanations and I suspect it might apply to other fields of investigation as well.

Alternatively, van Fraassen (1980) is famously known for adopting a stronger agnostic position by appealing to the notion of 'partial structure': only some elements of the theory are interpreted – in this case, those pertaining to space-time positions and derived notions, such as velocity and acceleration – and the target phenomena, or rather target models of data, are isomorphic only in respect to this interpreted 'fragment' of the overall structure provided by the model. The remaining, non-interpreted structure has no empirical relevance and therefore cannot be instantiated by any phenomenon. More recently, van Fraassen (1997; 2002) revised his position by arguing that even though many theories were developed on the premise of a realist interpretation, science could have very well evolved with no significant loss in the absence of realism. Elegantly formulated, van Fraassen's approach is drastically minimalistic and does not allow for a reframing of conclusions concerning theoretical terms as hypotheses about empirical reality. Most notably, talk about underlying mechanisms, microscopic structures and fundamental modes of action becomes largely irrelevant, as all that matters are the correlations between surface observables. While such a minimalist account may prove useful in some situations, there are also documented cases when the physical interpretation played a crucial role in the development of theories. For instance, the development of present-day molecular biology was largely motivated by an attempt to elucidate the physical nature of the entities postulated by classical genetics.

*fying the degree of similarity (either expected or actual) between the abstract model and the real system.*"

(2004 p. 648)

In short, the margins of experimental error, error bars, etc. are interpreted as a measure of the degree of similarity between an abstract object (the model of data) and the actual empirical description of the phenomenon (the raw data). The autonomous existence of an 'abstract object' is justified by the fact that, when developing theories and explanations, scientists drop the raw data and use the model of data as a substitute. (Note that Giere's insistence that a model is an abstract/ideal object used as a substitute for the empirical phenomenon suggests that resemblance accounts should be treated as a subset of broader-scope inferential accounts.)

Similarity intervenes a second time, at a higher level: the theoretical model as whole is also an abstract object bearing an overall similarity with a target phenomenon. According to Giere (1988; 2000), a model is constructed in accordance with a set of fundamental laws and principles and, just like a model of data, this higher-level theoretical construct has an independent existence as an object bearing a certain similarity with a target phenomenon. For example, Giere argues that, in respect to the resulting sinusoidal position as a function of time description of motion, both a mass-on-a-string and a pendulum oscillating at small angles are more or less perfect instantiations (similar within different margins of experimental error) of a theoretical object called the 'ideal linear oscillator' (1988 pp. 68-76). From a practical point of view, this allows for the modelling of a more complex phenomenon, in this case the two-dimensional motion of the pendulum, in terms of a simpler phenomenon, namely the weight on a string described by motion in only one dimension. However what interests Giere is the fact that the two systems are not "*special cases of a general relationship*" since each bears its own, specific degree of similarity with the theoretical construct.

Once again, little can be objected here. The defining feature of the 'ideal linear oscillator' is the absence of a dampening effect, hence its perfectly sinusoidal motion. This implies that no kinetic energy is lost during its forth and back transformation in potential energy via friction, deformation, heat, etc. Thus, the model is both ideal and abstract. It is ideal because it posits a very peculiar 'no loss of energy' condition, typically non-satisfied by real systems. And it is abstract or general enough because it does not specify what kind of potential energy is transformed forth and back into kinetic energy and what are the possible ways in which energy may 'leak' out of the system. It seems therefore that Giere's claim that a theoretical model is entitled to an autonomous existence as an abstract object is reasonably justified. I also agree with Giere's claim that the same theoretical model or theoretical treatment may be applied to different phenomena with different degrees of accuracy, and under distinct idealisation circumstances, as illustrated in his discussion of the linear oscillator.

This said, it is not clear yet how a theoretical model, understood here as a complex abstract object, can be compared with a phenomenon. Predictions are about limited number of aspects of the phenomenon under study and it is in those specific respects that individual predictions entailed by a deductive model are compared with raw measurements or models of data. Predictions about the orbit of Mars are not compared with the phenomenon of Mars revolving around Sun, but with the observed orbit of Mars; predictions about phenotypic distributions are not compared with the phenomenon of genetic inheritance, but only with the relative proportions of various offspring phenotypes; and so on. The actual phenomenon of Mars revolving around Sun, or again genetic inheritance comprises many more aspects, features and properties; the same applies to the models aiming to explain these phenomena. Things get even more complicated if we take Newton's gravitational model and Mendel's genetic explanation as representa-

tions of the 'inner workings' underlying the functioning of the Solar System and genetic inheritance.

Borrowing from the literature on exemplar and prototype theories in psychology (Medin, et al., 1978; Nosofsky, et al., 2000; Posner, et al., 1968), Giere's solution is to decompose the overall similarity between model and target in several respects, each having its own specific degree of similarity. For instance, in the case of Newton's gravitational model, we are told that

> "[t]*he positions and velocities of the earth and moon in the earth-moon system are very close to those of a two-particle Newtonian model with an inverse square central force. Here the respects are 'position' and 'velocity', while the degree is claimed to be 'very close'.*" (1988 p. 81)

Note however that in doing so, several problems arise. First, in the broader tradition of semantic approaches, it is assumed that a sharp distinction can be drawn between the relationship of 'satisfaction', which links theory and model, and that of 'resemblance', linking model and empirical reality. The former is explicitly and rigorously formulated, while the latter seems to belong to the realm of implicit judgments. In this respect, Sklar observes that

> "[i]*t is often emphasized that the degree of similarity of model to world, and the specification of the context in which the model is sufficiently similar to the world for the laws to have genuine predictive and explanatory value, are, once again not explicit in the theory itself. The application of lawlike theory to world, then, partakes of a kind of implicit learned scientific practice, something outside the explicit content of the theory in question.*" (2002 p. 42)

To simplify matters as much as possible, let us begin by considering a purely material model. Since we are dealing with perceptual objects it is not necessary to provide an explicit account of how the comparison is achieved; we tacitly rely of our natural ability to compare per-

ceptions, be them real or imaginary. For instance, it is possible to construct small-scale models reproducing planetary motions in the absence of any thorough knowledge of the principles of mechanics. Presumably, such small-scale material models preceded Newton's deductive model and it is very probable that Newton made use of the former in order to develop his own deductive model. To keep things simple, there is an analogy between a rock spinning on a string and Moon revolving around Earth. By changing the scaling factors for velocity, length and strength of the string, volumes, 'heaviness' and inertia, etc., it is possible to generate an imaginary movie of an object the size and 'heaviness' of Moon spinning around a fixed point marked by the center of Earth.

Compared with the observed motion of Moon relative to Earth, this imaginary perception is both similar and dissimilar. From a strictly perceptual standpoint, there is an analogy between the imaginary spinning of a rock the size of Moon and the actually revolving Moon since the shape of the orbits almost coincide. On the other hand, the analogy is faulty because there is no string connecting Moon and Earth and we have no empirical knowledge about an agent giving a first impetus to Moon. Granted, the analogy is fruitful because it provides an insight about how the quasi-circular motion of Moon could be achieved (e.g., God spins it around Earth on an invisible string), yet, just as in Hesse's examples, the overall degree of similarity between the alleged model and the actual phenomenon remains limited and rather difficult to assess rigorously.

Oddly enough, it is not possible to establish a stronger degree of similarity without introducing the concept of force and a law stating how forces determine motions. From a post-Newtonian perspective, the most striking similarity between a rock spinning on a string and Moon revolving around Earth is precisely a fundamental principle of classical mechanics stating

that, independently of their nature, forces of identical magnitude, direction and orientation cause, or at very least correlate with identical changes in motion.[23]

This preliminary hitch reveals a more general problem: explanatory and theoretical models in general can be compared with their targets only in theoretically-relevant respects. In the case of Newton's gravitational model, the target phenomenon – say Moon revolving around Earth – can be compared in terms of motion, mass, shape, volume and derived properties such as density, momentum, torque, etc., but not colour, chemical composition, possibility to sustain life, etc., since the latter are not considered to be relevant by the mechanical theory and therefore have no place in a mechanical model. On the other hand, a scale material model of the Moon-Earth system can be compared with its target in respect to colour and chemical composition. This indicates that there is more then one way of defining respects of comparison.

If this is the case, then it is not clear whether theoretically-irrelevant respects of comparison are simply to be dismissed or whether they count as respects having a high degree of dissimilarity. If the latter, Hegelian stance is adopted, then most models are bound to be highly dissimilar to their respective targets in an indeterminate number of respects; this must be the case since most theories specify which features of the target phenomena are relevant and not those which

[23] By themselves, most notably in the absence of the rules used to infer something about a target phenomenon from a substitute experimental setup, material models don't explain anything. For example, the rock-on-a-string substitute experimental setup explains the motion of Moon around Earth only if we appeal to the general notion of force (i.e., the tension of the string = gravitational force = centripetal force) and assume that the laws of classical mechanics hold true on Earth and in Heavens. Furthermore, without these assumptions, there is not enough resemblance between the model and target in some key respects and it is not clear in what sense the former can allow for surrogate reasoning, thus threatening not only the explanatory relevance, but also the representational value of the model. Similarly, the *in vitro* infection of T-cell lines by HIV is meant as a substitute system for studying the *in vivo* human infection of HIV. Despite some obvious surface similarities between the target and the substitute phenomena, the model explains some key aspects of AIDS and human HIV infection only in as much it is further assumed that the same mechanisms responsible for the *in vitro* infection are at play *in vivo*. Without this assumption, which explicitly attaches a theoretical dimension to an otherwise purely 'material' model, it is not possible to infer something about the target phenomenon from the model and the door is left wide open to major points of dissimilarity that can quickly neutralise the initial similarity. To use's Harré's terminology, a material model is a 'homeomorph' and does not bring new ideas in theory construction. In contrast, an explanatory model derived in light of a pre-existing theory is a 'paramorph' and posits an analogy "*between certain characteristics of different processes*" (Harré, 1960 p. 87).

are irrelevant. I conclude therefore that this option must be avoided at all costs. Although less damaging, the first strategy is also imperfect. If adopted, then Giere's approach retains an arbitrary character, for it assumes that the respects of comparison imposed by the theory are somehow more important or fundamental than those inherent, say, to visual perception.[24]

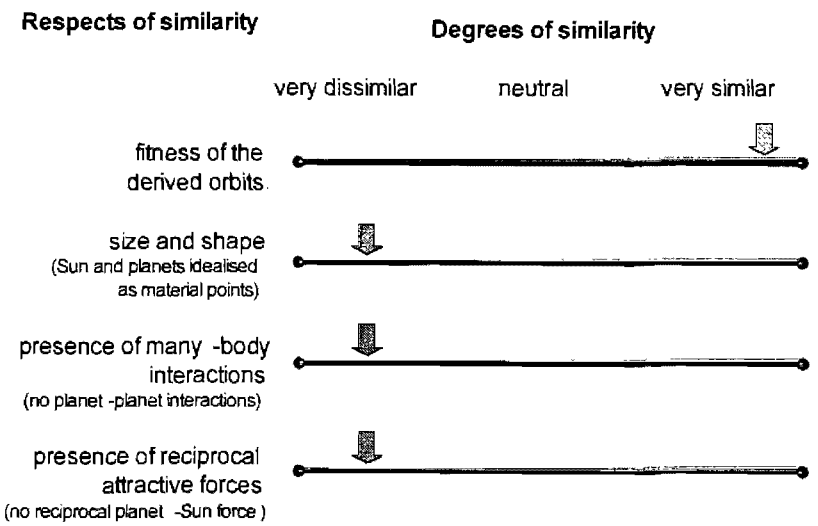
Finally, even if we limit ourselves to theoretically-relevant respects of comparison things can go bad. Again, back to the planetary motion example, Newton first considered a simplified model consisting of a point-like planet revolving around a point-like Sun, where Sun alone exerts a gravitational force over the revolving planet (Lakatos, 1970 p. 135). Thus, this preliminary model relies on idealisations concerning

i) the size and shape of Sun and planets;

ii) the number of gravitational interactions; and

iii) the third law of dynamics.

The three idealisations inherent to Newton's first model count as theoretically-relevant respects of comparison, yet the degree to which they differ from the intended phenomenon is maximal:




[24] Suárez (2003 p. 235) exacerbates the objection by remarking that if logically possible properties are possible, any object is trivially similar to any other object.

Figure 16. Respects of Comparison

Unless it is decided from the very beginning that the theoretically-relevant respects in which the model is idealised are not that important, the average tendency inclines towards an overall dissimilarity rather than an overall similarity. Giere doesn't say a word about how the various degrees of similarity proper to each respect of comparison are to be weighted in a non-arbitrary manner.[25]

### *9.5 A Hypothetico-Deductive Treatment of Idealisations*

A deductive approach to modelling faces its own difficulties that must be accounted for. The main advantage of the similarity approach is that idealisation is embedded into the similarity relationship; the bad part is that, as shown above, similarity brigs about its own set of problems. Conversely, a deductive approach doesn't have to deal with the problems associated to similarity, but it must give a clear answer to the issue of idealisation.

In my example, background assumption (I) in Putnam's list ("*No bodies exist except the sun and the earth*") is a direct allusion to a computational problem imposing an idealisation on Newton's gravitational model. In order to be thoroughly consistent with its fundamental and model-specific hypotheses, a gravitational model of planetary motion must take into account the gravitational interactions between all the masses in the Solar System; mathematically however, Newton could derive Kepler's laws only in as much the model reduced to a one or a two-body

[25] Sklar (2002 p. 42) openly complains that

> "*it is difficult to see how the introduction of this notion of model and the adoption of a relation of similarity of model to world will be of much help in understanding the applicability of law to world. We still are left with all the problems we may have had initially about explaining how lawlike assertions, if literally false, can be relevant to us in our predictive and explanatory tasks. For all the problems of characterizing just what the applicability of law to world consists in, and, in particular, all the original problems generated by inexactness, contextuality, and the ceteris paribus clause, still remain. These problems are now simply embedded in the notion of similarity, and the unpacking of that notion remains as obscure a task as was understanding the original notion of applicability of law to world.*"

problem. Unfortunately, not knowing how to solve a many-body problem is tantamount to recognising that it is impossible to derive the target empirical laws, in turn meaning that there is no model to begin with. It follows from here that, at least under a deductive treatment, any gravitational model is bound to be an idealised one. This simply shows that, within a deductive framework, the idealisation issue must be addressed frontally, and not laterally, as a side-effect of less-than-perfect similarity.

Lakatos provides a short description of the idealisation inherent to Newton's first gravitational model and of the subsequent attempts to alleviate the nefarious of idealisation effects by devising "*a chain of ever more complicated models simulating reality*":

> "*Newton first worked out his programme for a planetary system with a fixed point-like sun and one single point-like planet. It was in this model that he derived his inverse square law for Kepler's ellipse. But this model was forbidden by Newton's own third law of dynamics, therefore the model had to be replaced by one in which both the sun and planet revolved around their common centre of gravity.* [...] *Then he worked out the programme for more planets as if there were only heliocentric but no interplanetary forces. Then he worked out the case where the sun and planets were not mass-points, but mass-balls.* [...] *Having solved this puzzle, he started work on spinning balls and their wobbles. Then he admitted interplanetary forces and started work on perturbations.* [...] *It was then that he started to work on bulging planets, rather than round planets, etc.*"
>
> (1970 pp. 135-136)

Two common difficulties are associated with idealisation. First, Lakatos argues, Newton's 'faith' in his project rests on a 'methodological decision' to save at all costs the gravitational hypothesis by making the model more and more realistic. The suggestion here is that Newton's initial gravitational models are 'false', or, to be more precise, fail to contribute to the con-

firmation of the theory from which they are derived, not as much because they fail to entail the target empirical descriptions, but rather because they explicitly ignore data deemed relevant by the mechanical theory used to construct the model. This incompatibility between models and their parent theories (and, in this case between the model and some of its model-specific hypotheses as well) is particularly upsetting for deductive accounts, which hold that a model amounts to the derivation of target empirical laws from more general theoretical propositions.

And second, there is a logical problem often pointed out:

> "*On the deductive nomological model of scientific theories* [...], *a theory is a deductive scheme which uses laws and initial conditions to derive predictions of events or lower-level laws. A deductive scheme should yield true consequences when the premises are true. However, if idealizations are admitted as premises, the premises are, strictly speaking, false, and the conclusions need not be true even if the argument is valid.* "
>
> (Ben-Menahem, 2000 p. 169)[26]

Hopefully there is a possible solution to both difficulties. Laymon observes that, in deriving a solution for Einstein's field equations,

> "*an idealized description is employed: the Schwarzschild 'solution' assumes a perfect symmetrical non-rotating sun and no other masses. Together the field equations and the Schwarzschild idealization yield a solution for the metric.*"
>
> (1984 p. 109)

There are however more realistic models, such as the approximation of Sun to a rotating, spherical object used to derive the Kerr metric. Given a cascade of more and more realistic models, Laymon argues that

[26] In the context of a counterfactual interpretation, the logical problem disappears, since the premises are true in respect to a counterfactual description rather than the actual phenomenon. The strategy in this case is to argue that some counterfactual states of affair are more similar to the actual phenomenon than others, and that models true about counterfactual states closer to actual reality are more 'truth-like' than models true about more distant counterfactual states (Oddie, 1986; Niniluoto, 2000).

*"a theory is confirmed if it can be shown that it is possible to show that more accurate but still idealized or approximate descriptions will lead to improved experimental fit; a theory is disconfirmed when it can be shown that such improvement is impossible".*

(1984 p. 117)

In a analogue fashion, it can be argued that the main goal of Newton's model was to establish whether a mechanical model for planetary motion can be empirically adequate. The initial task was to approximate Kepler's laws. More elaborate versions of the initial model surpassed the empirical adequacy established by Kepler's laws and were able to explain additional phenomena. Since an increased experimental fit provides confirmation beyond the empirical adequacy aimed at initially, Newton was empirically justified to continue working on his project.

The above provides a patch for the inconsistency issue, but fails to really address the logical problem. The other half of the solution relates to what Sklar calls 'controllability', that is, to the notion that some idealisations are granted by the theory in light of which a phenomenon is modelled. This should provide a satisfactory closure to the logical problem. For example, in relation to the external interferences on an allegedly closed system, Sklar remarks that

*"the scientist believes, rightly or wrongly, that such interferences, even if unavoidable, are in general controllable. What does 'controllable' mean? The scientist believes that current scientific theory, including the substantial background theory that runs well beyond that part of current theory directly applicable to the system in question, possesses the resources necessary to tell the scientist in some cases that the outside interference is negligible."*

(2002 p. 44)

I further propose that an idealised model introduces its own hypotheses stating that certain features of the phenomenon, acknowledged by the theory to play a role in determining the overall behaviour of a phenomenon, have nevertheless a negligible contribution in the particular

case of the phenomenon under study. This approach takes into account the fact that, typically, a scientist doesn't just choose an idealised model because it is simpler this way, or in virtue of some vaguely specified analogy, but on the grounds that, assuming that the theory used to model the target phenomenon is true in respect to that phenomenon, then the disregarded features ignored by the idealised model must have a negligible contribution to the overall behaviour of the phenomenon.

Thus, the problem can be framed as follows. Technically, Newton was able to derive Kepler's laws – laws deemed to accurately describe the motion of Mars around Sun – from an idealised two-body model. The empirical adequacy of the model is insured. However, Newton and his successors considered a two-body model to be idealised in the sense that data that should have been relevant given the fundamental and model-specific hypotheses of the model is not used in the derivation of Kepler's laws. Note that the model is not idealised because it ignores the colour of Mars and Sun, their chemical composition, the fact that there Mars may or may not support life, etc., but because it ignores the gravitational influences of other bodies present in the Solar System, the density distribution of the two bodies, rotation effects and other theoretically-relevant aspects of the phenomenon of Mars revolving around Sun.[27]

The final solution reads as follows. Once he was able to derive Kepler's laws, Newton and his successors had to do one of the following two things:

i) theoretically justify the fact that some theoretically-relevant data is 'negligible' towards the derivation of the target empirical laws/description, that is, show that the theoretical approach used allows for certain idealisations; and/or

[27] Conversely, taking into account theoretically-irrelevant data is just as inconsistent with the theory as not taking into account all the theoretically-relevant data; in fact, if it can be shown that theoretically-irrelevant data plays a role in determining the target empirical description, then the theory is technically falsified. Neither Lakatos, nor Laymon discuss this aspect of idealisation.

ii) show that more realistic models taking into account the theoretically-relevant data neglected by idealised models allow for the derivation of more detailed empirical descriptions of the studied phenomena, not targeted by the initial models (Laymon's solution).

The advantage of (i) is that it allows for idealised models to directly contribute to the empirical justification of the theory from which they are derived. If it is possible to show that some data is negligible for theoretical reasons, then idealised models can be shown to be consistent with their 'mother theories' such that the empirical adequacy of such models can count towards the confirmation of the theory underlying them.

In reference to Newton's gravitational model, it can be easily argued that the three idealisations exemplified by Lakatos, and which must necessarily mark points of dissimilarity, become three additional model-specific hypotheses, as explained in the figure below:

**Idealisations**

| as **dissimilarities** concerning | as **model-specific hypotheses** | as **model-consistent hypotheses** |
|---|---|---|
| • the size and shape of Sun and planets | • the size and shape of Sun and planets is negligible | • the size and shape of Sun and planets is negligible in respect to the overall distance separating Sun and planet<br>• since $F_g \propto 1/d_{Sun\text{-}planet}^2$, if $d_{Sun\text{-}planet} >> r_{Sun}$ and $r_{planet}$, then the gravitational force acting on the poles is equal to that acting to the equatorial line |
| • the number of gravitational interactions | • the planet -planet interactions are negligible | • the planet -planet interactions are negligible in respect to the Sun -planet interactions<br>• since $F_g \propto m_1 m_2$, if the mass of Sun is considerably greater than that of any planet, then $F_{Sun\text{-}planet} >> F_{planet\text{-}planet}$ |
| • the third law of dynamics | • the third law of dynamics can be ignored | • the mass of any planet is negligible in respect to that of Sun, such that the third law of dynamics can be ignored<br>• since $F_g = m_{planet}\ a_{planet} = m_{Sun}\ a_{Sun}$, if $m_{Sun} >> m_{planet}$, then $a_{Sun} << a_{planet}$ (Sun is stationary) |

**Figure 17. Deductive Treatment of Idealisations**

All three idealisations can be theoretically justified in relation to some initial conditions. Undoubtedly, Newton simply assumed that Sun is significantly more massive than any of the planets, just as he made a number of assumptions about the Sun-planet distances based on the Copernican interpretation of astronomical observations. Nevertheless, he did not introduce any of the above idealisations as arbitrary dissimilarities justified by pragmatic concerns. Rather, he introduced them as consequences deductively granted by the model *if certain initial conditions hold true*. The end result is that instead of having to justify the assumptions themselves as additional, independent propositions required for the derivation – and therefore external to the theory –, it is only required to justify the conditions under which they follow as certain limit cases of the fundamental hypotheses used to formulate the model.

A hypothetico-deductive treatment of idealisations may provide some other advantages as well. For one thing, treated as model-consistent hypotheses, Newton's idealisations are not false premises, but consequences which obtain given specific initial conditions. Hence, they do not necessarily refer to a counterfactual state of affairs, but also to actual special cases or circumstances [what Sklar refers to as 'limit cases' (2002 pp. 61-62)]. Just like similarity, counterfactuals are useful, but not perfect, since they allow straightforward solutions for old problems while creating a bunch of new and potentially more difficult problems.

Secondly, under Giere's similarity treatment, simpler and more complex models contradict each other since they represent differently the same phenomenon. Under deductive treatment, simpler and more complex models equally contribute to our theoretical knowledge of the phenomenon. Knowing that some parameters have a negligible effect adds to the total knowledge about a certain segment of empirical reality since we are provided with a piece of knowledge not explicitly included in more complex models. Simpler models show that, in some conditions,

some parameters do not have a significant effect in the overall manifestation of a phenomenon, while more complex models progressively take into account more variables as they become relevant in various situations, thus yielding more generally applicable descriptions and predictions.

I think it is reasonable to assume that any theory claims that certain aspects and laws are more fundamental than others. For instance, classical mechanics states that all physical motion is determined solely by the three laws of dynamics. Although not always explicitly stated, this includes a clause of sufficient grounds of determination. Force, mass and speed suffice to describe and determine motion, while the colour and the chemical composition of moving bodies don't contribute to knowledge about motion. It is for this reason that I believe that many, if not most explanatory models are inherently reductive in the sense that they hypothesise that only some features of the target phenomenon are 'dynamically interlinked' and affect each others manifestation, while the remaining features are 'inert' and, if changed, do not affect other features (this distinction is usually part of the 'fundamental hypothesis' common to all the models associated with a given theory). Simplified models further add their own negligibility hypotheses stating that certain aspects, even if relevant from a theoretical point of view, have nevertheless a minimal impact in some particular circumstances.

Thus, even if, due to the introduction of model-specific hypotheses, simpler and more complex models of a theory cannot always amount to special cases of the same general relationship, they do share the same fundamental hypothesis, as well as some of the model-specific hypotheses. This common theoretical backbone should allow for a convergence of the predictions yielded by simpler and more complex models in those cases in which the initial conditions allow for idealisations:

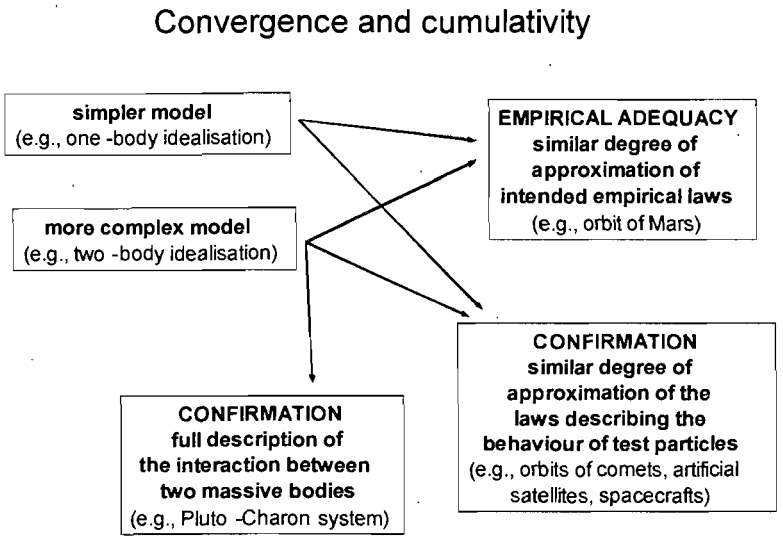



**Figure 18. The Convergence and Continuity of Scientific Knowledge**

In turn, the convergence of simpler and more complex models may provide an answer to the objection from ontological incoherence. As cited earlier, Morrison is concerned that the same chemical compound cannot be at the same time an ideal and a van der Waals gas. Presumably, some gases fit the equation describing ideal gases, in part, because the size of the particles is negligible in respect to the overall volume of the gas. It is clear however that no volume of gas can ever be compressed ad infinitum, meaning that no gas is an ideal gas, but rather behaves like an ideal gas under certain conditions. At some point during the compression the size of the particles is bound to become non-negligible in respect to the overall volume of the gas, hence the switch from $P \propto T/V$ to $P \propto T/(V-b)$, where $P$ is pressure, $V$ is volume, $T$ is absolute temperature and $b$ is the volume of a mole of particles.

### *9.6 Models as Representations of the 'Dynamic Structure' of Phenomena*

Thus far, I argued that a theoretical model must resemble its target phenomenon at the level of the empirical laws and/or predictions. I related this requirement to the fact that, in order

to function as a representation, a model must accurately represent some aspects of the target phenomenon. Under a strictly deductive framing, a model consists of mathematical propositions arranged in a deductive argument. Thus, it is not problematic to claim that some of these propositions are isomorphic, similar and many cases plain identical with target models of data also consisting of mathematical propositions.

However this essentially positivist point of view fails to clarify the main issue: How exactly scientific models represent empirical reality? Inferentialists argue that models represent by providing a means for surrogate reasoning, while adepts of the resemblance approach argue that models are abstract/ideal constructs mirroring some structural aspects of the target phenomenon. The dispute is primarily about which of the two is absolutely necessary for representation in general and to scientific representation in particular.

Although the most generally applicable and compatible with a strictly deductive or instrumentalist framing of the modelling practice, I think that Suárez's account fails to address the fact that any scientifically useful representation must allow for accurate inferences in at least one specific respect. I understand perfectly well that many scientific models and theories were used instrumentally in the past and that the recent proliferation of computer simulations and mathematical models in traditionally empirical domains of investigation such as biology and psychology justifies to a large extent Suárez's point of view. On the other hand, it seems to me that this point of view applies more accurately to theoretical science and since, quite obviously, theoretical science cannot constitute by itself all science, I tend to disagree with Suárez's suggestion that models are first and foremost conceptual devices allowing for surrogate reasoning. To use an example very dear to me, it is minimally required that an experimental setup is sufficiently similar with the target for which it is substituted in respect to the studied feature. I think it is quite obvi-

ous that the material models used for experimental purposes aim to reproduce some naturally-occurring counterpart, as opposed to denote it or metaphorically represent it. Furthermore, even the computer simulations used as low-cost alternatives to highly expensive or impossible to realise experimental setups still aim to reproduce available empirical laws and models of data given via instrumental algorithms. Granted, such models do not claim an overall resemblance with the target phenomenon, yet they have to yield at least some empirically accurate inferences.

At the other end of the spectrum, while attractive for its empiricism and ability to combine ideas from very diverse sources, I find that Giere's similarity approach needs to be further developed. On one hand, I assume that the notion of similarity can be further refined. By decomposing the overall similarity into a set of non-overlapping theoretically-relevant respects of comparison and assuming that, at least in some respects, the degree of similarity is 100%, similarity should collapse into a technical partial isomorphism. This suggests that the notion of similarity may receive a fully rigorous description. On the other hand, it is not clear what is the recipe for carving reality in 'respects of comparison' and how the degrees of similarity attached to each respect add together in order to yield an overall measure of the similarity between a model and its target.

Given these difficulties, I much prefer the less audacious, but easier to use traditional hypothetico-deductive scheme according to which a model hypothesises that only some properties of a phenomenon are 'dynamically linked' to other properties and play a role in determining the manifestation of the latter. That is, if the model is true, the variation of some features of the target phenomenon can be achieved by varying (or correlate with the variation of) some specific features and not other. The difference is subtle, but may suffice to save claims to partial resemblance. For Giere, theoretical models 'passively' reflect selected features of the target phenom-

ena with various degrees of similarity. The model is literally a map highlighting some aspects of reality while ignoring others; the 'similarity' part of the account is meant to take care of the fact that the map is not the same as a picture of the terrain in describes.

Note however that it is not clear why some features of the phenomenon are represented (or misrepresented) by the model while other features are not. In contrast, under a hypothetico-deductive treatment the model aims to represent the 'dynamic' or rather the 'experimental structure' of the phenomenon and usually succeeds in doing so every time it turns out the desired results can be achieved by experimentally manipulating only the theoretically-relevant variables. Suárez (2003 p. 237) complains that "*the case of representation of a well-established physical phenomenon by means of a differential mathematical equation is the hardest case for [iso] to accommodate*". My reading of representation aims precisely to take into account this shortcoming of resemblance accounts. While it might be true that in some individual aspects there might be a divergence between model and target, this is not necessarily a problem, because the model still accurately represents the 'dynamic interplay' between theoretically-relevant features. For example, the mass of an artificial satellite is negligible for theoretical reasons in respect to the Earth-satellite interaction, but certainly not in respect to the interaction between the satellite and an astronaut. In other words, the mass of the satellite is not negligible simpliciter, but rather in reference to its effects in the context of a given interaction. This kind of flexibility is impossible under Giere's approach: irrespective of the context, the mass of the satellite remains a rigid respect of comparison in which the model is either similar within a given margin of error or significantly dissimilar.

### *9.7 Concluding Remarks*

Ideally, the final goal would be to achieve a convergence between inferential and resemblance accounts of scientific representation. Suárez objects to the appending of an intentionality clause to resemblance accounts on the grounds that not all models resemble their targets and that isomorphism does not allow for misrepresentation; the main line of attack here is that resemblance is too strong, and therefore something must be subtracted from it rather than added to it (Suárez, 2003 p. 238). I think he is right. The amended resemblance approach fails to treat scientific representation as a special case of representation in general. I think that it is best to adopt the reverse strategy, namely start with Suárez's general conditions for representation and realise that, in order for surrogate reasoning to function reliably, representation must be further subjected to context-relative constraints fixing standards of acceptable inference. In particular, scientific representation further requires that a model directly represents some 'essential feature' of the intended target phenomenon. Thus, instead of appending intentionality and surrogate reasoning to resemblance accounts, it is resemblance that is appended to inferential accounts. In other words, resemblance doesn't represent by itself, although it is a necessary component of certain kinds of representation.

# PART III

# THE REALISM – ANTIREALISM DEBATE: THE CASE OF MOLECULAR BIOLOGY

CHAPTER 10

# REALISM AND ANTIREALISM IN CONTEMPORARY PHILOSOPHY OF SCIENCE

## *10.1 Scientific Realism*

I showed, via specific examples, how individual scientific explanations can be constrained, on one hand, by the available empirical data, and, on the other, by general patterns of explanation. I would like now to integrate these two levels of determination and address, once again by means of study cases drawn from the actual scientific practice, one of the hottest matters of debate in contemporary philosophy of science. The example I have in mind is the development of the present day genetic theory. The matter of debate is that of scientific realism.

In a well known essay, Boyd (1984) defines scientific realism as a doctrine embodying four central theses, which can be summarised as follows:

1) Theoretical terms, that is, unobservables to which scientific theories may appeal, refer to something existing in reality.

2) Scientific theories, interpreted realistically, are confirmable.

3) Science progresses towards more and more accurate approximations of the truth.

4) Reality is largely independent of thoughts, concepts and theoretical considerations.

Thus, scientific realism can be said to comprise three kinds of realism, namely metaphysical, semantic and epistemological realism. Metaphysical realism (thesis 4) states that the world exists outside us and has an intrinsic structure independently of our minds. Semantic realism (thesis 1) states that theories, models and the propositions of science in general assert something about reality and are true or false in respect to reality. Finally, epistemological realism (theses 2 and 3) states that it is possible to establish the truth or falsity of scientific knowledge and that it is possible to progress towards a more accurate and complete knowledge of reality.

### *10.2 Arguments for Metaphysical Realism*

Metaphysical realism stands in opposition to subjective and objective idealism. Subjective idealism – usually attributed to Berkeley, although Berkeley himself adhered more to a version of objective idealism where God and God's mind play a very important role – states that reality reduces to the mental activity of a given human subject. Objective idealism maintains that the human mind has access to non-material entities transcending the mental activity of any subject in particular; these non-material entities might be the ideas of Plato, the conceptual categories of Kant or again language for Hegel. Following a somewhat different approach, for contemporary metaphysical antirealists such as Dummett (1978) and late Putnam (1990), what we call the 'world' amounts to certain epistemic practices and conditions.

A metaphysical realist, especially in the context of scientific realism, will typically want to argue that there is an external reality in addition to or more fundamental than the reality of mental and linguistic items. This defines the metaphysical antirealist as someone who denies the existence of anything outside mental and linguistic reality. Many realists are also materialists, either rejecting mental activity as entirely inexistent and adopting a strong eliminative materialism (Churchland, 1988) or reducing it to a particular kind of material phenomena (Smart, 1968). It is worth noting however that the metaphysical realist can be content with a dualism accepting reality as being both mental and material, or rather extra-mental and extra-linguistic (Niiniluoto, 2000 p. 27).

Metaphysical realism is typically justified by the fact that some linguistic terms can be 'triggered' extra-linguistically. For example, a patient subjected to brain surgery tries to reconstruct the chain of events that led him on the operation table. He conjectures that he must have lost control of the vehicle he was driving, which led to a collision, which led to him knocking his head against some hard surface, which in turn would explain why he is now on the operation table. His internal discourse might be a purely linguistic entity subjected to a set of rules linking the term 'brain damage' with other linguistic terms such as 'car accident'. The discourse might refer or it might not. If it refers, it might be true or it might be false. In the mean time, one of the surgeons accidentally touches a specific area of the patient's brain with a needle. At the same moment, the internal discourse of the patient is interrupted by the vivid sensation of bright red and the linguistic term 'red' inserts in the middle of his mental discourse. The arbitrary insertion of the term 'red' is allowed neither by the rules of grammar, nor by the rules correlating the term 'brain damage' with other linguistic terms. The insertion is therefore something that transcends the internal logic of language. Completely absorbed by his thoughts, the patient failed to notice

that the surgeon touched his brain with a needle. In addition, he doesn't know anything about the structure of the brain and the mechanisms of perception. The patient might try to rationalise the sudden insertion of the term 'red' in the middle of his mental discourse. Note however how his mental discourse is unlawfully disrupted first and only afterwards the disruption is rationalised by appealing to further rules correlating the term 'red' with other linguistic terms. It is the sudden introduction of the term 'red' which calls these further rationalisation, thus determining the sudden change in the mental discourse, and not the initial discourse, with its terms and rules, which brought about the term 'red' and the new thread of thoughts that followed afterwards.

The example shows that some elements, even if represented in language and having an effect on the linguistic discourse, have an extra-linguistic origin. Metaphysical realism holds that intrusions occur independently of the will of the thinking agent and that they do not obey the rules of the language to which his or her thoughts are subjected. Thus, they fall outside any language game to which a reasonable agent might willingly be adhering.

The question that remains to be settled is whether the reality transcending the mind of the patient is objective. A common way to address the question is to appeal to the argument from intersubjectivity (Niiniluoto, 2000 pp. 36-41). Here is my own version of the argument. If the brains of two patients are simultaneously touched with a needle by the same surgeon, then, after consulting the testimonies of the two patients, presumably each thinking of different things, the surgeon can easily pinpoint the introduction of the term 'red' in the discourses of the two patients as correlating with the intrusion of his needle and not with the subject matter of the discourses of the patients. This is indeed a most interesting finding. Somehow, the surgeon's train of thoughts concerning the simultaneous brain surgery of two patients intruded the trains of thoughts of the two patients. Three independent discourses, each belonging to different minds and each involv-

ing their own terms and laws somehow intersected in the most miraculous manner. I say miraculous because the surgeon didn't communicate by means of language what he was doing to the patients, yet even in the absence of a linguistic connection, the three discourses intersected nonetheless. Further investigation from the part of the surgeon will most likely establish that he is able to repeat exactly the same intrusion over and over again in any patient independently of what the patient is thinking of and of how he rationalises the intrusion afterwards. The correlation between the term 'red' in the discourses of the two patients and the term 'needle' in the surgeon's discourse transcends at the same time the limits of any mind in particular.

Naturalistic accounts (Prinz, 2002; Kornblith, 2003; Murray, 2004) take advantage of this kind of examples and further argue, correctly in my opinion, that some linguistic terms, explicitly connected to one another by rules of language, are also extra-linguistically connected either to other linguistic terms (such as the 'red-needle' correlation), or to something transcending language altogether (the term 'red' irrupting in the discourses of the patients). Most often, the extra-linguistic connections are thought of in terms of causal chains mediated by perception mechanisms. Thus, perception anchors specific terms in an extra-linguistic reality and subjects the logical structure of a linguistic discourse to external influences.

### *10.3 Issues Related to Semantic and Epistemological Realism*

Whether the above considerations suffice to establish metaphysical realism is still a matter of debate. Since the realism-antirealism debate in philosophy of science is typically concerned with semantics and epistemology, for the purposes of this book, I will simply take metaphysical realism as a premise of the investigation.

The next three chapters are dedicated to the problem of semantic realism. As discussed previously, positivist verificationism amounts to an eliminative reductionism of theoretical terms

to observational ones. Strong verificationism usually implies that theoretical terms cannot refer to anything else than certain combinations of observables and, in this sense, it represents a form of semantic antirealism about the former. Of course, as statements about observables, scientific theories, models and propositions have a truth value. Nevertheless, they assert something about observables alone and do not point towards the existence of theoretical entities hypothesised by scientific explanations. Instrumentalists (Nagel, 1950; Fine, 1984) push matters further and argue that scientific theories and models serve solely pragmatic purposes and should not be taken literally, but rather as means to summarise and organise knowledge, as reliable methods or computational algorithms for predicting phenomena, etc. The argument here is that semantics doesn't add anything to scientific theories and models, which can function just as well without it. In contrast, semantic realists argue that semantics adds something to the scientific discourse. More specifically, it is thought that the entities, mechanisms or structures hypothesised by scientific theories and models underlie and determine the manifestation of empirical reality at the observable level (Bunge, 1973; Psillos, 1999; Niiniluoto, 2000). Thus, realists typically believe that a phenomenon can, or at very least could be, accessed and manipulated at two distinct levels, at the level of its observable manifestation and at that of its underlying structure, mechanism, etc.

Epistemological realism, discussed in more detail in the last chapters of the book, states that truth can be ascertained and contrasts with agnostic (van Fraassen, 1980; 1989; Laudan, 1984) and sceptical (Feyerabend, 1987) stances. Epistemological antirealists do not deny that theoretical terms refer, and not even that currently accepted theories and models are true, but rather that we cannot legitimately claim that we are in the possession of truth. In fact, most epistemological antirealists push matters further and argue that irrespective of whether theoretical terms refer or not, science can function by providing explanations alone, without bothering to

prove that the entities it postulates for explanatory purposes exist or not. Note that, unlike instrumentalists, epistemological antirealists endorse semantic realism in respect to theoretical terms contributes to the explanation; what doesn't contribute to the overall functioning of science is justification. On the realist end of the debate, it was initially argued that the claims made by scientists can be confirmed or falsified (Popper, 1959; Hempel, et al., 1965). However, since confirmation is usually only partial and since it is not clear to what extent the confirmation of some elements of a model justifies claims about the model being true, it is nowadays common to argue that the inferences and justification methods employed by scientists are reliable and tend to yield the truth (Maxwell, 1962; Smart, 1963; Hacking, 1983; Boyd, 1984; Psillos, 1999).

## CHAPTER 11

# THE PHYSICAL INTERPRETATION OF MENDEL'S GENETIC EXPLANATION

### *11.1 Conceptual Explanations: The Example of Mendelian Genetics*

The historical development of classical genetics provides a suitable study case for exploring issues related to semantic realism. The phenomenon under investigation is sexual reproduction. In a first time Mendel (1866) – as well as de Vries, Correns, and von Tschermak shortly after – observed patterns of phenotypic frequency from one generation to the next and represented them mathematically as proportions (Olby, 1985). The phenotypic frequencies constitute the empirical constraint on hypothesis formation. By analogy with the gravitational model, the patterns of frequency represent an essential empirical description of the phenomenon under study which that any theoretical explanation must entail as a conclusion in order to be empirically adequate.

Given this empirical constraint, Mendel hypothesised a rather complex story whereby entities called 'alleles' are somehow responsible for phenotypic traits. Each organism must possess two alleles, one inherited from each parent. The two alleles contribute to the manifestation of the phenotype, yet it is not always the case that an organism which inherited two different alleles from its parents has a mixed phenotype. Instead, some alleles are dominant, while others are recessive. The phenotype of the organism inheriting two dominant alleles is indistinguishable from the phenotype of an organism inheriting a dominant allele and a recessive one, meaning that, when present, the dominant allele determines the phenotype alone. The phenotype associated with recessive alleles manifests itself only if an organism inherits two copies of the recessive allele.

For reference, the classical case of complete dominance is illustrated in the diagram below:

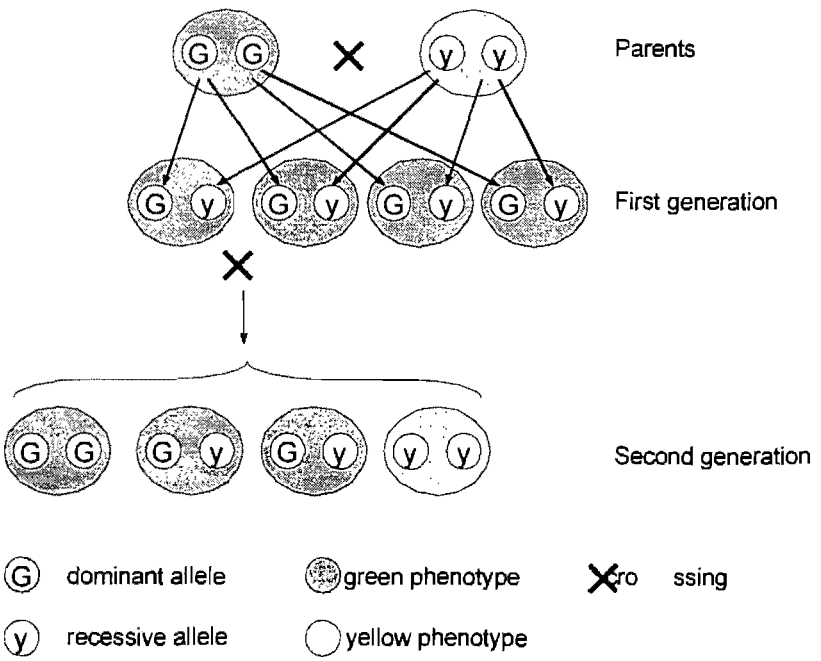



**Figure 19. Mendelian Inheritance**

Variations of the above explanatory strategy require that some alleles are partially dominant or that two alleles determine conjointly the phenotype.

By analogy with the Newtonian model discussed in the previous chapter, the 'allele story' constitutes the fundamental hypothesis. Since this is an entirely new hypothesis, custom-tailored for the needs of a particular body of empirical data, and not a pre-existing pattern of explanation applied to new phenomena, we don't have to worry about the distinction between fundamental and model-specific hypotheses. Different patterns of inheritance (partial dominance, inheritance for non-diploid organisms, etc.) are explained not by extending Mendel's initial hypothesis to these new phenomena, but by modifying Mendel's explanatory story in order to accommodate new phenomena case by case. Nevertheless, the family of explanations associated with Mendelian genetics shares a common theme, namely the supposition that there is something in each organism, namely the 'alleles', which determine its phenotype (this would be the fundamental hypothesis common to all Mendelian models); to this common supposition, different genetic explanations add some further assumptions specifying how many alleles are required and how they interact in order to determine the phenotype (the model-specific hypotheses).

Mendel's explanatory hypothesis is not explicitly formulated as a set of mathematical propositions and, as consequence the deductive character of Mendel's explanation is less evident. As noted on a previous occasion, Mendel's genetic explanation hypothesises a mechanism for heredity, and, at the same time, it is framed as a deductive consequence of a set of 'laws'. On the mechanistic side, Mendel hypothesises the existence of 'genetic elements', that is, of particle-like entities, that are transmitted via semen or pollen, that mix with their female counterparts and that later on segregate in a specific pattern before being carried on to the offspring. On the deductive side, the behaviour of the 'genetic particles', most notably their distribution along several

generations, is described by probabilistic 'laws'. These 'laws' 'dictate' how the genetic mechanism functions, that is, how the 'genetic elements' segregate following fertilisation.

From an intuitive point of view, if things happen as postulated by the hypothesis (i.e., each organism contains two alleles determining its phenotype, which segregate randomly between distinct gametes, subsequently fused together in order to produce the fertilised egg), the observed phenotypic distribution should obtain. Also, it seems that in developing his explanation Mendel applied the principles of combinatorial mathematics to discrete elements of inheritance called 'alleles' (Gayon, 2000). Furthermore, the hypothesis can be easily reframed in terms of combinatorial probabilities and given a rigorous mathematical formulation [see, for example, the Hardy-Weinberg law (Edwards, 1977)].

### *11.2 Instrumentalism and the Independence of Early Genetic Theories from Biochemistry, Molecular Biology and Developmental Biology*

The version of Mendel's genetic explanation presented in most genetics textbooks comprises no further hypotheses about what alleles are made of, where they might be located in the organism, how they replicate, segregate and determine the phenotype. In this sense, the mechanistic side of Mendel's explanation is incomplete.

It seems however that Mendel believed that microscopic 'elements' are transferred from parents to offspring during mating or fertilisation, and, if found in sufficient quantity, these 'elements' are responsible for the observed phenotypes (Gayon, 1998 pp. 105-146). Although the physical interpretation endorsed by Mendel was not retained during the subsequent development of genetics, this indicates that Mendel thought of his genetic theory along the lines of semantic

realism, that is, as hypothesising the existence of physical entities and of a mechanism of action, rather than a purely conceptual schema serving the sole purpose of explanation.

Since, initially, all the entities postulated by the genetic explanation (alleles, segregation, dominance, etc.) were purely hypothetical in nature, an instrumentalist interpretation was also envisaged. Johannsen was the first to use the term 'gene' in order to refer to the "*special conditions, foundations and determiners* [present in the gametes and by means of which the] *many characteristics of the organism are specified*" (1909 p. 124). What is less known is that, in defining the concept of 'gene', Johannsen also made a sharp distinction between the epistemic status of the notion of 'phenotype', pertaining to observable traits, and that of the notion of 'genotype', pertaining to the realm of what may be a strictly explanatory notion (Johannsen, 1909; Roll-Hansen, 1989). Throughout his life, Johannsen remained an agnostic about the material existence and constitution of the genotype, and, according to some authors, treated alleles as essentially instrumental constructs introduced for the purposes of explanation alone (Rheinberger, 2000).

The genetic explanation was also interpreted instrumentally in a second, weaker sense. Under this alternate interpretation it is not question of doubting the existence of a genetic material, but rather of understanding that the genetic theory proposed by Mendel simply assumes that 'alleles' are transmitted from one generation to the next and that they determine the phenotype without incorporating any further hypotheses as to how these two feats are achieved. According to Moss,

> "[w]*hat Johannsen called for in distinguishing between the genotype and the phenotype was a separation of the inheritance of Mendelian units from development, thereby constituting the study of genetics as an independent discipline.*"
>
> (Moss, 2003 p. 29)

It seems therefore that early geneticists understood quite clearly that, if interpreted along the lines of semantic realism, Mendel's genetic explanation is incomplete or, if we prefer to put it this way, has a limited explanatory scope.

This second sense in which the genetic explanation was used instrumentally appears quite clearly in Morgan's research:

> *"At the level at which the genetic experiments lie it does not make the slightest difference whether the gene is a hypothetical unit, or whether the gene is a material particle."*
>
> (Morgan, 1935 p. 3)

If Johannsen expressed doubts concerning the existence of a material substance responsible for inheritance, Morgan chose to abstain from formulating any hypotheses about the physical, chemical or biological 'mechanisms' via which genes are inherited and determine phenotypes, and concentrate instead on the more immediate problem of defining the relationship between genes and phenotypes.

The above considerations indicate that Mendelian and classical genetics are essentially about the probabilistic relationships between phenotypic traits and allegedly underlying genotypes (or what Moss calls 'gene-P') and much less about 'biological mechanisms'. This further indicates that

1) although few geneticists doubted the existence of a physical entities responsible for inheritance, early on, genetic theories received only a partial physical interpretation;
2) a sharp distinction was initially drawn between genetics and the budding new sciences of biochemistry, molecular biology and developmental biology.

### *11.3 A Partial Physical Interpretation for Mendel's Genetic Explanation and the Birth of Classical Genetics*

In the context of an instrumentalist approach, an explanation fulfils its goal in as much as it is empirically adequate in respect to some desired result – in this case, in respect to the ability to predict phenotypic distribution in a given population. The downside of instrumentalism is that, once a successful explanation is provided, research has no logical reason to continue. In the case of Mendel's explanation, to ask "How genes are inherited?" or "How genes determine phenotypes?" implies that there is more to be said, that a 'mechanism' of some sort must be specified, in short, that a an instrumental – or 'black-box' approach, if we so prefer – use of the genetic explanation is unsatisfactory or incomplete.

Historically, Mendel's explanation was received by the scientific community along the lines of semantic realism at least in the minimalistic sense that it postulated the existence of physical entities called 'genes'. Furthermore, the historical development of genetics blatantly contradicts instrumentalist views of science advocated by Carnap (1928; 1936) and, later on, by Nagel (1950). As noted above, this is not all that surprising since instrumentalism leaves no room for further experimental and technological developments, which is a rather unhappy consequence, considering that, most of the time, it is precisely the quest for experimental and technological control that motivates scientific investigation in the first place.

Half a century after the initial formulation of the genetic explanation, Sutton and Boveri (Sutton, 1903; Crow, et al., 2002) hypothesised a plausible physical interpretation based on the observation that, during meiosis, chromosomes segregate in a Mendelian fashion. A decade later, Morgan and his research group showed that some traits are specifically linked to the sex chromo-

somes, thus providing evidence supporting the Sutton-Boveri interpretation (Morgan, et al., 1915).

Morgan and his group are famously known for their work on jointly inherited traits. Morgan hypothesised that such traits are associated with alleles located on the same chromosome (hence 'genetic linkage'), meaning that each chromosome contains more than one allele (nothing surprising here since the number of chromosomes is extremely small in comparison with the number of phenotypic traits) and that a Mendelian distribution of phenotypes is possible only when genes/alleles are linked to different chromosomes. In this respect, Mendel's explanation was extended in order to accommodate the newly discovered fact that theoretically postulated alleles must be located on or associated with physically existent chromosomes.

By analysing minute divergence in the proportions of phenotypes associated with linked alleles, Morgan and his collaborators further discovered recombination phenomena and hypothesised crossing-over (process during which chromosomes exchange parts). Shortly after, Sturtevant and his collaborators showed that the frequency of recombination is proportional with the distance between the locations of the linked alleles, thus providing the first 'chromosomal maps' (Sturtevant, 1913; Morgan, et al., 1915). Painter further found a correlation between the displacement of genetic loci on chromosomal maps and visible changes in the banding pattern of giant salivary gland chromosomes of *Drosophila* (Painter, 1934), further strengthening the link between the genetic explanation and the Sutton-Boveri interpretation.

Eventually, experiments involving radiation-induced mutation (McClintock, 1929; Muller, 1951), chemical inhibitors of meiosis (Ravnik, et al., 1999), plasmid (Avery, et al., 1944) and chromosome (Dieter, et al., 2007) transfers, as well as a thorough classification of the syndromes associated with chromosomal aberrations showed beyond doubt that any interference with chro-

mosomal segregation and alterations of the chromosomal content of a cell/organism correlates with changes in the phenotype. As a result, it is now generally acknowledged that each set of alternate alleles in a diploid organism is associated with a precise place on a chromosome, namely with a gene – to be understood here in the classical sense of genetic locus –, and that several genes are arranged in a sequential order along each chromosome.

Some of the experimental data justifying the Sutton-Boveri interpretation is presented in the figure below:

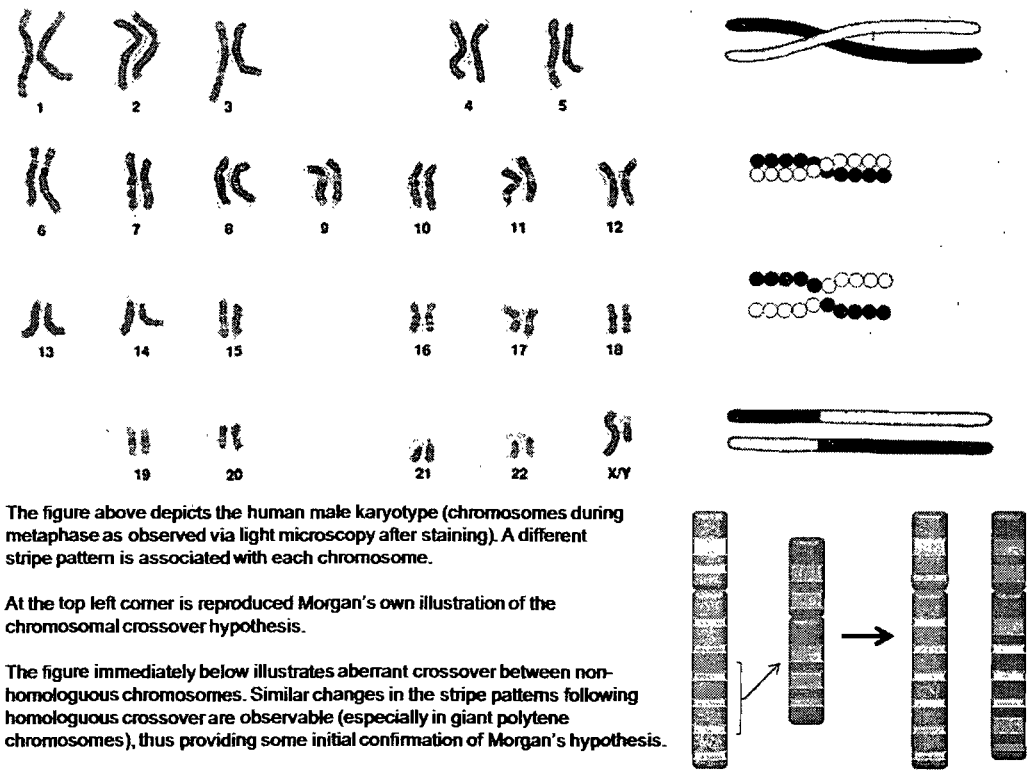



Figure 20. Chromosomal Maps

### *11.4 The Demise of the Instrumentalist Interpretation*

It is possible to imagine an alternative formulation of Mendel's explanatory story which does not claim to be a statement about empirical reality, but an algorithm or method of calculat-

ing phenotypic frequencies. By analogy with the mathematical representation of the correlation between the length of a metal rod and temperature as a linear function, we might view the genetic explanation along the lines of a probabilistic representation of available empirical data. Note however that an instrumentalist reformulation does not aim at the same goals as Mendel's explanation. An instrumentalist algorithm aims solely to yield the correct result; the main concerns here are the accuracy and the efficiency of the algorithm. In contrast, the genetic explanation, as understood by the scientific community, hypotheses the existence of entities named 'alleles'; therefore, the efforts are oriented towards identifying and further investigating the posited entities. The obvious questions in this case are "What are the genes made of?" and "How do they determine the phenotype of an organism?" Such questions have no meaning in the context of strictly instrumentalist interpretations.

By the third decade of the 20$^{th}$ century, researchers were able to physically locate the alleles in an organism and identify the physical path they follow during segregation and passage from parents to offspring. With these specifications, a chapter in the history of genetics was closed and another one begun. Classical genetics, as defined by Morgan, treats the observed phenotype as a function of the genotype. The explanation does not specify, not even at purely hypothetical level, how the genotype determines the phenotype. It is this question – a question transcending both the aim and the explanatory resources of Mendelian and classical genetics – that motivated the subsequent research in the field of genetics. Hence, a distinction should, and is usually made between classical genetics and the further advances of biochemistry and molecular biology.

The physical localisation of the alleles led naturally to the next step of the investigation: determining the material composition of the chromosomes. It is interesting to note that no further

over-arching explanatory story, at least none equalling the generality of Mendel's genetic theory, was ever proposed in the subsequent development of genetics. Rather, researchers relied on the expectation that an elucidation of the material composition of the genetic material, a matter of chemical analysis, will reveal essential clues as to the nature of the causal links joining genotype and phenotype. The project is clearly outlined by Muller in late 1920s. Convinced that genes must be material particles associated with chromosomes, Muller (1951) further conjectured that they are endowed with two properties: autocatalysis, or the ability to replicate themselves, and heterocatalysis, the ability to determine a phenotype. He also realized that "*the geneticist himself is helpless to analyse these properties further. Here the physicist, as well as the chemist, must step in*" (Muller, 1936 p. 214).

Chemists and biochemists did step in. In 1933, Brachet showed that chromosomes are made, among other things, of DNA. A decade later, Avery and his colleagues identified DNA as the 'transforming principle' capable of changing the phenotypes of certain bacteria (Avery, et al., 1944). Yet another decade later, Watson and Crick elucidate the chemical structure of DNA and predict that it must replicate semi-conservatively (Watson, et al., 1953), prediction soon confirmed by the Meselson-Stahl experiment (Meselson, et al., 1958). The structure and mechanisms responsible for the 'autocatalytic' property of genes were thus uncovered and shown to be heavily dependent on the specific chemical structure of DNA molecules. To this date, 1953 marks the year when one essential property of genes was shown to be essentially a matter of biochemistry. As one can easily guess, it is this strong dependence of 'autocatalysis' on the chemical structure of DNA that led several philosophers of biology to claim that the concept of 'gene' can be reductively defined as 'stretch of DNA' (Schaffner, 1969; 1967).

The mechanisms and processes responsible for the 'heterocatalytic' activity were likewise elucidated. Note however that, unlike 'autocatalysis', the 'heterocatalytic' activity cannot be characterised exclusively via biochemical analysis (i.e., as a function of chemical structure), but requires the further notion of 'molecular mechanism' (signalling, regulation, etc.). In a first time, Beadle and Tatum (1941) showed that genes code for proteins and hypothesised that the enzymatic activity of proteins is responsible for the phenotype. During the 1960s, various research groups work to uncover the mechanisms and processes leading to the 'expression' of DNA as proteins. Three decades later, Fiers and his colleagues determines the structure of the gene coding for the coat protein of the phage MS2 (Fiers, et al., 1971). Later on during the 1970s, it becomes clear that most genes are regulated at the level of transcription regulation; several other levels of regulation are also discovered, including splicing, frame shifting, etc.

The tables below highlight some of the milestones in the development of modern genetics. A schematic representation of the currently accepted model of the causal links between genetic makeup and phenotype is also provided. More can be found in Moran, part four (Moran, et al., 1994); for a discussion of the historical development of molecular genetics see Darden (1991), Carlson (1967), Waters (1994).

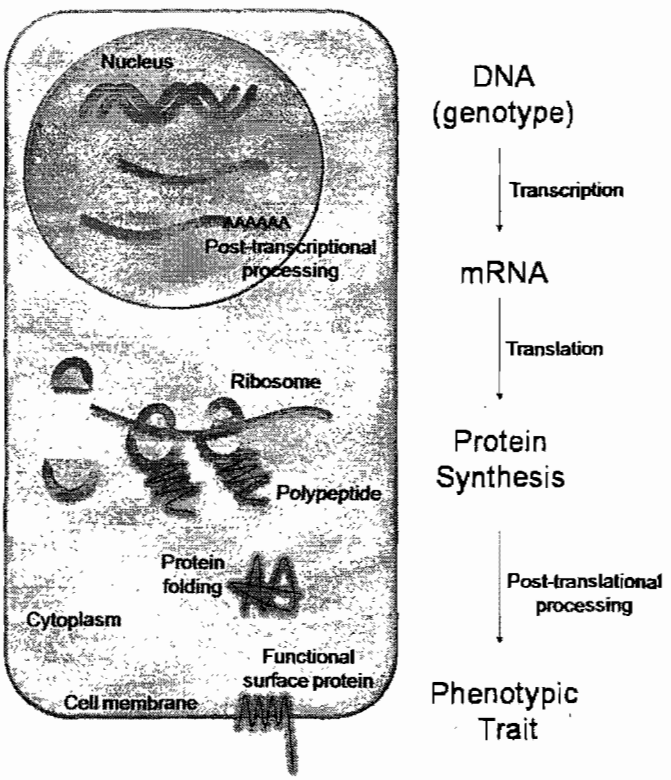



Figure 21. History of Genetics & Central Dogma of Molecular Biology

## CHAPTER 12

# THE CONTINUITY AND CONVERGENCE OF SCIENTIFIC KNOWLEDGE

### *12.1 Reference Commensurability vs. Theoretical Continuity*

The widespread experimental practice in sciences argues against instrumentalism, which must be confined to highly theoretical branches of science specifically aiming to provide an enhanced, easier to compute or just more elegant mathematical treatment of scientific problems. Nevertheless, it does not follow from here that semantic realism makes the unanimity. Quite on the contrary, the issue is still hotly debated, usually in relation to the continuity of scientific knowledge from one theory to the next.

The theories, models and explanatory stories accepted today evolved through several versions before reaching their current textbook formulation. Presumably, they will continue to change in the future; science is an ongoing process. But if theories change, then it is not clear

how reference, meaning, ontology and knowledge in general are continuous and compatible from one theory to the next.

Notoriously, Kuhn argues that terms appearing in both classical and relativistic mechanics, considered to be two successive theories explaining a common set of phenomena, do not have the same meaning and do not refer to the same things. He exemplifies his claim by discussing the status of mass in mechanical theories:

> "*Newtonian mass is conserved; Einsteinian is convertible with energy. Only at low relative velocities may the two be measured in the same way, and even then they must not be conceived to be the same*".
>
> (1970 p. 102)

The argument here is that mass cannot be the same entity in classical and relativistic mechanics because the 'classical mass' is conserved while the 'relativistic mass' is not, the underlying idea being that conservation defines some essential property of mass.[28]

Kuhn alludes at two distinct problems. The first one pertains to 'reference commensurability'. I showed on a previous occasion that we don't have to worry about the lack of meaning continuity between the cognitive and neurological manifestations of depression. To accurately evaluate contingencies does not mean to suffer from a serotonin-norepinephrine imbalance; in fact, to accurately evaluate contingencies doesn't even mean to be depressed. There is no reduction between theoretical terms, or between theoretical and observational terms. Nonetheless, in

[28] Against this conception, whereby mass is defined via to a list of properties, Mach (1893 pp. 266-267) proposes an operational definition whereby mass is an indicator of whether two physical objects can be interchanged such that changes in motion after a mechanical interaction remains identical. If a particle is accelerated from relative rest to a speed high enough for relativistic effects such as length contraction and time dilation to become manifest, does Mach's definition of mass still hold? Will the resulting motions of high-speed collisions still be the same for two bodies if it has been already established by low-speed collision experiments that the two bodies have the same mass? The answer is "Yes". Nowhere in the definition of mass is it further specified what are the resulting motions, what is the total mass of a mechanical system or that the numerical value of the measured mass must remain the same. This definition works equally well in classical and relativistic mechanics and refers to the same phenomenon of 'mechanical interchangeability' of two physical bodies.

the absence of a unified psychological theory, cognitive psychology and neuropsychology continue to refer, describe and assert something about the same subjects, operationally defined and selected by means of the same 'depression test' experimental protocol.

The other problem pertains to 'theoretical continuity'. Coreference does not guarantee the possibility to translate terms belonging to a theory in the language of another theory; in other words, coreference is not necessarily paralleled by theoretical continuity. Ideally, scientific progress is characterised by both coreference and theoretical continuity, such that new theories are able to recuperate the successes of older theories and ultimately provide a more general basis for understanding empirical reality.[29]

### *12.2 Arguments against the Convergence and Continuity of Scientific Knowledge*

According to Fine, the classical realist argument for the continuity and convergence of scientific knowledge makes use of the fact that, at any point in time, there are only a handful of related theories competing as true explanations for a given phenomenon. Fine frames a version of the realist argument, which he attributes to Boyd, as follows:

> "*it is reasonable to restrict one's search for successor theories to those whose ontologies and laws resemble what we already have, especially where what we already have is well confirmed*".

[29] Perhaps the most discussed example in the literature is that of Maxwell's electromagnetism. Unlike the dual cognitive and neurological description of depression, which merely corefer, Maxwell's electromagnetic theory is thought to have evolved with considerable modifications from previous ether mechanical models. Some realists argue that the ether, seen as the hypothetical medium responsible for the propagation of light, 'pointed to' or 'referred' from the very beginning to certain variations in the electromagnetic field (Psillos, 1999 pp. 130-143). The argument is that both the ether and the electromagnetic field share some features which explain the propagation of light. In contrast, other philosophers of science seem to think that the transition is characterised by theoretical continuity in the absence of co-reference and argue for structural realism, that is, the view that only the mathematical formalism survives from one theory to the next, and not a physical resemblance of the entities which the theories are concerned (Stein, 1989; Worrall, 1989). Finally, some argue that it is not reference which is conserved from one theory to the next, but rather bits and pieces of information amounting to an approximate truth (Saatsi, 2005).

This raises three questions:

> *"(1) why only a small handful out of the (theoretically) infinite number of possibilities? (2) why the conservative family resemblance between members of the handful and (3) why does the strategy of narrowing the choices in this way work so well?"*
>
> (1984 p. 87)

The realist's answer to all three questions is that scientific theories are approximately true and therefore no new theory can depart significantly from current theories.

Alternatively, Laudan reframes the realist argument in terms of empirical success:

1. *If scientific theories are approximately true, then they typically will be empirically successful.*
2. *If the central terms in scientific theories genuinely refer, then those theories generally will be empirically successful.*
3. *Scientific theories are empirically successful.*
4. *(Probably) theories are approximately true and their terms genuinely refer*

This argument from the success of science further links to the continuity of scientific knowledge:

1. *If the earlier theories in a 'mature' science are approximately true, and if the central terms of those theories genuinely refer, then later, more successful theories in the same science will preserve the earlier theories as limiting cases.*
2. *Scientists seek to preserve earlier theories as limiting cases and generally succeed in doing so.*
3. *(Probably) earlier theories in a 'mature' science are approximately true and genuinely refer.*

(Laudan, 1984 p. 220)

Nevertheless, Fine retorts, instrumentalism can account for these observations just as well and without relying on the further unjustified assumption that "*confirmation is a mark of an ap-*

*proximately correct ontology*". Presumably, similar theories, used instrumentally, say, as algorithms for generating predictions, will yield similar results, and therefore share similar degrees of empirical success. Hence, convergence does not necessarily entail realism (Fine, 1984).

Laudan attacks the realist argument from a different angle. He relies on the notion that if a theory is able to successfully explain, this does not automatically entail the existence of the terms postulated by its explanatory models. His argument consists largely in showing that the history of science provides a wealth of examples of theories that offered persuasive, yet utterly false explanations. Since the unobservables postulated by models of these theories were not incorporated in the ontologies postulated by later theories, theories about the same phenomena fail to consistently converge towards the same 'ontological picture' (Laudan, 1984; 1996).

### *12.3 Overlapping Empirical Constraints*

Typically, theories are thought to be similar in respect to their mathematical or logical formalism, to the explanatory strategies they introduce or again in respect to the unobservables they postulate. As a realist, Boyd has in mind primarily a similarity concerning the unobservables and the overall ontology of a theory (Boyd, 1984; 1990).

Beside the above criteria, theories can also be compared from the standpoint of the empirical constraints they must initially satisfy in order to achieve an initial level of empirical adequacy. Presumably, two theories aiming to explain similar phenomena cannot fail to satisfy common empirical constraints. More so, given a sequence of more and more comprehensive theories, it can further be argued that in as much as newer theories tend to cover at least some of the phenomena previously explained by the older theories, they inevitably have to take into account identical experimental possibilities and impossibilities.

From the standpoint of historical development, it can be further argued that a theory can be modified without being completely abandoned. New theories are built from scratch only if they refer to and aim to explain radically different phenomena, described in terms of different observables. Otherwise, already existing theories are modified by targeting specific elements within the initial theory. Theories are often mended and recycled, dismantled in a set of general principles and/or explanatory strategies, then recomposed back after adding, removing or altering some principles and strategies, as dictated the newly imposed empirical constraints.

The existence of a common background of 'input' empirical constraints provides a partial answer to Fine's objection. Aside the resemblance of theories in terms of explanatory strategies, formalism, postulated unobservables, etc., theories about the same phenomena must also satisfy common empirical constraints. I propose therefore that a progression towards truth does not follow from the convergence of theories alone – which, as Fine points out, speaks equally well in favour of realism and instrumentalism – but also from the fact that additional empirical constraints are taken into account by newer theories. Ideally, newer theories satisfy new empirical constraints in addition to the previously established ones constraining older theories; thus, newer theories tend to achieve a higher initial level of empirical adequacy than older theories. Now, assuming that there are no radical changes in terms of postulated unobservables from one theory to the next, then an increase of the level of empirical adequacy amounts to net increase of the total empirical content. In some cases, this suffices to give a realist interpretation a slight advantage over a purely instrumental one.

### *12.4 The Case of Genetics*

An increase of the overall empirical content of a theory is bound to occur every time a theoretical consequence given certain hypothetical conditions is subsequently verified and becomes an empirically established consequence that must be entailed by a model of the theory if a revised version of the theory in question is to be empirically adequate. Thus, the strategy here is to ensure that some parts of the theory acquire an empirical significance they didn't have before.

The immediate objection to this approach is that an increase of the overall empirical significance of a theory doesn't target key unobservables hypothesised by a converging pattern of explanation spanning several theories. The added empirical content concerns side predictions and aspects of the theory, increasing the instrumental value of the core explanatory pattern without supporting in any obvious way realism about the unobservables associated with it. In fact, in as much as theoretical terms cannot be successfully reduced an replaced by observable ones, key unobservables are bound to remain just as unobservable and, in this sense, an increase of the overall empirical content of a theory fails to tilt the balance in favour of realism.

The history of genetics constitutes a powerful counterexample to the above objection. It shows that it is possible to provide key unobservables with an empirical significance without reducing or equating them to a set of observable empirical phenomena. The figure below depicts some of the reference relationships (dotted arrows) between the theoretical terms (orange boxes) introduced by Mendel's initial explanation and empirical/experimental observations associated with classical and molecular genetics (blue boxes):

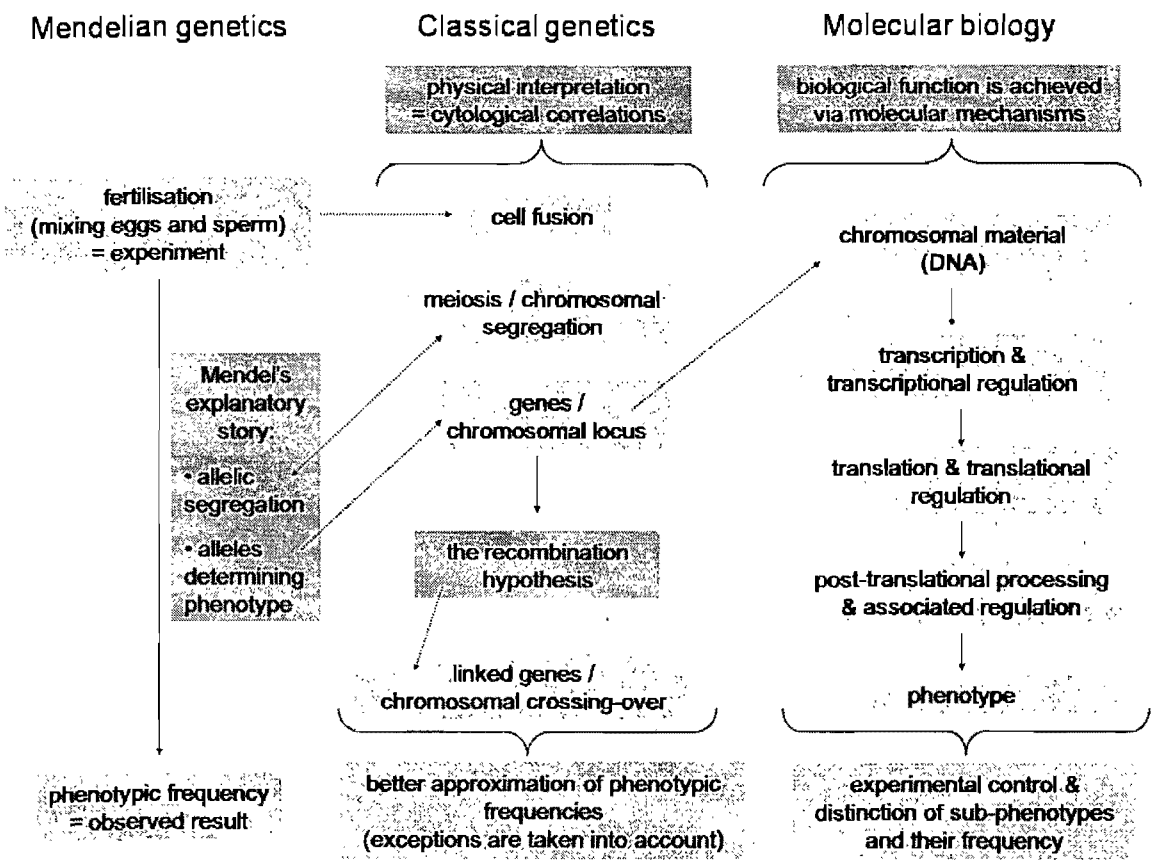



Figure 22. The Development of Genetic Theories

Interestingly enough, the physical interpretation offered by classical genetics is incomplete and non-uniform. The term 'allelic segregation' is reduced to and identified with the chromosomal segregation occurring during meiosis (the biological process whereby chromosomes are distributed among haploid gametes); in this case, a theoretical term is reduced to observational terms. The term 'fertilisation', which already has a physical interpretation, is further correlated with gamete fusion, whereby diploid organisms are generated; in this case, the network of experimental and empirical data is extended further without providing a physical interpretation and without involving a reduction of one term to another. Finally, the theoretical term 'allele/gene' is correlated, but not reduced to the notion of 'chromosomal locus'; in this case a non-reductive physical interpretation is provided. Of the three examples provided, only the first

matches the classical, logical positivist pattern of verification whereby a theoretical construct introduced for the purpose of explanation acquires a strictly empirical interpretation.

Classical genetics doesn't specify what exactly genes refer to, what they are from a physical point of view and how they fulfil the functional role genetics attributes them. Nevertheless, reference is not completely absent either. Whatever genes are, they are physically tied down to specific places on chromosomes, which they follow through the subsequent processes of recombination, meiosis and fusion. Likewise, from the standpoint of molecular genetics, we can say that, as a general rule – that is, leaving aside more exotic cases such as developmental processes mediated by gradients of transcription factors in the fertilised egg –, genes are inseparable from DNA. At the same time, it is also the case that the term 'gene' cannot be easily reduced to that of 'DNA sequence'; at any rate, it is impossible to derive the auto- and hetero-catalytic properties of alleles from the properties of that which is found at a certain chromosomal locus, in occurrence, a stretch of DNA (Rosenberg, 1978; Kitcher, 1982; 1984).

Even though the physical interpretation hypothesised and submitted to verification by classical genetics is a partial one, it succeeds in imposing a novel empirical constraint onto future genetic theories. It became clear very early that differences in the chromosomal makeup of an organism correlate with differences in phenotype; furthermore, experiments showed that any interference with the chromosomal makeup of an organism leads to radical changes in phenotype. This data served at the time as a partial confirmation of Mendel's conjecture and as a further empirical constraint on future genetic theories: future explanations must take into account the fact that the fate of genes and their ability to determine the phenotype is tightly linked to the fate of chromosomes.

It follows that the above considerations that the continuity from Mendelian to classical to molecular genetics doesn't hinge on a reductive physical interpretation, whereby theoretical terms are reduced to observables, but rather on something along the lines of 'approximate reference': it is true that we cannot have alleles without genes, yet it is not entirely the case that alleles are genes either. In this sense, many authors like to argue that alleles are 'localised' onto (Darden, 1991; Wimsatt, 2006), as opposed to identified or reduced to (Schaffner, 1969) chromosomes and DNA.

This has an interesting consequence with respect to the realist – instrumentalist debate. According to Boyd, the genetic three theories must be approximately true because they convergence towards a common explanatory pattern postulating the existence of something physical transmitted from parents to offspring and responsible for inheritance. Against this, Fine would retort that this kind of convergence may have nothing to do with the truth of the theories in question, as it can also be due to a desire to preserve successful patterns of instrumental explanation. I argue that the genetic explanation is preserved from a theory to the next not because it is instrumentally valuable in respect to the derivation of certain predictions, but because some core elements associated with the genetic explanation acquired an experimentally tangible empirical significance they didn't have before.

### *12.5 Fundamental vs. Model-Specific Ontologies*

Laudan's objection that science does not advance via a constant and continuous convergence towards the same ontological picture can be defused more easily. Consider for example an attempt to extend the genetic explanation to a phenotypic trait that is not inherited, but acquired, say, a complex psychological trait. The ontology associated with the explanation would then be

false. Geneticists would end up hypothesising a set of genes and possibly a whole chain of transcriptional/translational control as well, none of which really exists. Does this mean that there are no alleles, no genes and no molecular mechanisms? No. This simply means that there are no genes and no definite genetic molecular mechanisms underlying that particular phenotype. The general ontology associated with genetic theories is true in the sense that there are genes and that molecular mechanisms are responsible for expressing them as phenotypes, although some model-specific hypotheses whereby genes and genetic control are associated with an individual phenotype may turn out to be false.

In more general terms, there is something which Laudan overlooks in his exposition of the various mistakes of science, such as the hypothesised existence of the phlogiston or that of the ether. Strictly speaking, these are not new theories, but models aiming to extend the domain of application of already existing theories. The hypothesised unobservables concern solely the extension, not the initial, or core theory. The phlogiston was introduced as a mere generalisation of the observed fact that, in order to burn, a body must contain or be made of an inflammable material. The initial observation is not false. In most combustion reactions, an input of energy is required in order to weaken specific chemical bonds, whose final breaking releases more energy than the initial energy input. The empirical support for the phlogiston mini-theory relied on the observation that the burned leftovers weight less than the original body. As we all know, the measurements were hardly accurate. Similarly, the ether was introduced as a physical medium for the propagation of electromagnetic waves, that is, in order to provide a mechanical model for electrodynamics. In this case, experimental data favoured a very different way of unifying mechanics and electromagnetism, namely special relativity. Analogous comments apply to the vari-

ous other examples discussed by Laudan, most of which amount to mini-theories of the phlogiston type.

Often times, scientists try to extend the explanations a theory has to offer to new phenomena. Each extension of a theory to a new phenomenon provides a model for that phenomenon. Some of these models are verified, while others are falsified. Sometimes, models merely perpetuate the core ontology associated with a theory to new phenomena. In the above example of the extension of the genetic explanation to complex psychological traits, the model introduces new unobservables – namely, genes associated with acquired traits – which are of the same kind as the unobservables associated with the core explanation. In other cases, new kinds of unobservables, such as gravitational forces acting at distance, are introduced in order to model the target phenomenon. One way or the other, these unobservables pertain to model-specific hypotheses; they are not central to the theory, nor are they in any way permanently attached to the theory.

As discussed on a previous occasion, the falsification or lack of confirmation of some models does not automatically entail that the theory is false (Lakatos, 1970; Lakatos, et al., 1976; Putnam, 1991). That some models hypothesising the existence of specific unobservables are falsified does not entail that other models derived from the same theory and postulating the existence of different unobservables cannot be corroborated.

In the initial stages of modelling, there may be as many ontologies as there are models of a given phenomenon, and the final ontology associated with a theory is bound to change as some models are abandoned in favour of others. More so, in as much as the models in question are mutually exclusive, their ontologies will also be incompatible. For example, in the case of planetary motion, there are several possibilities: gravitational forces, frictional forces and their associated

media, jets, angels or volcanoes, etc. At the same time, beneath the divergence of model-specific ontologies, there is an identity of the 'ontological picture' associated with the fundamental hypothesis common to all models. In this particular example, anything that might exert a force constitutes a viable option, meaning that the divergence is at the level of individual models, while the core theory, the explanation it provides (in this case, something along the lines 'all accelerated motion is caused by a force acting on the moving body') and the basic ontology associated with it (an ontology of forces, masses and motions) remain the same across all models.

In summary, the overall ontology associated with a theory doesn't have to be, and often just can't be a coherent one converging towards a unique 'ontological picture'. For one thing, it cannot be known a priori how far a theory can be extended. And second, it is often the case that different preliminary models point towards different and potentially incompatible ontologies. Still, despite this uncertain diversity of unobservables attached to each model in particular, they are all built on the stable ontology associated with the fundamental hypothesis common to all the models of a theory. I think it is this core ontology which Boyd has in mind when he observes that, in any given domain of investigation, theories tend to converge towards the same general 'ontological picture'.

## CHAPTER 13

# REDUCTIONISM

### *13.1 Two Kinds of Reduction*

In the philosophical literature on genetics, and biology in general, the problem of convergence and continuity of scientific knowledge is tightly linked to the issue of reductionism.

Nagel (1974 p. 907) defines inter-theoretical reductionism as the possibility of deducing the predictive statements entailed by a theory starting from the premises of a different theory. More specifically, theory $T$ is reducible to theory $T'$ if $T'$ is at least as well systematized as $T$ and all observational statements explained by $T$ are also deductive consequences of $T'$ (Kemeny, et al., 1956). If the theories in question are formulated in a mathematical language, the deduction is mathematical, as exemplified by the derivation of Kepler's laws from Newton's mechanics and the inverse square law for gravitational attraction. If the theories are not formalised or only partially formalised, which is the case of most, if not all theories in biology, then cannot be a deduc-

tion since there is no explicit deductive system allowing us to prove that one statement entails another. Nevertheless, we may still speak of a reductive derivation assuming that it is possible to replace all the terms of reduced theory with terms proper to the reducing theory and show that under this substitution the reducing theory can account for all the relationships between these terms as postulated by the reduced theory.

In addition to this criterion of interchangeability, which merely makes the two theories equivalent respect to their ability to save the phenomena, the reducing theory must also have the advantage of accounting for cases in which the reduced theory fails to agree with empirical observations or at least provide a more universal formulation applicable phenomena extending beyond the reach of the reduced theory (Nagel, 1961 p. 136; Schaffner, 1969 p. 340).

In contrast to the above 'successional' reduction, usually illustrated by examples drawn from physics, stands an 'inter-level' kind of reduction (Nickles, 1973; Schaffner, 1967; Wimsatt, 1976; Wimsatt, 2006) whereby "*reductive explanations are driven by referential identities or localizations*" (Wimsatt, 2006 p. 450). The relationship between Mendelian genetics and the classical genetics is considered to belong to this second category, and is viewed either as a localization of alleles to chromosomal loci (Darden, 1991), or as a straightforward gene – chromosomal DNA identity (Schaffner, 1969).

Interestingly enough, in as much as classical genetics offers a particular physical interpretation of Mendel's explanatory story, the former can be said to be a special case of the more general pattern of explanation provided by the latter. More so, since it is possible to provide a series of valid physical interpretations for Mendel's notion of allele – alleles can be associated with chromosomal inheritance, mitochondrial inheritance, plasmid-based inheritance, etc. – and as-

suming, as Nagel does, that generality is one of the hallmarks of reducing theories, we might be tempted to conclude that classical genetics reduces to Mendelian genetics rather than vice versa.

It seems however that this counterintuitive result can be defused if we take into account the fact that further investigation can be conducted only in as much as an explanation receives a physical interpretation. If multiple physical interpretations become available, then this simply means that several avenues of research are open. Conversely, if cut from the prospect of acquiring a physical interpretation, a theory is bound to remain purely instrumental, and therefore experimentally sterile. Presumably, the advantage stems from the fact that each physical interpretation renders the theory more apt to account for minute empirical observations. For example, Mendelian genetics can explain linked phenotypes only by further hypothesising that alleles are somehow linked via a yet to be defined mechanism; in contrast, classical genetics provides a more complete explanation by specifying how alleles are linked and how they segregate together. Thus, although an inter-level reduction leaves the door open to the possibility of multiple realisability, physically interpreted theories are typically considered to be a preferable alternative to purely conceptual explanations.

### *13.2 The All-Important 'Molecular Details'*

Waters points out that, as a general rule, "*geneticists did not understand how (i.e., by what mechanism) individual genes made their contributions to phenotype*" (1994 p. 171). The historical development of biochemistry and molecular biology is tightly linked to the elucidation of the physical connection linking genotype and phenotype.[30] It seems therefore legitimate to

[30] Schaffner (1994 p. 201) makes a similar point about immunology: on the background of Burnet's 'clonal expansion' theory, the 'genetic recombination' hypothesis further postulates the mechanism whereby genetic diversity is achieved. For Burnet, the notion of a 'generator of genetic diversity' is essentially a theoretical term serving an ex-

conclude that the relationship between classical genetics and molecular biology can likewise be described as being primarily a matter of 'inter-level' reduction (Wimsatt, 2006 pp. 450-452).

In line with this remark, Hull points out that, initially, classical genetics and biochemistry/molecular biology started as disciplines covering different aspects of heredity (Hull, 1974). However, adverse to the reduction thesis, Hull further concludes that the clarifications brought about by molecular biology specify the causal chain linking the genotypes of an organism to its phenotype without directly contributing to the genetic explanation (Hull, 1979). Hence, Rosenberg concludes,

> "[a]*ntireductionism must claim that there are at least some explanations in functional biology that cannot be completed, corrected, or otherwise improved by adducting considerations from molecular biology*".
>
> (2007 p. 129)

It seems that Hull is not against the idea that molecular biology offers the tools required to experimentally manipulate some of the entities hypothesised by classical models and therefore offers some experimental proof of the models postulated by classical genetics, but he is vehemently opposed to the idea that molecular biology offers a more comprehensive theoretical basis from which the classical genetic models and probabilistic laws of phenotypic and genotypic distribution can be derived as special cases.

Several authors back up this conclusion by arguing that it is impossible to derive the laws of phenotypic distribution associated with classical genetics from the laws of biochemistry (Rosenberg, 1978; Kitcher, 1984). Given this impossibility, which rules out the Nagel-style reductionism defended at some point defended by Schaffner (1969), Hull pushes the argument a

---

planatory purpose. In contrast, the various hypothesis concerning as to what this generator might be provide a physical interpretation.

step further and dismisses the 'inter-level' reduction of classical genetics to molecular biology as 'trivial'. Now, I take it that Hull is not aiming here to dismiss the achievements of molecular biology. I think that what he means to say is that the 'inter-level' reduction of classical genetics to molecular biology is 'trivial' in the sense that despite its contribution towards the elucidation of the mechanisms behind the auto- and hetero-catalytic properties of alleles, it does not provide a more accurate or more generally applicable explanation of observed phenotypic distributions from one generation to the next.

Hull successfully conveys the idea that although molecular biology has a wider explanatory scope which includes, or at very least extensively overlaps with the narrower explanatory scope of classical genetics, in respect to the initial explanatory scope of classical genetics, molecular explanations don't have much to add. The core idea behind the argument is of extreme value for anyone interested in the study of the history of genetics. On the other hand however, Hull's antireductionist argument is accurate only at the first approximation. The truth is that the 'molecular details' turned out to be relevant in respect to the explanatory scope of classical genetics in the most surprising ways.

Allow me to clarify my remarks by means of a simple example. The 'red eye' phenotype in *Drosophila* depends on the production of a pigment whose synthesis requires a chain of biochemical reactions involving more than one enzyme. This shows that the expression phenotype depends on two kinds of constraints:

a) the synthetic pathways responsible for their synthesis must present and functional; building materials, such as amino acids, must be available; finally, the general transcriptional/translational machinery responsible for the synthesis of the required enzymes must be intact

b) all the enzymes specifically required for the synthesis of the pigment in question must be present and functional.

Condition a) is usually met by default, for any organism incapable of sustaining essential metabolic pathways as well as transcription and translation is bound to die early in the development.[31] Thus, although these fundamental functions of the cell are determined genetically, loss of functionality mutations in the genes encoding for the basic metabolic apparatus are lethal mutations; since they are lethal, they never reflect in the phenotype of populations and notoriously fail to enter the domain of study of classical genetics. Alone, this observation points out that there are phenotypes which elude the methods of investigation of classical analysis. Specifically, all phenotypes that result in the death of an organism before it reaches sexual maturity cannot be investigated via the breeding techniques proper to classical analysis and that despite the fact that the phenotypes in question are genetically determined. It follows from here that molecular analysis has something to say about minute discrepancies in the actual phenotypic distributions, discrepancies which classical analysis systematically fails to explain.

The limitations of classical analysis don't stop here. Even assuming that the explanatory scope of classical genetics excludes juvenile lethal phenotypes, problems can still arise. From a molecular perspective, constraint b) states that a loss of functionality mutation in the gene encoding for enzyme E1 or a mutation in the gene encoding for E2 or .... or any combination of these mutations is bound to result in a loss or considerable reduction in the production of the pigment and therefore in a loss of the 'red eye' phenotype.

From the standpoint of molecular analysis, the mapping of genes via classical techniques cannot always pinpoint individual genes. In fact, classical analysis often identifies clusters of

[31] The only exception to this rule would be, say, a fly which has all the essential molecular apparatus intact, as well as all the enzymes required for 'red eye' phenotype, yet fails to display this phenotype because it temporarily lacks a some amino acids or vitamin co-factors in the diet.

genes jointly required for the manifestation of a phenotype. Many classical explanations are adequate despite the fact that they provide only a crude genetic map simply because the populations used in most genetic studies carry a loss of functionality mutation in only one of the genes required for the expression of a given phenotype. Furthermore, many genes required for a given metabolic pathway are clustered together in what classical genetics identifies as a unique chromosomal locus and often depend on common mechanisms of gene activation and expression.

In short, from a molecular point of view, classical analysis happened to yield the correct answers because some special requirements happened to be met. The diagram below illustrates the molecular explanation of a hypothetical phenotype dependent on the expression of a pigment requiring enzymes E1, E2 and E3 for its synthesis:

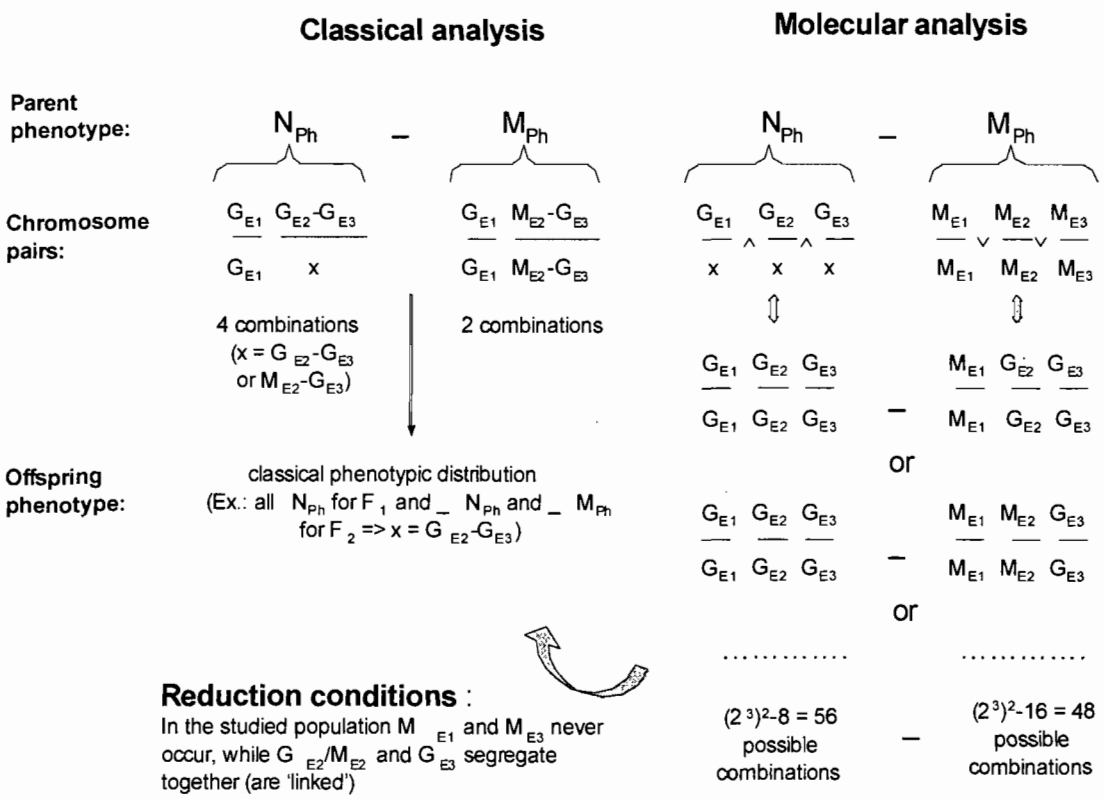



Figure 23. Classical vs. Molecular Analysis

Knowledge about the biochemical pathway leading to the synthesis of the pigment has a huge impact on the genetic analysis. Since three enzymes are involved in the synthesis pathway, it is hypothesised that at least three distinct genes must be expressed in order for these enzymes to be produced. $G_{E1}$, $G_{E2}$ and $G_{E3}$ are identified as the genes encoding for functional enzymes, while $M_{E1}$, $M_{E2}$ and $M_{E3}$, occurring naturally or created in the lab, encode for mutated, non-functional versions of the enzymes. According to a molecular analysis, the normal/wild-type phenotype $N_{Ph}$ must have at least one copy of the genes $G_{E1}$, $G_{E2}$ and $G_{E3}$ (in classical terms, the wild-type allele is 'dominant'), while a mutant phenotype $M_{Ph}$ must have at least one gene mutated on both chromosomes (the mutant allele is 'recessive'). Consider now that $G_{E1}$ and $G_{E3}$ are never mutated in the populations accessible to classical analysis and that $G_{E2}/M_{E2}$ and $G_{E3}$ are located close to each other on the same chromosome (they are closely linked, to the point that recombination is extremely infrequent). It follows from here that classical analysis, which is limited to naturally available genotypes, cannot distinguish between $G_{E2}/M_{E2}$ and $G_{E3}$, nor can ever establish that there is a third gene $G_{E1}$ also involved in the expression of the phenotype.

This example demonstrates two things. First, it shows that that classical genetics can make mistakes about the number of genes necessary for the expression of a phenotype. This is mainly due to the fact only molecular biologists have in their possession the experimental tools necessary to individually mutate, restore and insert genes, that is, to create mutants which do not occur in any known population, although they could exist in unstudied populations and can always arise in any population as a result of spontaneous mutations in the genetic material of the gametes. Molecular analysis spells out the mechanism leading to the expression of a phenotype, and in doing so it often ends up saying something about the genotype explanation of the studied phenotype. It follows from here that molecular biology explicitly competes with classical genet-

ics within the same explanatory scope and therefore the view according to which classical genetics and molecular biology run in parallel is false.

The second point is equally important. By introducing three genes instead of one, the fundamental explanatory hypothesis initially introduced by Mendel is not in any way altered. Instead, molecular analysis simply ties it down to a more accurate and more detailed physical interpretation imposed by the partial elucidation of the mechanism responsible for the expression of the phenotype. This fits a typical, Nagel-style reductive scenario whereby

i. molecular biology can explain the bulk of actual phenotypic distributions already explained by classical genetics, as well as make successful predictions about mutants and populations thus far inaccessible to classical analysis;

ii. molecular biology can explain why classical explanations work despite the fact that they where wrong about some key element (in this case, the precise number of genes associated with the phenotype under investigation).

### *13.3 The Issue of Simplicity*

A closely related anti-reductionist objection capitalises on the one – many correspondence between classical and molecular gene. For instance, Rosenberg (1985 p. 101) gives the same 'red eye' example in order to show that molecular is "*hopelessly complicated*", while classical genetics is so elegantly simple. Even more perplexing, Kitcher (1984) is deeply concerned that "*our feeble human brains*" might not be able to handle the increased number of genetic combinations associated with the genetic maps provided via molecular analysis. Finally, Hull (1972) talks about an infinite disjunction problem in relation to the one – many relationship between the classical and molecular understanding of the term 'gene'.

This family of objections points out that the classical analysis is simpler than the molecular analysis. I couldn't agree more. In the hypothetical case illustrated above, instead of having 4 genotypes underlying 2 possible phenotypes, there are 64 potential genotypes assuming all genes segregate independently. Things get more complicated, although not to the point that we have to worry about our 'feeble' capacities of computation. Genes are not in an infinite number and do not code for omnipotent proteins, hence one cannot appeal forever to the argument "we know so far that there are *n* genes determining this phenotype, but there could always be another gene involved". Besides, even if this were the case, our current knowledge of molecular mechanisms and their genetic component is not in any way invalidated or falsified.

Above all, these authors seem to forget that the results yielded by a classical analysis, while successful when applied to a given population today, may be completely false when applied to a different population or even to the same population a couple of years later. When simplicity goes against empirical adequacy, it is simplicity which must be sacrificed, and not vice versa.

Classical genetics defines dominance via macroscopic observations unaided by any objective standards of measurement. Roughly speaking, a mutation in the coding sequence rendering an enzyme dysfunctional is equivalent to a complete truncation of the promoter region resulting in a total loss of enzyme synthesis: since both situations result in approximately the same phenotypic symptoms, classical genetics hypothesises that they are caused by the same gene. It must be understood however that this identity holds only in as much as we make abstraction of a host of minute variables, such as the severity of the symptoms, the onset of the disease, secondary complications, difference in response to treatment, etc. A protein, even if dysfunctional, still alters the chemical environment of the cell by interfering with a number of other processes. A

mutated enzyme may bind its substrate without processing it, thus competing with functional version of the enzyme, as well as with other enzymes using the same substrate for other metabolic purposes. In contrast, if the mutation results in the enzyme not being produced at all, there is no impoverishment of the substrate. Thus, in this particular case, the phenotypic condition associated with a loss-of-function mutation in the coding sequence is more severe than the phenotypic condition associated with a truncation of the regulatory sequence. The two conditions are very similar, but not identical. Only the more detailed molecular analysis is able to account for this diversity of sub-phenotypes. Once again, this proves that the 'molecular details' are not superfluous, but add to the overall empirical adequacy of the genetic explanation.

### *13.4 The Molecular Definitions of the Term 'Gene'*

The most serious matter of concern remains however the bridging of the classical notion of 'allele'/'gene' to its molecular homologues. Such definitions are required for the formulation of 'bridge laws' or 'translation rules' allowing the substitution of the terms belonging to the reduced theory with terms from the reducing theory.

Falk observes that there is no well defined entity that plays the role of term 'gene' in molecular biology. The 'molecular gene' is

> *"neither discrete – there are overlapping genes, nor continuous – there are introns within genes, nor does it have a constant location – there are transposons, nor a clearcut function – there are pseudogenes, not even constant sequences – there are consensus sequences, nor definite borderlines – there are variable sequences both 'upstream' and 'downstream'."*
>
> (Falk, 1986 p. 169)

To this I would add that there is no well defined set of molecular mechanisms that plays the role of the classical notion of 'gene' either. In a recent paper, Gerstein provides a comprehen-

sive overview of the various historical definitions of the term 'gene', of the problems associated with these definitions, as well as an attempt to redefine the concept in order to account for newly discovered regulation and diversity-generation mechanisms associated with gene-expression (Gerstein, et al., 2007). The definitions of 'gene' discussed by Gerstein, as well as some of the problematic unaccounted for by these definitions are briefly summarised in the table below:

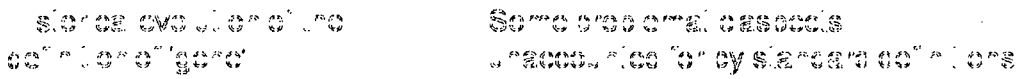


**Figure 24. The Concept of 'Gene'**

The bulk of the problems associated with the definition of the term 'gene' fall in three categories:

i. First, despite the initial successes of biochemistry, it turned out the auto- and hetero-catalytic properties of a gene cannot be defined solely as function of the chemical properties of the material associated with a certain genetic locus. These properties of genes reduce to the DNA sequences associated with certain chromosomal loci only in the extended the context of a general replication & transcription-translation biochemical machinery. In this sense, Ruse (1971) rightly points out that molecular biology takes a gene to be a "*functional*" (i.e., biologically functional) stretch of DNA rather than just the chemical structure and composition of that stretch of DNA. The 1940s definition, whereby a gene is a 'blueprint' for a protein takes into account the fact that genes must be

expressed, usually as proteins, in order for a phenotype to become manifest; by the same token, this definition disqualifies non-coding DNA sequences, such as pseudo-genes and structural DNA.

ii. Second, the general machinery responsible for replication and gene expression merely explains how the auto- and hetero-catalytic properties can be realised in terms of molecular mechanisms. The precise knowledge of heredity as it manifests itself in a living organism requires to take into account the associated regulatory mechanisms modulating the activity of the general replication & gene expression machinery in response to a predefined 'genetic program' or in response to environmental cues. This category includes problems related to the various levels of regulation (cell-cycle regulation, chromatin structure, transcriptional and translational regulation, inducible promoters, etc.). The 1960s definition is still the most widely used definition, for the very simple reason that the most potent and most widely studied mechanism of regulation is that of transcriptional regulation. The 1970s-1980s definition handles a fairly common mechanism of translational regulation.

iii. And third, the heterocatalytic activity of several genes maps (localises) onto overlapping chromosomal loci. This includes problems related to mRNA and protein splicing & trans-splicing. Gerstein's newly proposed definition aims to take into account the fact differential patterns of expression of the same genetic locus/DNA sequence can yield several RNA or protein end products sometimes serving distinct biological functions, and therefore are responsible for distinct phenotypes.

i) and ii) point to an inability of biochemistry to directly derive the auto- and hetero-catalytic properties of alleles/genes from the chemical structure of the material found at a certain locus.

iii) points out a further complication, namely the impossibility to associate an allele with a unique genetic locus.

### *13.5 Gene Expression*

Classical genetics draws a net distinction between genes (or alleles in Mendelian genetics) and their chromosomal locus (Morgan, 1935; Wain, et al., 2002). A locus is the place on a chromosome where the allele responsible for a given phenotype is to be found (Sturtevant, 1913; Painter, 1934); physically, this locus consists of a specific DNA sequence (Brachet, 1933; Watson, et al., 1953). The classical approach remains open to the idea that genes are localised, rather than identical to, a certain chromosomal locus (Wimsatt, 2006). The gene itself is 'that which ultimately leads to the manifestation of a given phenotype'; in slightly more modern terms, the gene is 'that which has autocatalytic and heterocatalytic properties' (Muller, 1951). In contrast, biochemistry and molecular biology aim to define the term 'gene' – which is a primitive, undefined term in classical genetics – as a function of what one finds at a the locus associated with a certain gene, that is, as a function of the properties of certain stretch of DNA (Schaffner, 1969; Darden, 1991; Waters, 1990; 1994).

Now, as Kitcher (1982; 1984) remarks, the difficulty of the molecular project stems from the fact that the properties of a gene are not identical to and do not reduce in any obvious way to the properties of the DNA sequence found at the locus associated with that gene. At any rate, it seems impossible to derive the auto- and hetero-catalytic properties of genes from the biochemical properties of the DNA found at a certain chromosomal locus (Rosenberg, 1978; Kitcher, 1984; Hull, 1972).

The initial ambition of biochemistry was to show that the chemistry of the living material alone suffices to explain and determine biological function. Supporting this initial ambition, it can be argued that, up to some point – that is, by making abstraction the fact that replication turned out to be a tightly regulated process – the 'autocatalytic' property of alleles is highly dependent on the peculiarities of the chemical structure of DNA; this realisation constitutes one of the most spectacular successes of biochemistry. Schaffner's argument that a gene reduces to what one finds at the specific chromosomal locus associated with that gene, namely a stretch of DNA, relies almost exclusively on the notion that the 'autocatalytic' property of alleles is granted by the chemical structure of the genetic material (Schaffner, 1969).

Based on the fact that enzymes act as catalysts in biochemical reactions in virtue of their chemical structure (three-dimensional structure, affinity for substrates, etc.), Beadle and Tatum (1941) entertained the hope that the 'heterocatalytic' property of a gene can likewise be defined in strictly biochemical terms. Unfortunately, it became clear fairly soon that certain biochemical mechanisms must be present and functional in order for DNA to be 'converted' into proteins.

Typically, classical analysis establishes that one or more loci are associated to each gene/allele. It does not follow from here that genes reduce to the said loci. Even if a stretch of DNA 'encodes' the sequence of a protein, this doesn't guarantee the expression of that protein, and therefore may have nothing to do with phenotypes and their genetic inheritance. In the absence of a mechanism linking the DNA sequence to the manifestation of a phenotype, a stretch of DNA does not 'code' for anything. For instance, DNA floating in an aqueous solution inside a centrifuge tube is simply a chain of phosphate-liked deoxyribonucleotides. In the context of a living cell, there are pseudogenes lacking the promoter region, just as there is huge amount of

highly compacted structural chromosomal DNA which doesn't seem to code for anything despite the fact that it consists of sequences of deoxyribonucleotides.

In order to be attached to a phenotype, as required by the genetic analysis which led to their discovery in the first place, genes must be expressed. In the vast majority of cases, gene expression amounts to DNA being transcribed into mRNA, the translation of mRNA into proteins, themselves ultimately responsible for phenotypes. Investigation of gene expression at the level of transcription revealed that genes consist of regulatory sequences, such as promoters and enhancers, and coding sequences, which are transcribed into mRNA and eventually translated into polypeptides. Without a basic promoter to which the basic transcriptional apparatus, consisting of the RNA polymerase and transcriptional factors, can bind, a gene is never transcribed into mRNA and therefore never expressed. Thus, it cannot be the case that a gene is a DNA sequence, and not even a coding DNA sequence. At the molecular level, a gene must be composed of a promoter regulatory sequence and a (usually adjacent) coding sequence.

The figure below illustrates a typical case of transcriptional regulation. As we can see, without an intact promoter region, gene expression is lost and, as a consequence, the phenotype changes. This clearly indicates that a phenotype is dependent not only on gene sequence (the chemical composition of a particular chromosomal locus), but also on gene expression, that is, on of the details of the 'how' leading from a certain to genotype to the corresponding phenotype.

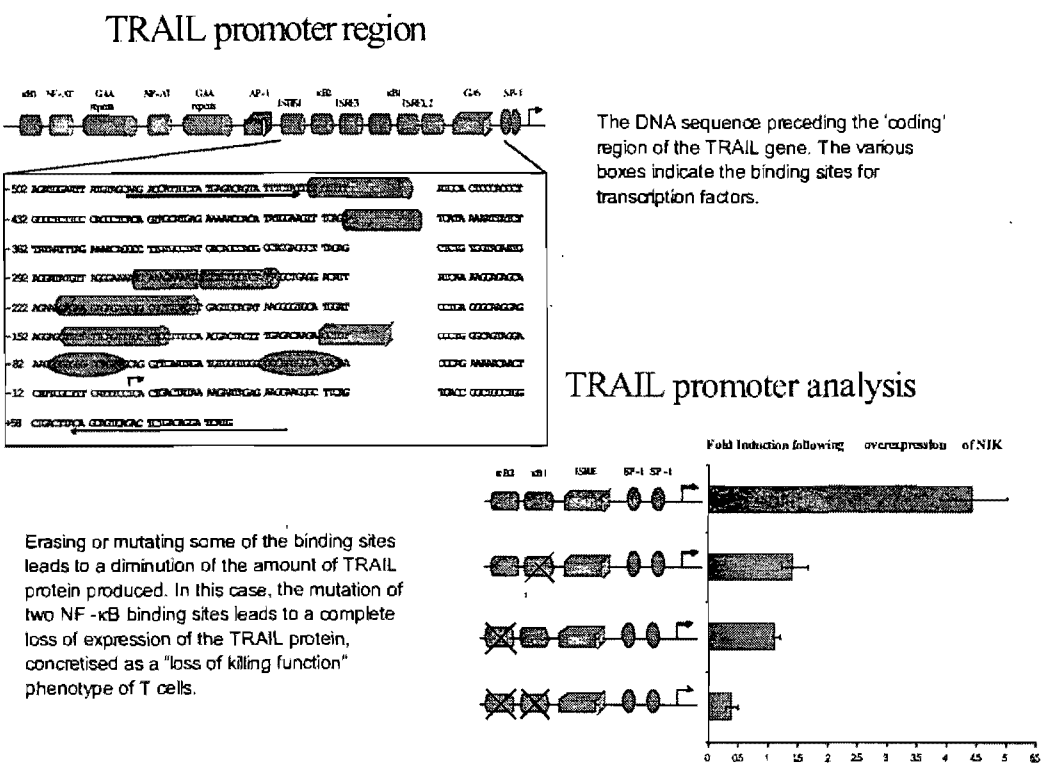



**Figure 25. Promoter and Coding Sequences within a Gene**

DNA sequences that do not have this structure are transmitted to the offspring following DNA replication (i.e., are autocatalytic, in Muller's terminology), but cannot be identified as genes via classical analysis since they don't have a phenotype associated to them (do not have a heterocatalytic property). If a coding sequence is treated as a gene, for example in order to translate some case of 'dominant-recessive' relationships in molecular terms, it is also immediately specified that it is a defective gene in which the minimal regulatory sequence necessary for the expression of the gene has been truncated (i.e., the gene has a dysfunctional regulatory sequence and there is a functional, promoter-endowed version of it in the population). In contrast, a pseudogene is a potentially 'coding' sequence (i.e., it would code for something if we would artificially add a promoter in front of it) devoid of regulatory sequence on both chromosomes and in all the members of a species.

The above considerations provide the necessary elements for understanding the standard, 1960s definition: a gene reduces to the DNA found at a certain chromosomal locus + transcription apparatus (promoter DNA sequence and biochemical machinery necessary for transcribing DNA into mRNA).

### *13.6 Gene Expression: A More Complete Definition*

This definition is satisfactory in regard to most intents and purposes, but remains incomplete. Ultimately, the whole 'DNA unpacking → DNA transcription → mRNA translation → protein localisation & post-translational processing' sequence of events must be reflected in a complete definition of a gene. Note that some of these steps – most notably transcription, translation and localisation – depend on specific sequences and therefore are said to be 'encoded' in the gene, yet none of these sequences means something in the absence of a biochemical apparatus recognising and processing them. Other processes, such as differential translation frames or alternative splicing seem to depend more on the stability of nucleic acid molecules and their interactions with proteins rather than specific sequences 'encoded' originally by the gene.

For example, a more accurate translation of talk about the TRAIL gene, understood in the classical sense of the 'ability-to-induce-apoptosis allele located at a certain chromosomal locus', in molecular terms is presented in the figure below. The data provided illustrates the correlation between the levels of expression of TRAIL mRNA, TRAIL protein inside the cell, TRAIL protein on cell surface and, finally, TRAIL protein function, which is to induce apoptosis (programmed cell death) of activated T cells.

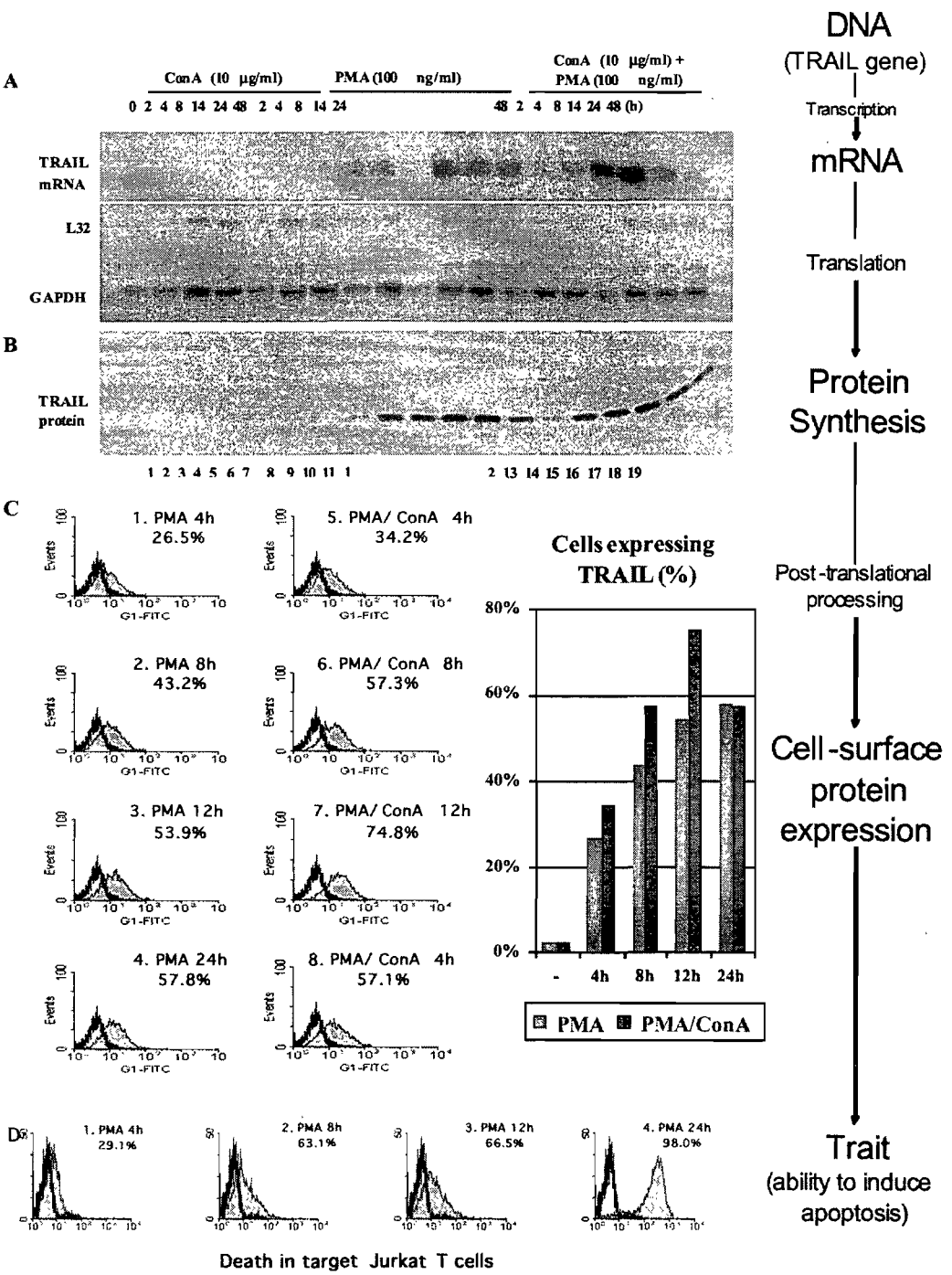



Figure 26. Causal Chains Linking Genotype and Phenotype

Talk about 'genes' and 'alleles' can be successfully replaced by talk about transcription, translation, etc.; depending on the gene, the molecular mechanism behind its heterocatalytic activity may be more or less complex. Note also that the first kind of talk, proper to Mendelian and clas-

sical genetics, is theoretical and serves only explanatory purposes. In contrast, the second kind of talk is backed up by experimental data; at the very least, by simply examining the figure above, the reader will be convinced that the molecular redefinition of Mendel's initial notion of 'allele' is associated with a whole network of observational outputs and empirical knowledge of causes and effects.

The above considerations show that the chemical structure of a stretch of DNA is insufficient to ground biological function; rather, a whole molecular mechanism, in this case DNA in conjunction with a whole replication & transcription/translation mechanism and their associated regulatory mechanisms are required in order for a biological function to obtain. This speaks against a straightforward reduction whereby "$gene_1 = DNA\ sequence_1$" (Schaffner, 1969 p. 342).

The bidirectionality of the equation proposed by Schaffner is highly problematic. It is indeed the case that, set aside a handful of exceptional cases, a gene is always associated with a certain chromosomal locus, itself consisting of a certain DNA sequence. However, we cannot know a priori if a random polymer of deoxyribonucleotides, or even better, a freshly sequenced strand of DNA extracted from a cell, is a gene, for neither is necessarily associated with a certain allele and its corresponding phenotype. In order to count as a putative gene, a DNA sequence must also present certain functional domains, most notably, a promoter responsible for recruiting the basic transcriptional machinery (the promoter region) and a sequence homology with some protein or RNA known to have some distinct biological function (the coding region).

### *13.7 Gene Expression Regulation*

Although a key element, the DNA structure of a gene cannot determine alone the phenotype of an organism. In order to count as a gene, a DNA sequence must at the very least be able

to interact with the basic transcriptional machinery of a cell. However, this is not the only requirement.

The 1960s definition applies very well to genes 'coding' for enzymes, structural proteins, cell-surface receptors, ligands/hormones/neurotransmitters, etc. It is less clear how it applies to what Waters (1990; 1994) calls "*regulatory genes*", that is, genes coding for transcription factors, as well as genes coding for the various protein and RNA components of the general transcription-translation, DNA compacting/uncompacting, post-translational modifications, general metabolism machinery.

As I mentioned on an earlier occasion, mutations in these genes have systemic effects which usually amount to a unique 'embryonic death' phenotype. Less radical phenotypes associated with mutations in genes 'coding' for transcriptional factors include 'severe immunodeficiency' syndromes and developmental aberrations. In these cases, the mechanism whereby the genotype determines the phenotype is not fundamentally novel, but just more complicated than the usual transcription-translation mechanism, as it usually involves a whole cascade of transcription-translation cycles.

One of the main differences between biochemistry and molecular biology hinges precisely on the notion of 'regulation'. While biochemistry uncovers the relationship between biological function and chemical structure, typically reducing the former to the later, molecular biology studies the mechanisms via which the biological function of a protein is modulated or turned off/on via an alteration of its chemical structure. In other words, biochemistry shows how certain chemical structures allow for certain processes/functions/properties, while molecular biology shows how living organisms literally function as biochemical automatons, that is, how the various chemical structures articulate together in order to form chemical mechanisms.

In order to understand how 'regulation' applies to gene expression, let us consider a typical example. The figure bellow provides a schematic representation of the chemical structure of the I□B polypeptide (the structure of the unfolded protein):

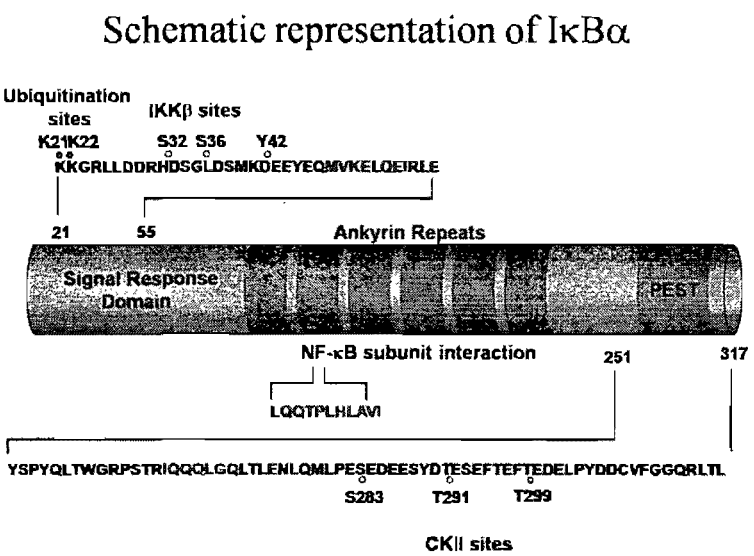



**Figure 27. The Chemical Structure of the I□B Polypeptide**

The structure, chemical composition, as well as the chemical properties of the various functional domains of the protein is determined by means of physical, chemical and biochemical techniques ranging from X-ray crystallography to immunoprecipitation assays. Note that although biochemistry reveals the chemical properties of the protein, most notably its ability to bind other proteins with high affinity, it tells us very little about its role in the actual functioning of a living cell.

In contrast, the figure bellow provides a schematic representation of a typical molecular mechanism:

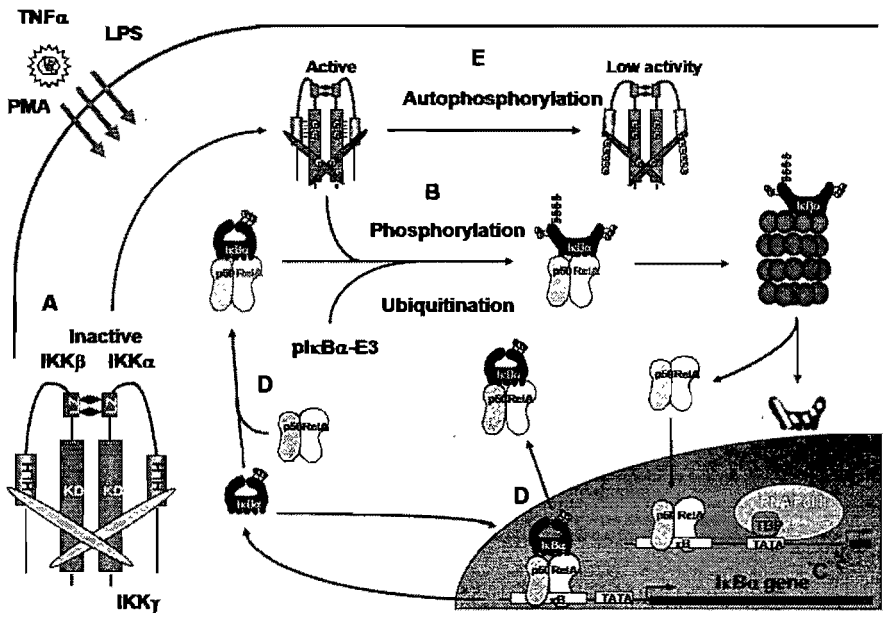



Figure 28. The Regulation of the I□B Gene

Briefly, in mammalian cells, unphosphorylated I□B binds a DNA-biding transcription factor known as NF-□B, forming a bulky complex that cannot reach the nucleus. When the cell is exposed to certain stimuli, I□B is phosphorylated. Phosphorylated I□B binds a complex of proteases and is degraded. As a result, NF-□B is freed, can translocate to the nucleus and bind specific DNA sequences. The binding of NF-□B results in an increased transcription of the target genes, in turn causing an increase in the expression of certain ligands, hormones, cell-surface receptors, antibodies and other key players responsible for a successful immune response. Interestingly enough, NF-□B also binds the promoter of the I□B gene, causing an increased production of I□B. The newly synthesised I□B binds NF-□B, trapping it back in the cytoplasm. Thus, following stimulation, the cell secrets a number of chemicals and expresses cell-surface receptors essential

for cell-cell communication, and then the system automatically turns itself off by means of a relatively simple molecular mechanism.

Note how in moving from biochemistry to molecular biology there is a loss of resolution in terms of chemical structure details, but something is gained in terms of a 'bigger-picture' understanding of biological function. Ideally, both a complete understanding of chemical structure and the wider understanding of molecular mechanisms is required in order to achieve complete knowledge. In practice, there is always a trade-off between the two.

Molecular biology shows that biological functions don't reduce to plain chemical structures, but rather to complex molecular mechanisms made possible by certain chemical structures.[32] In this particular example, defects of the NF-□B signalling and regulation pathway sometimes manifest as 'immunodeficiency syndrome' phenotypes. In order to understand the genetic basis of these syndromes it is not enough to understand that certain proteins directly responsible for immunity, such as antibodies, are produced following the expression of some genes. The geneticist must also have an overall understanding of the mechanism responsible for the regulation

[32] In this respect, I would like to reiterate a point made on an earlier occasion: all mechanisms are mechanisms of a certain kind. The Mendelian mechanism for heredity is a probabilistic one about mixing particles and then redistributing them. Molecular mechanisms are chemical; they function in virtue of chemical interactions. Physiological mechanisms are classical mechanisms relying on the notion of contact force and the laws 'dictating' its behaviour. And so on. It is impossible to talk of mechanisms simpliciter. A set of rules stating some fundamental ways of action must also be specified.

For example, it is often the case that a certain chemical structure, say, a phosphorylated protein, 'means' something for the overall functioning of the organism in the context of its newly acquired ability to bind DNA (or to bind more of the same protein in order to form a polymer playing a structural function; or again, to serve as a mediator in some signalling transducing pathway, etc.). The phosphorylated protein performs a certain function, namely biding DNA, while the unphosphorylated version of the same protein fails to perform the said function. Thus, in this case, phosphorylation is the molecular mechanism responsible for the 'turning on/off' of a certain biological function. Note however that the phosphorylated protein can bind DNA because the negative charge carried by the newly attached phosphate group causes a change in the three-dimensional structure of the protein, exposing a DNA binding domain hidden inside the unphosphorylated protein. Thus, the biological function of proteins depends on their chemical structure and, in this sense, the molecular biologist endorses the biochemist's credo that all biological function is consequence of chemical structure.

I make this point in order to emphasise the fact that just as the study of levers, pulleys, screws and other mechanical devices is a branch of classical mechanics, molecular biology is a branch of biochemistry. Many authors rightly emphasise the differences between biochemistry and molecular biology [see, for example, (Darden, 2006)]; I think however it is equally important to keep in mind that there cannot be molecular explanations in the absence of the theoretical and experimental framework of biochemistry.

of NF-□B. Mutations in the genes coding for NF-□B, I□B, the upstream signalling receptors and ligands, the kinases responsible for the phosphorylation of I□B, etc. are just as important as mutations in the genes directly responsible for immunity. It follows from here that, in many cases, talk about the heterocatalytic activity of certain genes cannot be successfully replaced by talk about transcription/translation alone, but requires talk more complex molecular mechanisms involving multiple cycles of transcription/translation, as well as signal transducing pathways, phosphorylation regulation mechanisms, etc.

### *13.8 Overlapping Chromosomal Loci*

In the context of classical analysis, research starts at the level of an inherited phenotype/biological function, hypothesises the existence of a gene responsible for the phenotype and attempts are made to localise the gene as a chromosomal locus/DNA sequence. Molecular analysis usually follows the reverse pattern. Chromosomal DNA is sequenced, putative genes are defined by matching probable coding DNA sequences homologous to known genes/RNA/protein sequences and possible promoter/enhancer regions, preliminary research shows that the putative genes can be expressed in artificial systems such as plasmid transfected-cells, and, eventually, the DNA sequences in question are shown to be essential to the determination of some biological function *in vivo*, say, via the generation of transgenic/knockout organisms.

On one hand, a gene, as defined in the context of classical genetics, is a chromosomal locus/DNA sequence associated with the inheritance and the expression of a certain phenotype. In contrast, many molecular biologists speak of genes while in fact they refer to putative genes whose precise biological function is still unknown or in the process of being elucidated (Fogle,

2001). The Human Genome Nomenclature Committee definition (the 1990s-2000s definition in Gerstein's list) attempts to alleviate the divergence between these two uses of the term 'gene' by redefining a gene as

> *"a DNA segment that contributes to phenotype/function. In the absence of a demonstrated function a gene may be characterized by sequence, transcription or homology."*
>
> (Wain, et al., 2002 p. 464)

The HGNC definition is the perfect complement to the molecular definitions preceding it. Biochemistry and molecular show that, in order to result in the expression of a certain phenotype, the stretch of DNA associated with a gene must have certain characteristics that allow it to interact with the transcriptional/translational machinery of a cell and its associated regulatory mechanisms. Hence, genes are defined as transcribed & translated stretches of DNA, or again as stretches of DNA participating in a larger and more complex mechanism such as the NF-□B pathways, etc. Conversely, if a stretch of DNA consists, say, of a promoter and a coding sequence showing some strong homology with known RNA/protein sequences, the stretch of DNA in question will most likely interact with the general transcription/translation machinery, thereby resulting in the synthesis of proteins which must one way or another affect the overall functioning of an organism. In other words, the presence of sequence motifs provides preliminary evidence that a stretch of DNA is most likely a gene responsible for a yet to be determined phenotype/function.

Gerstein and his collaborators point out that, although satisfactory for most intents and purposes, the HGNC definition fails to provide unique names for genes responsible for distinct phenotypes, but associated with overlapping chromosomal loci. For example, it was long known that the differential splicing of mRNA results in the synthesis of different proteins, which, in

some cases, may have distinct functions and therefore be responsible for radically different phenotypes. In order to account for this difficulty, Gerstein proposes that

> "*1. A gene is a genomic sequence (DNA or RNA) directly encoding functional product molecules, either RNA or protein.*
>
> *2. In the case that there are several functional products sharing overlapping regions, one takes the union of all overlapping genomic sequences coding for them.*
>
> *3. This union must be coherent – i.e., done separately for protein and RNA products – but does not require that all products necessarily share a common subsequence.*"
>
> (Gerstein, et al., 2007 pp. 676-677)

The figure bellow illustrates how this definition would work in practice:

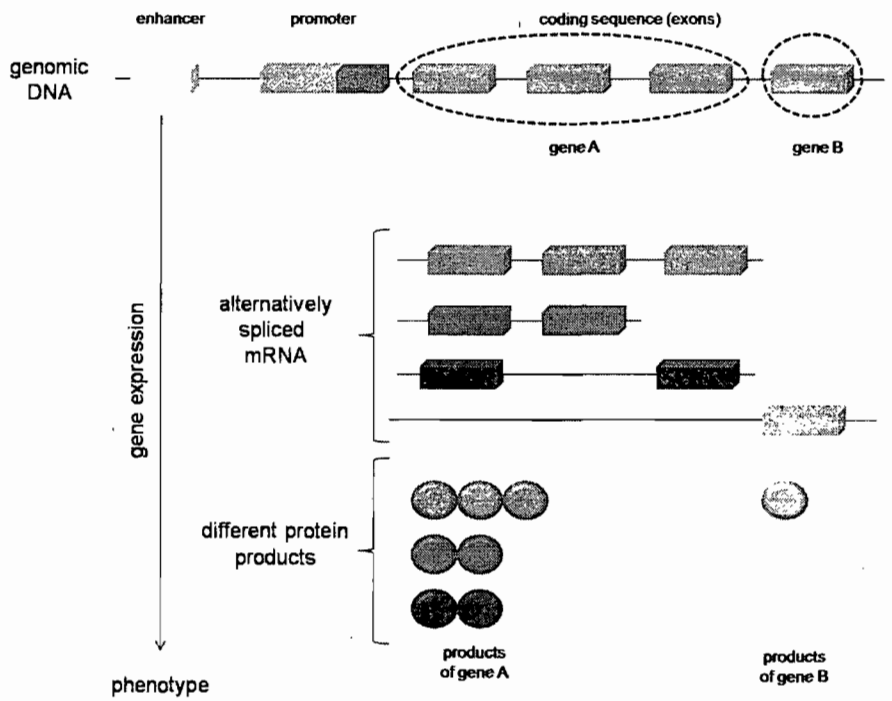



**Figure 29. Genes as 'subroutines in the genomic operating system'**

Note that the promoter region is no longer considered to be part of the gene. Just as several genes can share the same enhancer, Gerstein proposes that several genes can also share a common promoter.[33]

### *13.9 The Reduction of Classical Genetics to Molecular Biology is Complex, but not impossible*

The above survey indicates that there is no universally applicable allocution formulated in the language of molecular biology, that can replace the classical understanding of the term 'gene'. Based on this observation, Hull (1974), Rosenberg (1978) and Kitcher (1984) argue that the transition from classical genetics to molecular biology is marked by a discontinuity that renders intertheoretical reductionism impossible.

I think their conclusion is unjustified. Classical analysis localises genes at the level of the chromosomes. Biochemical analysis further establishes that specific DNA sequences are found at the chromosomal loci associated with genes and establishes that the chemical structure of DNA can account for certain properties of genes. Finally, molecular biology shows exactly how the auto- and hetero-catalytic properties of genes are actually realised via molecular mechanism in the living cell/organism. Since all these disciplines study the same spatio-temporal reality, we can safely conclude that there is at least a partial reference continuity from classical genetics to molecular biology.

---

[33] It should be clear though that even if the promoter regions are ignored in order to simplify the classification and nomenclature procedure, a gene does not reduce to a collection of exons (i.e., fragments of genomic DNA transcribed and eventually translated into proteins). As discussed in the previous sections, a stretch of DNA cannot function as a gene in the absence of a minimal transcription/translation machinery with which it must be able to interact via the promoter and enhancer sequences. Thus, whether the promoter/enhancer regions play a role in the classification and naming of genes, they are absolutely required for the molecular explanation of the heterocatalytic activity of genes, and therefore must be taken into account in a complete definition of the term 'gene'.

As a general rule, reference continuity is not a matter of dispute between reductionists and anti-reductionists; I mention it only to defuse possible charges of reference incommensurability. The first point of debate is whether classical genetics and molecular biology investigate different aspects of the same phenomena. It is question here of establishing whether there is a theoretical connection between the two disciplines in addition to mere reference continuity.

There are at least two good reasons to suspect that there is a tight theoretical relationship between classical genetics and molecular biology. First, the constant updating of the definition of 'gene' indicates that molecular biologists deploy relentless efforts to preserve the continuity between classical genetics and molecular biology. The presence of such efforts is inconsistent with the view that that molecular biology replaced, in a non-reductive sense, classical genetics. Second, I showed how the so-called 'molecular details' play a role in determining the genotype of an organism; this refutes the view that molecular biology and classical genetics have mutually exclusive explanatory scopes.

This brings us to the core of the debate. If there is a theoretical relationship between classical genetics and molecular biology, does this relationship amount to a reduction? Hull argues that reductionism is impossible because there is no simple, one-to-one relationship between classical and the molecular terms:

> "*Phenomena characterized by a single Mendelian predicate term can be reproduced by several types of molecular mechanisms [...] conversely, the same type of molecular mechanism can produce phenomena that must be characterized by different Mendelian predicate terms.*"
>
> "*To convert these many-many relations into the necessary one-one or many-one relations leading from molecular to Mendelian terms, Mendelian genetics must be modified extensively. Two problems then arise – the justification for terming these modifications 'corrections' and the transition from Mendelian to molecular genetics 'reduction' rather than 'replacement'.*"

(Hull, 1974 pp. 39, 43)

Part of the argument is that different kinds of molecular mechanisms explain the same classical notion of 'heterocatalytic' activity of genes (i.e., gene expression/regulation). This is true, but also irrelevant to the issue of reductionism. All that matters is that talk about the 'heterocatalytic' activity of genes can be successfully replaced by talk about molecular mechanisms without any loss, and often with a net gain of explanatory power. Even assuming that each individual instance of the term 'gene' in classical explanations corresponds to a different molecular mechanism, in as much as the sum total of classical explanations postulates a finite number of genes, to this finite number of genes corresponds an equally finite number of molecular mechanisms. The absence of a simple algorithm for converting talk about genes/alleles into talk about molecular mechanisms pertains to a technical difficulty, not to an impossibility of principle.

In fact, the reduction of classical genetics to molecular biology may not even be as complex as anti-reductionists like to believe. Most anti-reductionists fail to realise that despite their diversity, all molecular mechanisms consist of combinations between a handful of simpler sub-mechanisms such as transcription, translation, phosphorylation, etc. This suffices to defuse the misconceived idea that molecular biology introduces an indefinitely large number of explanatory strategies, each hypothesising a distinct kind of mechanism. In truth, molecular biology relies on a very limited number of basic explanatory mechanisms, which it combines as needed in order to provide a huge diversity of higher-level mechanisms (Darden, 2002; Darden, et al., 2002).

The second half of Hull's argument states that the same kind of molecular mechanisms must sometimes be translated by different classical terms. For instance, if we equate the molecular understanding of the term 'gene' to 'transcribed & translated piece of chromosomal DNA', then it can be argued that the same mechanism of transcription/translation results in the synthesis of proteins, yet mutation in one protein is 'dominant', while a similar mutation in the other pro-

tein is 'recessive'. Note however that the argument hinges on the implicit assumption that the molecular definition of 'gene' must be relatively rigid. Thus framed, half of Hull's argument goes head against the other half of his two-part argument. If we agree that the unique classical term 'gene' relates to a variety of molecular mechanisms, then the molecular biologist is free to further investigate and find out whether there is some minute molecular difference between the two cases that may explain the results of the classical analysis. In this case, the divergence may be eliminated by elucidating the mode of action of the protein in question (say, the first mutation may increase the affinity of some enzyme for its substrate, which would thus bind all the available substrate and render it unavailable for the 'wild-type' version of the enzyme; in contrast, the second mutation may render an enzyme totally incapable to bind its substrate, which can be processed by the non-mutated enzyme). Hull's argument works only if there are documented cases where mechanistic differences were not found despite the best efforts of molecular biologists to understand the preliminary data provided by classical analysis. To my best knowledge, there are not such cases.[34]

I conclude therefore that the argument from the absence of a one-to-one relationship between the classical and molecular understandings of the term 'gene' does not prove the impossibility of reduction. Instead, it simply points to a technical difficulty due to its inherent complexity.

[34] Outside the narrower issue of the reduction of classical genetics to molecular biology, it can be argued that certain biological functions and phenomena cannot be reduced to molecular explanations. According to Kitcher (1999), natural selection would be a good candidate. Following Lewontin and Levins (1985), Rosenberg (2007) also points out that genic reductionism is flawed. For example, he argues that even if the cause of haemophilia can be traced to a mutant gene, a complete explanation of the disease requires a wider understanding of genetics, biochemistry and human physiology. Note however that we are not concerned here with the general problem of reducing biological functions to molecular mechanisms or genetic explanations, but only with the particular problem of reducing the auto- and hetero-catalytic properties of genes to molecular mechanisms, as outlined in Muller's initial project.

### *13.10 A Simplified Approach to the Reduction of Classical Genetics to Molecular Biology: Differences in Genotype Typically Reduce to Differences in DNA Sequences*

It is interesting to note that despite the painfully elaborate definitions discussed above, most molecular biologists continue to do what everybody else is doing, namely call certain pieces of chromosomal DNA 'genes' and claim without the shadow of a hesitation that the DNA makeup of an organism is responsible for its inherited traits. The reason behind this widespread belief seems to escape most authors writing on the topic of reduction in genetics. In this section I will elaborate on a simplified and yet very powerful approach to reductionism initially outlined by Waters (1994).

Technically speaking, what is passed from parents to offspring is DNA coming from both parents and the egg from the female parent. The egg contains the molecular machinery required for gene expression and its associated regulatory mechanisms. If the latter were not passed along with the DNA to the next generation, there wouldn't be a next generation to talk about, less alone its phenotype. From the standpoint of the reductionist project, it is likewise clear that the 'genes/alleles', as defined in classical and Mendelian genetics, must have auto- and hetero-catalytic properties. As discussed in some detail, these properties cannot be accounted for by the chemical properties of DNA alone.

This said, it is also the case that, with very few exceptions, the basic biochemical machinery responsible for energy production, transcription, translation, essential metabolic pathways, etc. is highly conserved not only across the members of the same species, but often for all eukaryotes (Weber, 2005 pp. 162-164). Thus, in as much as the general context of a machinery responsible for 'reading' and 'expressing' the 'genetic code' is implicitly granted, it is correct to

say that what makes the difference between the inherited phenotypes of two individuals of the same species, or even different species is their DNA makeup.

This simplifies matters considerably. If we are interested in providing a molecular definition for the classical notion of 'gene', then we have to accept the fact that there is no unique molecular description that fulfils the explanatory role of the term 'gene'. However, if we are interested in comparing the genotypes of two or more organisms, then the difference between the 'genes' of various organisms can almost always be described in molecular terms as a difference between various DNA sequences.

Most commonly, the biochemical analysis of the material found at the chromosomal loci associated with a mutant and wild-type gene typically reveals only a difference in the DNA sequence. Furthermore, the experimental manipulation of the DNA sequences almost always establishes that the phenotype changes from wild-type to mutant or vice versa by making the appropriate changes in the DNA sequence. But if genes change as dictated by their associated DNA sequence, then not only the difference between the two genes localises at the level of point mutations, deletions or insertions in their DNA sequence, but it genuinely reduces to the said mutations, deletions or insertions.

There is therefore a clear sense in which classical genetics reduces to molecular biology. Note that this is not some lateral approach to genetics. This is the sense in which genetics is understood as a branch of medicine. What a clinical geneticist does most of the time is compare genotypes and their relation to phenotypes. In respect to clinical applications and counselling, this amounts to explanations of the type 'condition *X* is linked to a recessive mutation of gene *Y* to be found at chromosomal locus *Z*'. In other words, there is a difference between healthy subjects and patients affected by condition *X*, and this difference correlates with some difference of

the DNA sequence to be found at *Z*. The geneticist informs his patients about the risk of their progeny being affected by condition *X* given the presence of a mutant version of gene *Y* in one or both of the parents, or given the incidence of the condition in the family, etc. The molecular biologist hypothesises, and then verifies in cellular and animal models, that if the DNA sequence found at *Z* is restored back to the wild-type version of *Y*, condition *X* is eradicated. For those unfamiliar with the concept, this is the basic principle behind gene therapy.

### *13.11 The Convergence and Cumulativity of Scientific Knowledge from Classical Genetics to Present-Day Molecular Biology*

Despite the complexity issues related to the molecular definition of 'gene', there is an extensive overlap between classical genetics and molecular biology, where molecular biology succeeds in explaining most of what classical genetics explains while adding an increased degree of empirical adequacy, experimental control and overall degree of confirmation:

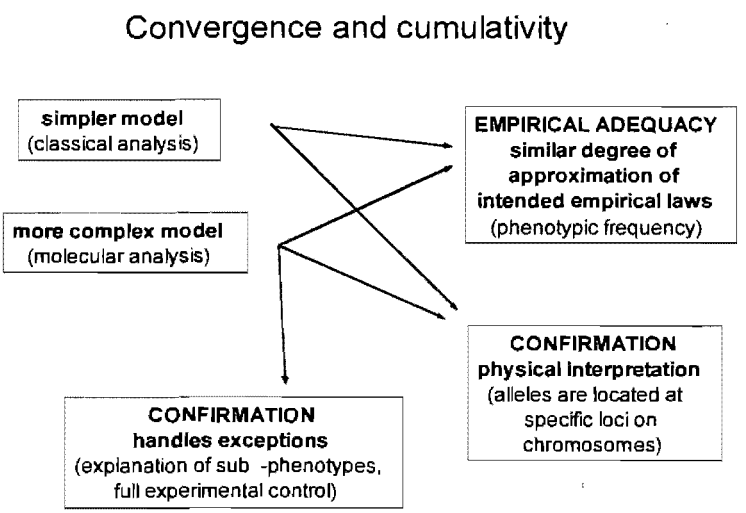



**Figure 30. The Convergence and Continuity of Genetic Theories**

The dual effect of convergence and cumulativity is demonstrated by the possibility to reduce classical analysis to molecular analysis and by the fact that differences of genotype typically re-

duce to differences of DNA makeup strongly suggests that classical genetics reduces to molecular biology.

## CHAPTER 14

# THE CONTRIBUTION OF EXPERIMENTAL DATA TO THE EMPIRICAL ADEQUACY OF SCIENTIFIC EXPLANATIONS

### *14.1 The Realism – Antirealism Spectrum*

Strictly speaking, the assumption that terms like 'allele' and 'segregation' refer does not contribute significantly to Mendel's initial explanation. Nevertheless, this assumption shaped the subsequent development of the theory and channelled the investigation on issues that are meaningless from an instrumentalist point of view. Thus, even if a theory in particular may allow for an instrumentalist interpretation, the subsequent development of the theory may deny it.

If instrumentalism is not a serious issue in this particular field of investigation, things are quite different when it comes to epistemology. The first thing worth noting is that epistemology allows for many degrees of confirmation, as well as truth approximation, in addition to plain

truth. As a consequence, epistemological realism and antirealism range towards the two extremes of a relatively wide spectrum of intermediate positions:

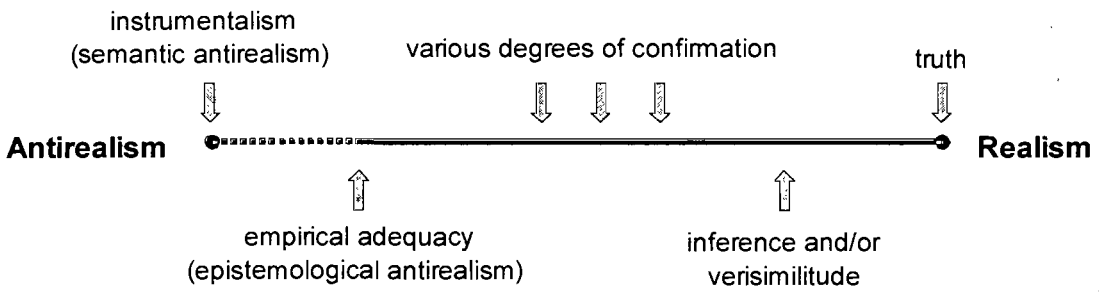



Figure 31. The Realism – Antirealism Spectrum

On the above spectrum, realists tend to defend at least the possibility of inference, that is, the possibility to conclude that a model is true because some elements it postulates are confirmed in experience. At the other end of the spectrum, antirealists tend to defend at least some form of epistemological antirealism.

This leaves us with the obvious question: "Is currently available confirmation enough to conclude, predict or reasonably expect realism?" Part of the answer must have something to do with the fact that, throughout the historical development of a theory, model or explanation there is a constant shift from empirical adequacy towards higher degrees of confirmation. Explanatory stories that don't shift towards the realist end of the spectrum are not necessarily abandoned, yet they invariably retain the lower status of possible routes of investigation in comparison with the higher status of acceptance associated with explanatory stories that receive some amount of confirmation (Lakatos, 1970; Laymon, 1984). The other part of the answer must have something to do with the possibility of inferring truth from partial confirmation. The traditional *HD/DN* approaches (Popper, 1959; 1965; Hempel, 1945; Hempel, et al., 1965), as well as abductive accounts allow for such inferences (Maxwell, 1962; Smart, 1963; Psillos, 1999), while similarity-based accounts allow a model to be truthful about a target phenomenon only in as much as indi-

vidual elements of the model resemble individual aspects of the target phenomenon (Hesse, 1966; van Fraassen, 1980; Giere, 1988; 2004).

### *14.2 Constructive Empiricism*

Van Fraassen argues that a distinction must be made between strictly empirical data and the further conceptual structure unifying empirical data into a larger theoretical structure. For example, in the case of celestial mechanics, we are told that observed or apparent motion amounts to "*relational structures defined by measuring relative distances, time intervals, and angles of separation*". It is only to these observables that we have direct epistemic access. In contrast, for Newton "*bodies are located in absolute space, in which they have real or absolute motions*". Absolute space and absolute motion are unobservables postulated by Newtonian mechanics. Nevertheless, within a Newtonian model of celestial motion "*we can define structures that are meant to be exact reflections of those appearances, and are* [...] *identifiable as differences between true motions*" (1980 p. 45). It is in virtue of this identity or isomorphism between actual appearances and structures within the model that Newton's mechanical theory can claim empirical adequacy.

The constructive empiricist approach can be applied to other disciplines as well, in particular to genetics and biology. As Johannsen (Johannsen, 1909; Roll-Hansen, 1989; Rheinberger, 2000) seems to have realised with a surprising clarity, Mendel's genetic explanation consists of two distinct stories: a smaller, empirical story concerning the frequency of phenotypes, and a bigger, explanatory one about the segregating alleles. The bigger story contains the smaller story; conversely, the smaller story is – to use van Fraassen's terminology (van Fraassen, 1980; 1989) – 'embedded' in the bigger story. One is part of the other. Unfortunately, if we know the

smaller story to be empirically true, we cannot say the same about the bigger story. The bigger story doesn't contradict the smaller story and in this sense it saves the empirical content of the latter, yet its own contribution towards explanation remains purely theoretical.

The diagram below summarises the main idea behind constructive empiricism:

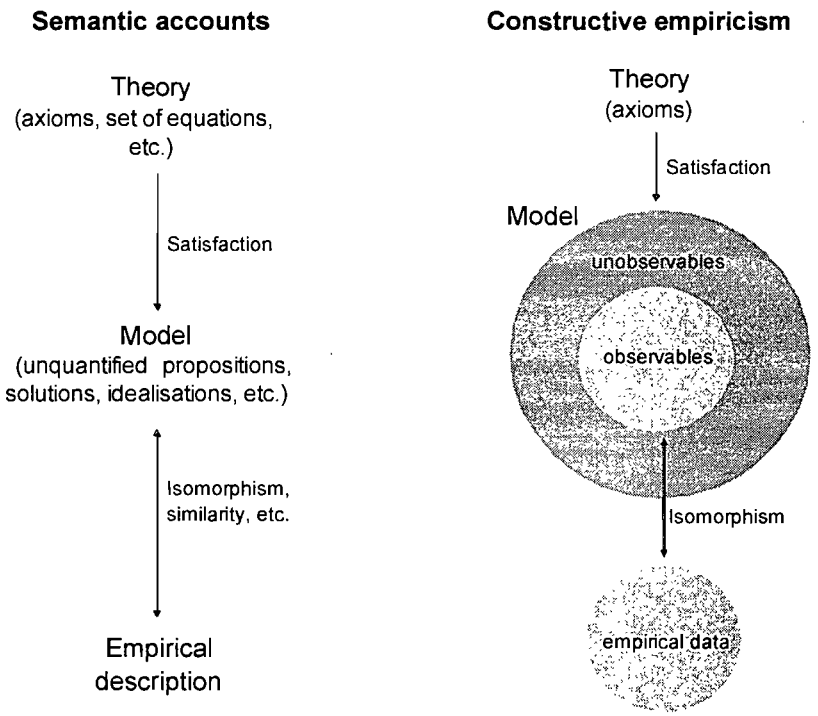



**Figure 32. Constructive Empiricism**

Given the antecedent provided by such examples, van Fraassen pushes the argument a step further and concludes that the aim of science is to provide empirically adequate explanatory stories:

> "*Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate.*"
>
> (1980 p. 12)

In his more recent writings, van Fraassen adopts a less ambitious formulation, according to which science could be just as intelligible and successful without further assuming that it suc-

ceeds in ascertaining truth about its theories, models and explanations. Under this weaker formulation, constructive empiricism claims to portray an equally viable alternative to science rather than science as it is actually practiced (van Fraassen, 1994; van Fraassen, et al., 1997).

### *14.3 Direct Observability*

The epistemological antirealism associated with constructive empiricism stems, in part, from the premise of a rigid distinction between 'observable' and 'theoretical' terms. Van Fraassen proposes that an entity is observable in fact or in principle if and only if observers can or could perceive it by unaided senses (van Fraassen, 1980 p. 16). In the case of genetics, this means that, irrespective of the experimental control and technological advances in terms of diagnosis and therapy brought about by molecular analysis, the genes and other entities associated with molecular mechanisms are bound to remain unobservable because, in all probability, there will never be nano-scientists observing directly these microscopic entities.

This conception of positivist descent was and still is heavily criticised (Hempel, 1965; Maxwell, 1962; Hacking, 1984; Churchland, 1985; Salmon, 1985; Menuge, 1995; Psillos, 1999). Here are some common matters of concern:

First, it has been noted that van Fraassen's criterion of observability presupposes that new evidence can never raise the probability of t-assertions (Psillos, 1999 pp. 187-188). This presupposition is untenable, especially if applied to models and explanations. An explanation, such as Mendel's genetic explanation of inheritance, is constructed in such as way as to entail an empirically adequate phenotypic distribution of the offspring. In the context of a Bayesian account of confirmation, the initial empirical adequacy of the explanation ensures that the explanation has non-zero prior probability. Further evidence, such as Griffith's discovery of the 'transforming

principle' raises the probability of Mendel's initial hypothesis that material 'elements' are transmitted form parents to offspring and are responsible for the phenotype of the latter (Avery, et al., 1944). It seems therefore reasonable to assume that hypotheses containing unobservables are confirmed the same way as hypotheses formulated only in terms of observables.

Van Fraassen (1985 p. 253) claims that "*experience can give us information only about what is both observable and actual*". But, as Rosen (1994) observes, 'observability' itself is a modal property. Since van Fraassen (1980 pp. 59-60) professes agnosticism about modal/dispositional facts, which he distinguishes from direct observations, it is not clear how it is possible to ascertain observability in the first place. Even if 'observability' is operationalised by correlating it with the presence of certain observable properties, there is no stable vocabulary of strictly observable terms. 'Red' is directly observable in reference to cars, but not in reference to blood cells. This raises questions as to what kind of things and properties are directly observable (Maxwell, 1962). Also, the notion that instrument-based observation is weaker than direct observation because we must further assume that the instrument of observation is reliable does not stand, for the same requirement applies to direct sense observation: the observer must not be blind, drugged, etc. (Menuge, 1995). Thus, there is a sense in which even direct observational data is conditional to a certain kind of experiment.

Finally, van Fraassen's criterion for distinguishing what is observable in principle remains quite vague. Our conception of what is possible in principle depends on our past experience and our current understanding of the world; to use Hempel's terminology, it hinges on the notion of 'nomological possibility' (Hempel, 1965; Hempel, et al., 1965). For example, it is not all that clear why there cannot be nano-observers, nor is it in any way clear how current theories in physics and astronomy allow for an observer travel to other end of the Solar System in order

to directly observe the satellites of Saturn or to the other end of the universe in order to directly observe a quasar (Psillos, 1999 p. 190).

### *14.4 Observational vs. Experimental Data*

The above objections to constructive empiricism are certainly damaging, yet they remain insufficient. Van Fraassen has a 'secret weapon' so to speak, a far-reaching argument that must be carefully considered, as it targets a key piece of scientific reasoning.

Van Fraassen backs up his agnostic position by making a sharp distinction between direct observation and experimental data. The argument is that experiments provide a fundamentally different kind of information about the world, not to be put on the same level as strict observational data. In relation to Newtonian mechanics, van Fraassen argues that

> "[i]*n the context of that theory, and arguably in all of classical physics, all measurements are reducible to series of measurements of time and position. Hence let us designate as basic observables all qualities which are functions of time and position alone. These include velocity and acceleration, relative distances and angles of separation – all quantities used, for example, in reporting the data astronomy provides for the celestial mechanics. They do not include mass, force, momentum, kinetic energy.*"

On the background of this general distinction between observables and unobservables, van Fraassen further exemplifies the status of inertial mass:

> "*if we postulate with Newton that every body has a mass, then mass is not definable in terms of the basic observables (not even if we add force). For, consider, as simplest example, a (model of mechanics in which a) given particle has constant velocity throughout its existence. We deduce, within the theory, that the total force on it equals zero throughout. But every value for its mass is compatible with this information.*"

In order to determine the mass of a particle, a force must be applied on the particle; only then its mass is determined. This kind of concerns suggests the following conclusion:

> "*the core behind them* [definitions of mass] *is that mass is experimentally accessible, that is, there are situations in which the data about basic observables, plus hypotheses about forces and Newton's laws, allow us to calculate the mass. We have here a counterfactual: if two bodies were brought near a third body in turn, they* would *exhibit different accelerations. But, as the example shows, there are models of mechanics – that is, worlds allowed as possible by this theory – in which a complete specification of the basic observables does not allow to determine the values of all the other quantities. The same observable phenomena equally fit more than one distinct model of the theory. (Remember that empirical adequacy concerns actually phenomena: what happens, and not, what would happen under different circumstances).*"
>
> (1980 pp. 59-60)

### *14.5 Modal & Dispositional Properties*

Stripped from its counterfactual formulation, van Fraassen's argument states that experimental knowledge is not about the studied objects or phenomena, but about their behaviour in 'contrived' or 'artificial' circumstances. This indicates that experimental data is not about properties of objects or phenomena, but about properties of larger experimental contexts in which these objects or phenomena are considered. In turn, this distinction further suggests that experimental data may not contribute to the total empirical adequacy of a model aiming to explain a naturally-occurring phenomenon.

Van Fraassen's insight is quite powerful. Experimental properties of an object or phenomenon can be said to be dispositional in nature. For example, in light of van Fraassen's analysis, it can be concluded that mass is a dispositional property of physical bodies that manifests

itself only when a force is applied onto the bodies in question. According to most analyses of dispositionality, dispositions are hypothetical properties making an implicit reference to a counterfactual state of affairs (Prior, 1982). Hence, in as much as talk about the experimental properties of an object considered outside the experimental context responsible for rendering these properties manifest is bound to refer to counterfactual states of affairs, it seems that van Fraassen is justified to conclude that experimental data does not contribute to the empirical adequacy of explanations concerning the actual state of the object.

Another point of interest to the issue pertains to the fact that many scientific models and explanations point to the 'possible' rather than the 'actual' mechanisms or structures underlying the empirical manifestation of the phenomenon under study. Many successful and fairly well confirmed explanations show how lab-produced phenomena relate to certain underlying mechanisms and structures; since the lab conditions are realised in naturally-occurring phenomena, scientists hypothesise that it is physically possible, and even probable, that the same mechanisms and structures underling the lab-produced phenomena are responsible for determining the naturally-occurring phenomena. Note however that even though scientists may know for sure that certain lab-produced phenomena are determined by specific underlying mechanisms and structures, they are not always sure whether these mechanisms and structures actually determine similar naturally-occurring phenomena. Depending on the domain of investigation, it is possible that the same phenomenon can be produced via several redundant mechanisms; this is especially true in biological sciences. Thus, a certain dose of agnosticism seems to be justified when successful explanations of lab-induced phenomena are extrapolated to naturally-occurring phenomena.

### *14.6 The Empirical Status of Experimental Data*

In the case of classical genetics, van Fraassen's objection would be that there are no chromosomes outside an experimental setup involving microscopy. There is a distinction to be made between directly observable patterns of phenotypic frequency and the further physical correlation of the alleles/genes with chromosomal loci, which is deemed to depend counterfactually on microscopy experiments. Thus, the empirical adequacy of classical genetics is guaranteed by direct observations while the further confirmation, in this case amounting to a suitable physical interpretation of some theoretical entities, relies on experimental data having a lesser empirical status.

The objection may seem sound, but if we look up closer there is something unmistakably absurd about the whole approach. If we take van Fraassen's objection seriously, we would have to conclude that Sutton redefines the theoretical term 'allele' in function of the theoretical term 'chromosome'. This is an extremely odd conclusion! It would further follow that Morgan provides empirical evidence for the well founded of the redefinition of a theoretical term in function of another theoretical term. But how can it be possible to provide empirical evidence for a theoretical definition? More so, why would one need to conduct an empirical investigation if we are concerned only with theoretical matters from the very beginning? There is something incongruous about arguing that chromosomes are theoretical entities because they are not observable by the unaided senses. Quite evidently, even if chromosomes are not observable by van Fraassen's standards, they do not stand on the same level as purely explanatory devices such as talk about alleles, nor do they play the same role in the overall economy of the explanation. At the very least, we should consider a three-way distinction between observables, experimental data and purely theoretical terms introduced for explanatory purposes.

Van Fraassen is entirely justified in arguing that we should not confuse properties of the observational outcome of an experiment, or properties of an experimental setup in general, with properties of the objects or phenomena subjected to experimentation. Nevertheless, it can hardly be said that the latter are actual, while the former are counterfactual. Rather, what is under scrutiny is an inference about whether the objects or phenomena in question behave the same way in different experimental setups. If this is the case, then researchers typically conclude that the properties belong to the objects/phenomena rather than to the experimental context or the interaction with the experimental context in which they are studied. If not, the properties in question remain just as empirical and actual, only they are not said to belong to the object/phenomenon under investigation (i.e., that they would have been there in the absence of a larger experimental setup).

This kind of inferences is undoubtedly very important, yet it is not something researchers must establish at all costs, especially in the early stages of a scientific inquiry. At the stage of the development of the genetic theories with which we are concerned in this particular example, it simply doesn't matter whether chromosomes exist outside microscopic observation or not. What matters is that a difference in phenotype (say, sex, mental retardation or cancer) is consistently associated with a difference in the observational output of microscopy (XX vs. XY karyotype, trisomy or chromosomal deletion) (Morgan, et al., 1915; Painter, 1934). Morgan did not know and did not claim to know what genes and chromosomes are 'in and by themselves' (Morgan, 1935). For him and his contemporaries, chromosomes were what one observes via a microscope after harvesting, staining and observing cells in a certain way. In the absence of substantial hypotheses about the nature, makeup and biological function of the chromosomes, there is nothing

theoretical about observing chromosomes and correlating these observations with certain phenotypes.

Also, given the right experimental setup, the correlation between certain phenotypes and chromosomal loci constitutes actual empirical data. In these specific setups, it is not necessary to translate this empirical correlation into talk about genes having an intrinsic, dispositional property of being associated with a chromosomal locus. It follows from here that experimental data is not (necessarily) theoretical, as it sometimes overlaps with actual empirical data.

### *14.7 The Continuity between Observational and Experimental Data*

Molecular analysis further correlates phenotypes not only with a more precise chromosomal mapping, but also with a host of other experimental data, such as the results of artificial mutations experiments, transgenic/knockout experiments, production of RNA, proteins, etc. By analogy with the mass example, van Fraassen might want to argue that this additional knowledge is counterfactual in the sense that it describes the objects of study as they behave in artificial experimental setups and does not reflect a knowledge of the world as it is by itself, that is, in absence of the experimenter's intrusion.

Intuitively, there seems to be a discontinuity between the direct, passive observation of phenotypic distributions from one generation to the next, classical analysis, whereby naturally occurring phenotypes are correlated with chromosomal loci via microscopy observations, and molecular analysis, characterised by an aggressive intervention whereby new mutants are artificially created. There is however a partial continuity as well. For example, although the mutants produced by the experimenter do not exist in the original populations studied by the classical geneticist, they

1) could exist in populations not currently under investigation
2) can always occur in virtue of 'spontaneous mutation' (replication and transcription errors, radiation, exposure to carcinogens, etc.)
3) if released in the population under study, they contribute to its overall genetic pool and cause a change in the phenotypic frequency of the studied population.

There is a variable overlap between 'naturally-occurring situations' and 'experimental setups' that doesn't fit very well van Fraassen's rigid actual-counterfactual distinction between experimental and passive observation data.

To use a popular, but quite adequate analogy, one of the most common experiments, the so-called 'knock-out' experiment, whereby genes are literally deleted from the genome of an organism, is similar to an attempt to identify the function of the various parts of a TV set by removing them one by one and assessing the impact of their absence on the overall functioning of the TV. Since electronics stores don't usually sell TVs with missing parts, van Fraassen would say that this or that component of the TV has a function only in the context of a counterfactual 'knock-out' kind of experiment. Nevertheless, TV sets occasionally break down and, on these occasions, by comparing functional and non-functional TVs, it is possible to actually observe the function of this or that part of the TV set. What is actual and what is counterfactual depends on what phenomenon we are considering. If we limit ourselves to the study of a brand new TV during one hour, then it might be impossible to observe the function of its various pieces. However, if we study the TV during its entire 'lifetime', the time it works, the time it breaks and the time after it is repaired, or even better, the 'lifetime' of the many TVs to be found out there, then we can observe the function of its parts quite easily. Seen from this angle, active experimentation in molecular biology, or any other field of research, simply accelerates the process of knowledge

gathering rather than providing a radically distinct kind of knowledge. Instead of waiting for TVs to break, something that invariably happens sooner or later, scientists break them first.[35]

Typically, the actual – counterfactual distinction arises only when empirical correlations (which may or may not amount to universal and/or necessary laws of nature) are generalised and then translated into talk about entities and properties in order to make predictions about specific

---

[35] A very similar comment applies to mass. If it is possible to determine the mass of a moving particle only by means of a counterfactual experiment, then it must be equally true that in the case of a particle actually deflected by a series of instantaneous or sustained collisions with various bodies, mass is actually observed. By considering a series of naturally occurring collisions, we find out that the particle may collide with perceptually indistinguishable bodies, yet the resulting change in motion varies considerably. Alternatively, the particle may collide with perceptually dissimilar bodies, yet the resulting change in motion is identical. To simplify the scenario, let us assume that the colliding bodies have all identical motion, say, they all enter in frontal collision with the particle and always at the same speed. Although it would require a significant amount of time, it is possible to gather all the required observations from naturally occurring phenomena alone. The results of the observations can be summarised as follows: of those bodies that undergo identical change in motion and cause an identical change in motion of the test particle following the collision, we say that they have the same mass; of those bodies that undergo and cause a different change in motion, we say that they have a different mass. Mass, which was defined so far as the empirically established equality or inequality of motion following collision, doesn't correlate with any visually detectable attributes of the objects and therefore cannot be seen no matter how hard we look at any object in particular, yet it remains an observable feature of a series of collisions. If we look at the overall event consisting of a series of collisions, we literally see the equality or inequality of mass; we didn't quantify it yet, but we definitively see it as clearly and distinctly as we see motion.

The above approach is a variation of Mach's definition of mass:

> "*If* [...] *mechanical experiences clearly and indubitably point to the existence of in bodies of a special and distinct property determinative of accelerations, nothing stands in the way of our arbitrarily establishing the following definitions: All those bodies are bodies are bodies of equal mass, which, mutually acting on each other, produce in each other equal and opposite accelerations. We have, in this, simply designated, or named, an actual relation of things.* [...] *The ratio of the masses is the negative inverse ratio of the counter-accelerations.* [...] *In our concept of mass no theory is involved; 'quantity of matter' is wholly unnecessary in it; all it contains is the exact establishment, designation and denomination of a fact.*"(1893 pp. 266-267)

In respect to his definition of mass, Mach remarks that

1) it is purely empirical, and
2) it is inseparable from a measure of mass, for it is impossible to define mass without saying at least if the masses of two bodies are equal or unequal (this second remark eliminates the satellite problem raised by van Fraassen: in as much as mass is observable, it is impossible to define mass without determining at the same time a quantitative or semi-quantitative measure of mass).

Given this definition – which, as Mach points out, is the mere naming of an empirical fact – the argument establishes that some particles, namely those actually colliding, actually have a mass. Pretending that collisions do not occur, and therefore their mass is forever hidden to observation, is a counterfactual statement. Pretending that the universe consists of a single, uniformly moving particle is a highly counterfactual statement referring to a very distant possible world. Van Fraassen focuses our attention on the fact that during the overall event of an experiment consisting in looking at a particle in motion, we do not observe mass. However, during the overall event of a second observation experiment, consisting of the initial motion as well as the collisions with other bodies, mass is observed and describes the event. In both cases, all we do is observe. As van Fraassen suggests, mass in not an observable property intrinsic to individual objects, but a property of a larger system consisting of colliding objects. However, it does not follow from here that the property is counterfactual or artificially created: in as much as the universe is actually made of colliding objects, there is nothing counterfactual or 'unnatural' about mass.

objects or phenomena. In this simple example, for any single TV at time *t*, we infer based on wide-range of correlations covering many TVs observed during long periods of time and/or in particular experimental setups that, although the TV is actually functioning perfectly, if some piece *x* were to removed, the TV would start displaying a specific malfunction. The original correlations pertain to actual empirical knowledge, and only the further inferences about individual units at specific moments in time refer to counterfactual states of affair.

### *14.8 Experimental Data: Actual or Counterfactual? A More Problematic Example*

This said, there are more problematic cases. Sooner or later, predictions must be made. Not only these predictions are dependent on inductive inferences, but quite often they also involve extrapolations to allegedly similar phenomena. As a study case, I will present an example drawn from my own research.

As part of my graduate curriculum, I worked in a molecular oncology lab on the regulation of T-cell activity. It was known at the time that, following direct stimulation, potent antigens, exposure to inflammatory agents, cell-to-cell contact, etc. primary and immortalised T-cells are activated *in vitro* and *in vivo* for a brief period of time, during which they multiply and produce a variety of chemicals leading to a strong immune response. After this period of activation, T-cells die via apoptosis (programmed cell death, to be contrasted with necrosis, or damage-induced cell death) and, as a consequence, the immune response shuts down. Crucial to the activation of the T-cells is a dimeric transcription factor known as NF-□B (nuclear factor □B; in its active form, it is found in the nucleus where it binds specific DNA sequences known as □B binding sites). Various stimulators work via distinct transduction pathways, all leading to the ac-

tivation of NF-□B. In the resting cell, NF-□B is held away in the cytoplasm by a small protein known as I□B (inhibitor of NF-□B). However, when the cell is stimulated, a chain of protein-protein interactions leads to the degradation of I□B, such that NF-□B is now free to translocate to the nucleus, bind the □B sequences in the promoter regions of target genes and drastically enhance their transcription, process ultimately yielding in the production of the various chemicals secreted by activated T-cells.

Paradoxically, it was shown that NF-□B is responsible for the transcription of several genes encoding for proteins responsible for inhibiting apoptosis. Presumably, this up-regulation of inhibitors of apoptosis is responsible for the early activation and proliferation of T-cells. It remained however unclear how T-cells die following activation. Eventually, it was shown that a family of ligand proteins – TNF□ (tumour necrosis factor □, originally characterised as responsible for killing cancer cells) being the first discovered member of the family – are secreted or expressed on the surface of activated cells. These ligands bind receptor proteins found on neighbouring cells and cause them to die through apoptosis. Thus, once a pool of T-cells is activated, they end up by producing apoptosis-inducing ligands which eventually kill neighbouring T-cells.

The figure below summarises these findings [for a more complete review of TNF, FAS and TRAIL-induced apoptosis, see (Baetu, et al., 2002)]:

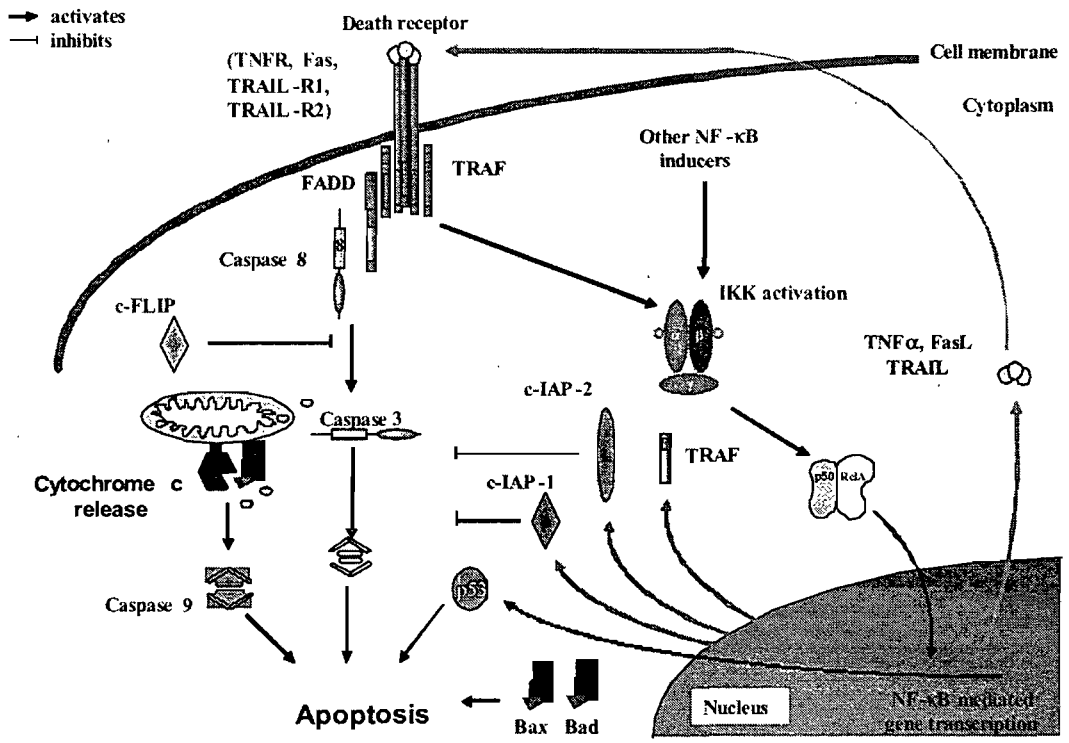



Figure 33. The NF-□B Signalling Pathways and Apoptosis

In a series of preliminary experiments, a former member of the lab noticed that the activation of NF-□B correlates with an increase of TRAIL (TNF-related apoptosis-inducing ligand) mRNA. As the name suggests, TRAIL is highly similar to TNF and was shown to be an extremely potent apoptosis inducing protein. This proved to be an excellent opportunity for research, as it was becoming more and more clear that TNF and other previously investigated members of this family of ligands, although involved in apoptosis, are not the most important players in as much as T-cell death is concerned. My task was to show that NF-□B activation is responsible for TRAIL induced apoptosis.[36]

In order to do so, I divided my project in two main steps. All the cells of an individual contain the same genetic material. Furthermore, genes essential for survival display little varia-

[36] For a complete listing and discussion of the results discussed in the following pages, see (Baetu, et al., 2001).

tion from one member of the species to another. Thus, it is reasonable to assume that any human cell would contain the TRAIL gene and the associated □B sequences in its promoter. Nevertheless, in order to be transcribed, the DNA must be un-compacted in the relevant cells, in this case, T-cells. In order to show that TRAIL is expressed following an activation of NF-□B, I had to reproduce the preliminary results in T-cells. The easiest way is to use an immortalised T-cell line, such as cells extracted from lymphoma patients. Primary cells, that is, normal, non-immortalised cells extracted from a healthy individual, die soon after their removal from the body; although not impossible, it is very hard to keep them alive in artificial media. In contrast, immortalised cells contain specific mutations leading to a loss of regulation of the cell cycle and an uncontrollable proliferation; in particular, they can proliferate *in vitro*, making them ideal targets for study.

Figure 26 shows the results in Jurkat T-cells (an immortalised T-cell line). Following stimulation with a variety of chemical agents known to induce NF-□B (in my study, PMA and ConA), the amounts of TRAIL mRNA, cytoplasimc TRAIL protein, surface TRAIL as well as the levels of apoptosis increases. As expected, this increase correlates with an increase in nuclear NF-□B, DNA-bound NF-□B, as well as the disappearance of I□B. The results were very encouraging, but insufficient. The experiment relied on very potent stimulators known to activate NF-□B, as well as many other transcription factors. Therefore, although highly probable, it was impossible to conclude that NF-□B is specifically required for TRAIL expression.

Without entering the technical details, I will simply state that it is possible to create artificial cell lines in which the expression of I□B can be drastically enhanced on command, thus leading to a complete or almost complete inactivation of NF-□B. The technology relies on the engi-

neering of a special bacterial plasmid (short, circular piece of DNA) responsible for Doxycycline (Dox; an antibiotic) resistance. Bacteria replicate and exchange this plasmid among them, thus gaining resistance to the antibiotic. The regulation of the Dox resistance gene ('encoding' for an enzyme responsible for digesting Dox) is very peculiar, as it relies on a transcription factor capable of biding the promoter of the gene only in the presence of Dox. By replacing the original Dox-resistance coding sequence with that of a constitutively active mutant of I□B (in my experiments, a truncated version of the protein that cannot be degraded following T-cell activation) and integrating the plasmid in T-cells, it is possible to create artificial cells in which NF-□B activity can be inhibited and restored at will by adding and removing Dox from the cell growth medium (this technology is known as the 'rtTA system').

There is a risk associated with this procedure. The bacterial plasmid is eventually integrated in one of the chromosomes of the cell. The integration can disrupt chromosomal genes, thus creating mutants of various kinds. In addition, the integration is essentially random, meaning that in a pool of transfected T-cells (cells having absorbed the artificial plasmid) integration may occur in a different place for each cell. The first thing to check is whether TRAIL regulation is affected in the newly created T-cell lines. In the figure below, the left columns (labelled rtTA-Neo; these are control cells transfected with an empty plasmid) indicate that mRNA and cell-surface TRAIL expression is increased following activation of artificially engineered cells the same way as in the original T-cell line. This indicates that the integration did not affect overall cell viability, the transcription/translation machinery, the TRAIL gene or any of the players involved in the NF-□B regulation pathway.

Now, if TRAIL requires NF-□B for its expression (the hypothesis under test), then by keeping NF-□B inactive a loss in TRAIL expression following stimulation of T-cells should be expected. This is indeed what happens, as shown in the right columns (labelled rtTA-2N□4 and referring to cells transfected with a truncated version of I□B):

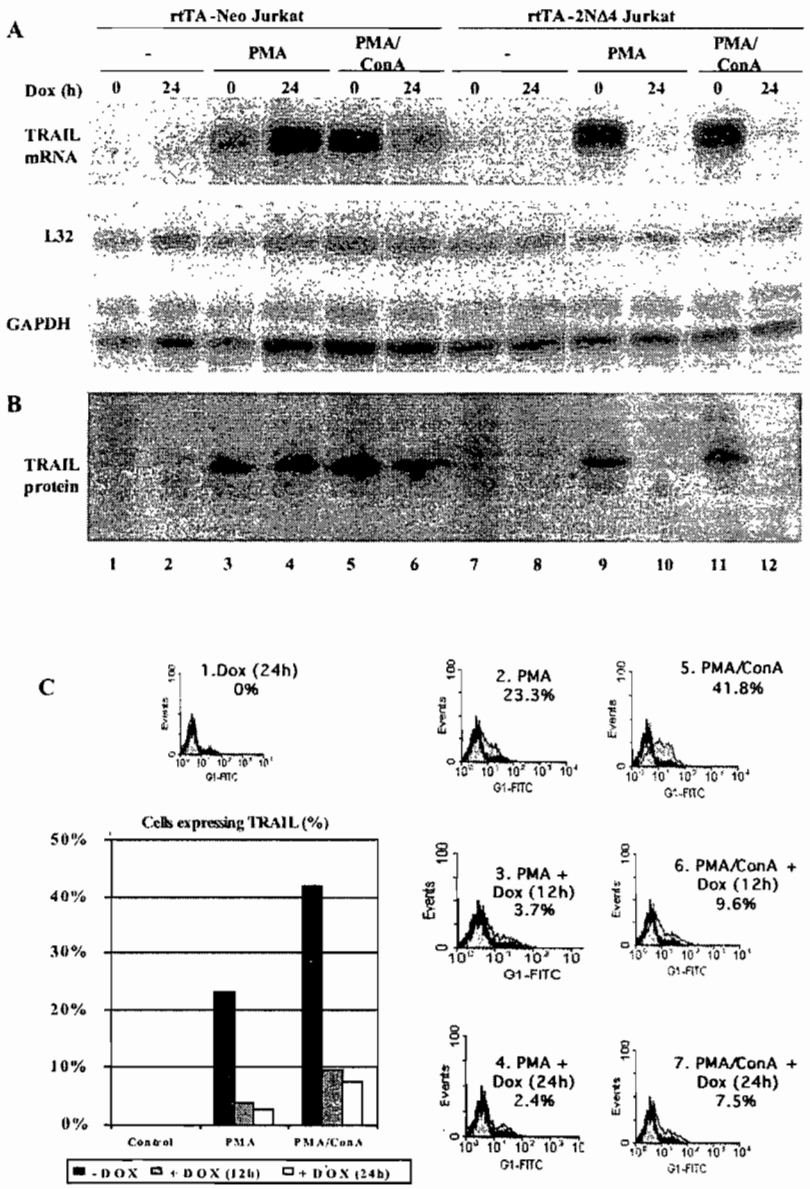



Figure 34. NF-□B Dependent Regulation of TRAIL

This shows that NF-□B is indeed necessary for the expression of TRAIL. Note however that in using a T-cell line instead of primary T-cells, I took a risk, namely the risk that the transcriptional regulation of TRAIL may be dysfunctional or different in immortalised cells than in primary cells. In order to show that TRAIL is, or at very least might be regulated by NF-□B in normal cells, the same results had to be reproduced in primary T-cells. In addition, all the work done so far was exclusively *in vitro*. In order to ensure at least a partial continuity with 'natural' *in vivo* conditions, it is required to switch to a different kind of stimulation, namely a cell-cell interaction. This posits a number of problems. First, to this date, it is impossible to use the rtTA technology on primary cells; they simply die too fast in order for the procedure to work. Second, cell-cell interactions are very difficult to reproduce *in vitro* (*in vivo*/clinical studies/trials on animal/human patients are simply out of question at this stage of the investigation). I had therefore to resort to a compromise. Instead of using the highly specific rtTA system targeting only the inhibition of NF-□B, I had to rely on less specific anti-inflammatory drugs (aspirin, Bay11, MG132) known to work mainly by knocking down NF-□B activity. Likewise, instead of performing cell-cell interaction assays, I relied on the fact that, in the case of T-cells, these interactions are known to be mediated principally via specific surface receptors known as CD3 and CD28. I used therefore antibodies known to bind these receptors with high affinity and cause a cellular response highly similar to that of natural ligands of these receptors or direct, cell-cell interaction. Once again, it was possible to show that TRAIL expression is greatly enhanced following stimulation of T-cells and that this stimulation can be inhibited by NF-□B inhibitory drugs:

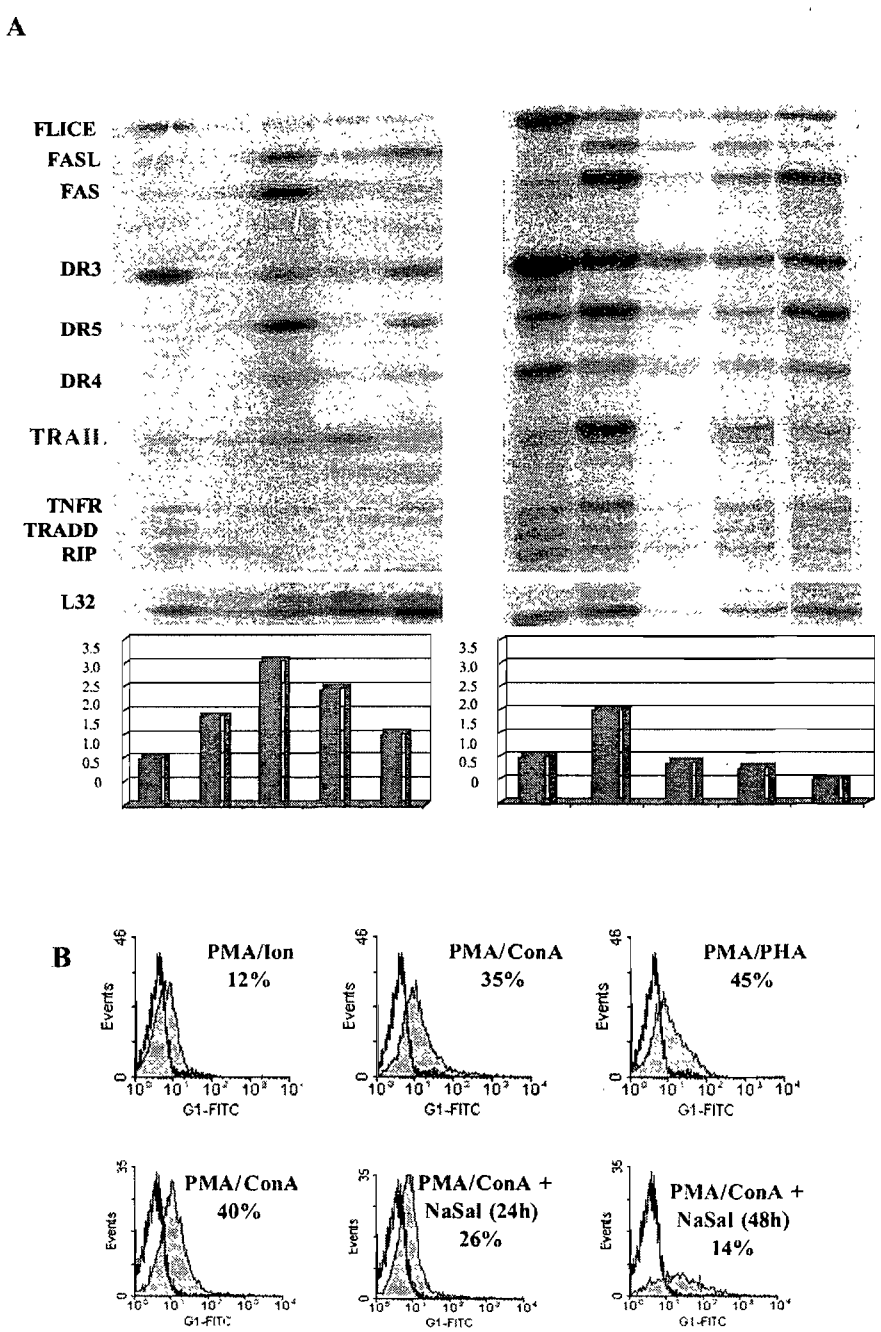



**Figure 35. NF-□B Dependent Regulation of TRAIL in Primary T-cells**

I hope the reader can see without any difficulty that my project consisted in reproducing the same results, using essentially the same techniques, in systems approaching more and more the actual, *in vivo* conditions of the human body. I also hope the reader can see that there is no

single huge jump from one context to another, but rather a whole series of extrapolations, each involving a modest inference about a slightly more complex, or sometimes just a less controllable, experimental setup. Some of these inferences can be justified theoretically, or at very least are coherent with the general theoretical framework of molecular biology (e.g., the notion that, with very few exceptions, all cells of an organism possess the same genetic material), while others rely on brute empirical correlations (the resemblance between primary and immortalised T-cells, between the effects of CD28 ligand stimulation and those of the anti-CD28 antibody stimulation, etc.).

### *14.9 Experimental Data is not used counterfactually and Similarity does not alleviate the Need for Confirmation*

My study does not show that T-cells actually die *in vivo* following activation due to an up-regulation of TRAIL. Rather, it shows that the up-regulation of TRAIL following the activation of NF-□B is responsible for their death *in vitro* and that it is possible that the same scenario happens *in vivo* (i.e., it is possible that the same conditions are met *in vivo*, thus leading to a sequence of events culminating with TRAIL-induced death of T-cells). This is the officially accepted interpretation of my results.

Since data obtained *in vitro* and/or on cell lines is not thought to contribute directly to the empirical adequacy of models about *in vivo* primary cells, the data obtained in the lab is not used counterfactually. Also, despite the similarity between various experimental setups (amounting, in some cases, to material models) and a target phenomenon, it does not follow, at least not from a scientific point of view, that what is true about the experimental setups is also true or approximately true about the target phenomenon. If that were the case, there would be no need for clini-

cal trials. Unfortunately, clinical trials are still needed, meaning that similarity does not replace the requirement for confirmation.

### *14.10 Models Consisting of a Chain of Extrapolative Hypotheses*

Now, if data obtained *in vitro* and/or on cell lines does not describe what actually happens *in vivo* and/or in primary cells despite an increasing degree of similarity, the reader may wonder what might be the value of studies on cell lines in respect to the understanding of the functioning of the human body. In my experience, it helps formulate further hypotheses and more complex models.

Work on cell lines tells researchers where to look first: they hypothesise that whatever happens in cell lines also happens in primary cells. In my case, a preliminary justification of my hypothesis relies, it is easy to guess, on the fact that immortalised T-cells resemble morphologically and functionally primary T-cells more than, say, epithelial cells or even B-cells resemble T-cells. As noted above, this kind of preliminary justification does not count as conclusive proof or evidence that the hypothesis is true, approximately true, probably or possibly true. Further empirical research must be conducted, else the project is bound to remain a mere hypothesis, maybe a very attractive one, but nevertheless just a hypothesis.[37]

The experimental control associated with cell line research is a lot tighter and more diverse; in addition, the confirmation of hypotheses is stronger and more complete. Not to mention that working on primary cells from the very beginning would have meant draining myself of half a litter of blood every three days until I figure out the right concentrations of activators and inhibitors of NF-□B, the best RNA extraction protocol, etc. From a practical point of view, work-

[37] Some epistemologists consider such hypotheses to have a higher probability of truth. In contrast, scientists usually understand 'epistemological probability' along the lines of a 'project worth pursuing'.

ing first with a suitable cell line was the only viable strategy. Once the research on cell lines finally gave results, I was able to formulate a reasonably acceptable hypothesis about what goes on in a living body and switch to the much more difficult and expensive work on primary T-cells.

Similarly, what happens *in vitro* serves to formulate a hypothesis about what happens *in vivo*. For example, it was hypothesised that during AIDS, HIV infected T-cells die because of an up-regulation of TRAIL. The elements for the above hypothesis were two-fold: it was shown, by my study and other studies as well, that NF-□B activation leads to an increased expression of TRAIL and a subsequent increase in apoptosis; the other piece of the puzzle came from a study of a colleague of mine who showed that HIV infection leads to an activation of NF-□B. Both studies were done *in vitro* with some partial emulation of *in vivo* conditions. The challenge was to show that the AIDS stage of the HIV infection correlates with a massive increase of TRAIL expression. This turned out to be true and several groups published a number of highly quoted papers on the subject.

Work on cell lines and/or *in vitro* is the equivalent of a mini-theory subsequently extended to similar phenomena occurring in primary cells and/or *in vivo* conditions. Due to the lack of a mathematical formalism, most models in molecular biology amount to something along the lines of the pre-mathematical 'rock spinning on a string' model for planetary motion initially suggested by Newton. It does not necessarily follow from here that these models are of an analogical kind. It is not question of determining to what extent two phenomena resemble each other in order to conclude that, if a high degree of similarity is established, what is true about one phenomenon must be approximately true about the other phenomenon as well [as exemplified in (Hesse, 1966; Giere, 1988; Giere, 2004)]. Rather, explanatory models derived from such mini-theories hypothesise that, just as the circular path of a rock spinning on a string is due to a ten-

sion force acting perpendicular to its motion, the circular planetary orbits are also due to gravitational forces acting perpendicular on motion; or again, that just as NF-□B is responsible for the regulation of TRAIL in T-cell lines, NF-□B is also responsible for the regulation of TRAIL in primary T-cells. These hypotheses need to be confirmed, at least partially, before researchers can claim to posses any kind of knowledge about the modelled phenomena.

The figure below summarises the overall modelling procedure associated with the example from TRAIL research:

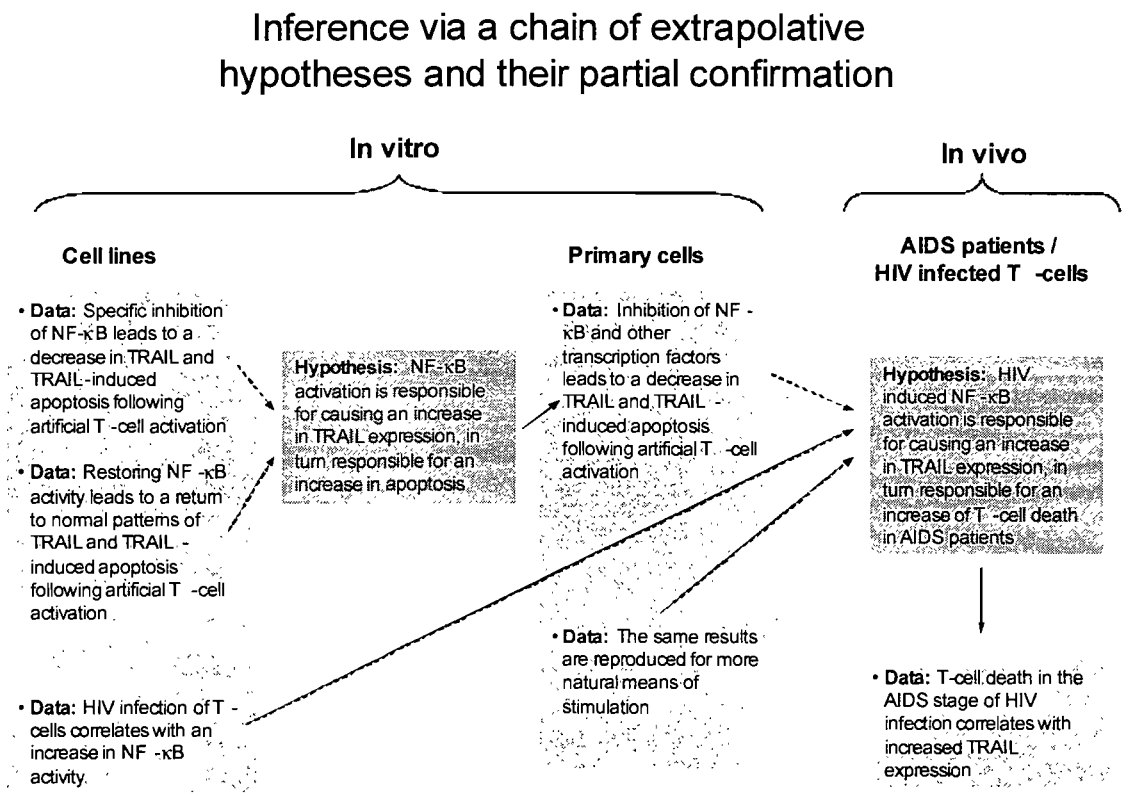



**Figure 36. The 'TRAIL-mediated' Model of T-cell Death during HIV Infection**

The final picture whereby HIV infection leads to NF-□B, in turn leading to TRAIL expression and cell death is still a hypothetical one. As shown in the figure above, the various pieces of the puzzle belong to different systems and experimental setups. Only by transgressing the differ-

ences between immortalised cells and primary cells, as well as the differences between *in vivo* and *in vitro* environments, can scientists pretend to have a complete, rounded up explanatory story. In this particular example – much more so than in van Fraassen's original mass example – the objection concerning the [counterfactual] attribution of properties proper to one system to another system is a most serious matter of concern.

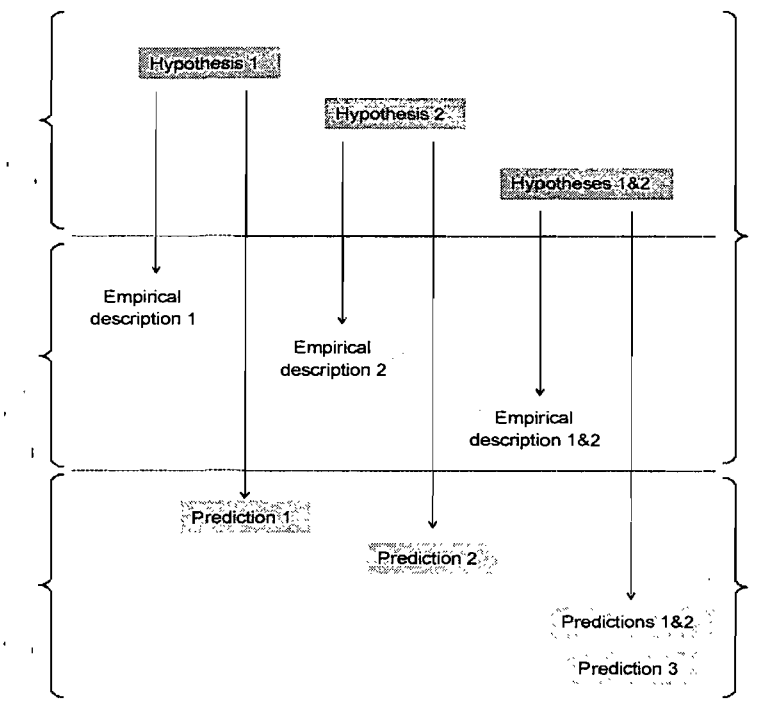
### *14.11 Holistic Confirmation of a Chain of Extrapolative Hypotheses*

Hopefully, there is a way out of the difficulty. Thus far, researchers have in their possession a reasonable hypothesis explaining the studied phenomenon: they have a theoretical model in which data obtained *in vitro* (a set of premises) entails massive T-cell death (the conclusion, which is description of the phenomenon under study).

Science is an ongoing process. Given the above model, researchers can design future treatments for AIDS and some autoimmune diseases based on the development of highly specific NF-□B inducing/blocking drugs, as well as various other means targeting directly or indirectly the production of surface TRAIL protein. The eventual success of such treatments will confirm the whole series of inferences whereby data from *in vitro* cell lines is extrapolated to living systems. Conversely, the failure of such treatments will show that somewhere the chain of extrapolations is broken (for instance, there might be an essential difference between stimulation *in vivo* and stimulation *in vitro*; or it may turn out that primary cells are, under some relevant aspects, very different from their immortalised cousins; etc.). If this is the case, then what is true about experimentally modified cells and of the lab models of a disease is not true of the cells and of the disease as they occur 'naturally'.

The above scenario illustrates the main idea behind the 'conjunction argument' (Friedman, 1983 pp. 244-247). The simplest and most straightforward confirmation strategy is to test each model on an individual basis. However, most of the time, the confirmation is only partial and the fact that some elements of the model are confirmed provides insufficient grounds for inferring that the model is true, most likely true or true to a high degree of approximation. Things change for the better when, instead of having to make a judgement about the truth of a single model, it becomes possible to assess the truth of a collection of models cross-referencing each other. The moment several partially confirmed models combine together in order to yield new confirmed predictions, the conjunction of the models receives a higher degree of confirmation than each individual model.[38]

The figure below summarises the concept of holistic confirmation:




[38] It has been argued that constructive empiricism is incompatible with the ability to derive novel predictions from the conjunction of several models (Psillos, 1999 pp. 205-211). The incompatibility targets van Fraassen's isomorphism approach to empirical adequacy [illustrated in (1989 pp. 218-220)]: the fact that two theories have models isomorphic with two distinct sets of empirical data does not entail that there is a model of the conjunction of the two theories isomorphic with the conjunction of the two sets of empirical data. This indicates that constructive empiricism is not a perfectly equivalent alternative to the actual scientific practice.

Figure 37. Holistic Confirmation of a Conjunction of Models

In as much as two partially confirmed models combine together and entail new, eventually confirmed predictions, scientists can conclude that they are on the right track. For example, the conjunction of classical genetics and chemistry (Figure 22), or again the conjunction of the NF-□B dependence regulation of TRAIL model with the NF-□B dependence of HIV replication model (Figure 36) entail new predictions, derivable only from the conjunction of two theories or models. In more general terms, given the larger context of an extended research project whereby more and more complex models are built on the premises of simpler models, modest inferences from lab setups to naturally-occurring conditions can be hypothesised and subjected to confirmation.

### *14.12 The Conjunction Argument and Abductive Reasoning*

As a final point, I would like to make a link between holistic strategies of confirmation and abductive reasoning. It has been suggested that the only way to defeat the problem of the 'underdetermination of theories by empirical evidence' (Quine, 1975; Laudan, 1996) is to allow for the explanation itself to decide between equally adequate theories. The argument goes as follows: if a theory amounts to a body of theoretical statements $T$ from which a body of statements about observables $O$ is derived, then no matter how much we extend $O$ (to include, for example, further experimental data in addition to direct sense observations), there will always be, in principle, a second body of theoretical statements $T'$ which entails $O$; the only way to effectively adjudicate between $T$ and $T'$ is to assess the extra-empirical explanatory 'efficiency' of the two sets of theoretical statements (McMullin, 1987; Lipton, 1993). The conclusion here is that justifica-

tion cannot be exclusively empirical. Given this need to rely on extra-empirical justification, it has been further suggested that the practice of inferring is a reliable form of reasoning conducive to truth (Boyd, 1984; Psillos, 1999).

Unfortunately, the approach is problematic. In order to conclude that abductive reasoning is a reliable practice conducive to success, one must provide some previous successful instances of abductive reasoning. Presumably, these earlier successes are established in light of some other form of justification, in occurrence, empirical justification (van Fraassen, 1980; Fine, 1984). This suggests that reliability cannot be a primary mode of justification, comparable to or more important than empirical confirmation. To make things worse, reliability doesn't justify the results, but rather the ability of a given practice to yield the desired results. That the practice in question aims to truth doesn't change much to the situation. In order to associate a practice with truth, one must first recognise truth. It follows from here that truth the criteria in virtue of which we justify something as being true are independent of the reliability of the methods used to obtain truth. Given these difficulties, it appears that explanationist approaches can work only in as much as they make place for a priorism in matters of justification. Despite some sympathetic voices (Miller, 1987), a priorism in matters of justification is deemed fundamentally incompatible with the standards and aims of scientific knowledge (Boyd, 1990). For the time being, it is hoped that naturalised accounts relying on evolutionary-like reliability might provide a way out of this impasse (Psillos, 1999).

The 'conjunction argument' may provide the patch this approach needs. The confirmation of more and more complex models constructed on the premises of simpler models retrospectively justifies a multitude of modest inferences required in order to combine models issued from various experimental setups. Presumably, once justified, the same inferences can be used again,

this time with a higher degree of confidence, as it has been already established that they are conducive to confirmable predictions.

## *Conclusion*

As announced in the introduction, the primary purpose of this thesis is to defend and update the hypothetico-deductive account of the scientific practice. Although it is tacitly understood that when it comes to empirical confirmation scientists invariably follow the same general strategy, namely formulate hypotheses and test for specific observational consequences, contemporary philosophers of science raised a number of doubts and objections as to the logical validity and efficiency of the method. The resulting philosophical reactions to this critique turned out to be very diverse, ranging from a denial of the possibility of confirmation, anti-empiricism and anti-realism, to an exploration of new aspects of scientific reasoning neglected by positivist currents, such as explanation, modelling and thinking in terms of mechanisms.

A first set of difficulties, addressed in Part I of the thesis, pertains to the possibility of distinguishing between the theoretical component of scientific hypotheses and the experimental data which often plays a role in the formulation of hypotheses and against which specific consequences of hypotheses are tested. Needless to add, if there were no such thing as experimental knowledge, then the whole practice of formulating hypotheses wouldn't make any sense. Ac-

cordingly, my first concern was to defend the possibility of gaining experimental knowledge in the absence of higher-level theoretical interpretations. I think the discussed examples, some in fairly minute detail, amply demonstrate this possibility.

Beyond the mere illustration of the possibility of experimental knowledge, an effort was made to relate it to some leitmotifs of positivism and provide a fresh and more charitable understanding of some aspects of the logical positivist project. Set aside its immense relevance from a practical point of view, philosophically, the most interesting use of experimental knowledge is perhaps operationally-defined coreference, mainly in relation to its implications for the non-theoretical unity of scientific knowledge. At a more preliminary level, I also explore the impact of available experimental knowledge on the formulation of explanatory hypotheses and conclude that, in some cases, the former posits a constraint onto the latter. At this stage my analysis is incomplete, and a more extensive investigation of the historical development of molecular techniques is required in order to determine how exactly experimental knowledge shapes molecular explanations; nevertheless, my immediate goal was to show that scientific explanations must be coherent with a relatively large and quite complex web of experimental correlations which favours some explanations at the detriment of others. I think that, the latter, more modest goal was successfully completed.

A second set of difficulties, discussed in Part II, is tightly linked to the issue of confirmation holism. The core concern here is that a hypothesis is usually formulated within a much larger and more complex theoretical context to which a host of initial conditions, idealisations and auxiliary assumptions/hypotheses are added in order to allow for the derivation of predictions specific enough to allow for empirical confirmation. Although there is no general solution to the holist conundrum, I show by means of a relatively simple example that specific elements

of a model can be confirmed and that it is possible to provide falsification conditions for a given theoretical approach, both necessary for successful science. The goal here is to analyse a simple study case in order to derive partial solutions applicable to the more complex examples. Most notably, I argue that models inconsistent with the theory but supported by empirical data provide powerful falsification conditions for that particular theoretical approach.

On the same occasion I discuss newer approaches to confirmation, issued form the so-called 'semantic approach', holding that a model or explanatory story is compared as a whole against empirical reality. The idea is to circumvent the holistic conundrum altogether by removing the requirement of a piece-meal confirmation strategy advocated by logical positivists. The semantic approach is credited to Suppes, who claimed that scientific models are empirical instantiations of theoretical structures, and is at the origin of present-day structuralist accounts. Although instantial models are fairly common, I argue that they pertain primarily to the subsuming of empirical data under a suitable mathematical formulation. At any rate, the confirmation strategies used to justify Newton's gravitational model of planetary motion are not compatible with Suppes' suggestion and I couldn't find a suitable illustration in the field of molecular biology and associated sciences. As far as I could gather via my formation and personal experience as a scientist, it is virtually never the case that an explanatory story, even a fairly modest one, can be shown to share an overall structural similarity with a target empirical phenomenon; rather, once specific elements of the explanation are shown to share a structural similarity with specific elements of the target by means of a multitude of confirmation experiments, it is further inferred, usually as a mere hypothesis, that the model as a whole is structurally similar with the target phenomenon. In particular, even though typical molecular explanations rely on elaborate biochemical mechanisms, in the end it is not the whole mechanism that is subjected to confirmation,

but rather very particular elements which entail specific consequences in terms of direct observation and experimental manipulation. As a consequence, I chose to defend a deductive approach to the modelling practice in science according to which theoretical models, as they are typically used in the scientific practice, amount to derivations linking more general patterns of explanation to specific bits and pieces of experimental data. The upshot of this view is that, although it is absolutely true that most explanations in molecular biology are framed in terms of mechanisms, the mechanisms in question are reconstructed bit by bit, like a puzzle, via the testing of much more modest, very specific and often indirect empirical consequences (or what Suppes calls 'models of data'). This observation strengthens my initial claim that experimental knowledge plays a significant role in shaping the formulation of explanatory hypotheses. Instead of constructing a mechanism and then showing that the mechanism bears an overall structural similarity with some empirical phenomenon, the mechanism is constructed step by step, via a succession of revisions starting from a relatively simple initial scheme and towards a more and more complex picture, as dictated by new empirical findings.

Finally, in Part II of the thesis I discuss difficulties pertaining to the realism – antirealism debate in philosophy of science. The deductive approach to explanation and modelling defended in the second part of the thesis is compatible with instrumentalism. It seems however that the historical development of science favours a realist interpretation. The study case on the basis of which I reach this conclusion is the development of present day genetic theory. The argument rests on the observation that the initial hypothesis formulated by Mendel was interpreted realistically, namely as postulating the existence of physical entities possessing what Muller called 'auto-' and 'hetero-catalytic' properties; the subsequent development of the theory consisted es-

sentially in identifying these entities and elucidating the mechanisms and modes of action responsible for their properties.

Despite the claims of some authors, my investigation revealed that hypotheses issued from a realist interpretation of the initial genetic explanation were partially confirmed and, interestingly enough, the physical interpretation favoured via confirmation turned out to contribute back to the explanatory story by adding further constraints the explanation must satisfy in order to be empirically adequate. This suggests that earlier versions of the genetic theory reduce to present day molecular genetics. In this respect, I provide a very detailed and up to date analysis of the theoretical relationships between classical and molecular genetics.

In the final chapter of the thesis I discuss a problem that, just like the observational-theoretical distinction and the holist conundrum, was left in suspense for quite some time. Generalising the principle of incertitude in quantum mechanics, it has been argued that experimental data is counterfactual and therefore cannot contribute to the empirical adequacy of an explanation. This scientific practice under attack here is that of inferring something about the structure and nature of naturally-occurring phenomena from knowledge of the structure and nature of lab-induced phenomena. As van Fraassen observes, strictly speaking, such inferences are unjustified. Nevertheless, I show via actual example that such inferences can sometimes be empirically justified via the confirmation of consequences entailed by a conjunction of models. The argument here is that by viewing science as an ongoing project, constantly integrating idealisations, assumptions and inferences taken for granted into the formulation of more complex explanations, it is possible to frame the latter as hypotheses that stand or fall as the more complex models they make possible are confirmed or falsified.

In conclusion, I hope the present thesis achieved a double goal. On one side, by using detailed examples, I showed how the use of the hypothetico-deductive method integrates the larger context of the scientific practice. In this sense, I depict science as dependent onto the generation and eventual confirmation of fairly specific observational consequences, yet try not to fall into the trap of extreme empiricism or naive positivism. This is part of a descriptive side of my project. In parallel, I showed how deductive accounts compare with alternative philosophical accounts of the scientific practice and in what measure they are apt to answer a number of key philosophical concerns. This is part of a critical assessment aiming to establish in what measure the use of the hypothetico-deductive method supports claims to truth and knowledge.

## Bibliography


**Achinstein, P. 1964.** On the Meaning of Scientific Terms. *Journal of Philosophy.* 1964, Vol. 61, 17, pp. 497-509.

**Asimov, I. 1966.** *Understanding Physics.* New York : Walker, 1966.

**Avery, O., MacLeod, C. and McCarty, M. 1944.** Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 1944, Vol. 79, 2.

**Baetu, T. M. and Hiscott, J. 2002.** On the TRAIL to Apoptosis. *Cytokine & Growth Factors Reveiws.* 2002, Vol. 13, pp. 199-207.

**Baetu, T. M., et al. 2001.** Disruption of NF-kB Signalling Reveals a Novel Role for NF-kB in the Regulation of TNF-Related Apoptosis-Inducing Ligand Expression. *The Journal of Immunology.* 2001, Vol. 167, pp. 3164-3173.

**Barnes, E. 1992.** Explanatory Unification and the Problem of Asymmetry. *Philosophy of Science.* 1992, Vol. 59, pp. 558-571.

**Baskar, R. 2008.** *A Realist Theory of Science.* New York : Versa, 2008.

**Beadle, G. W. and Tatum, E. L. 1941.** Genetic Control of Biochemical Reactions in Neospora. *Proceedings of the National Academy of Science.* 1941, Vol. 27, pp. 499-506.

**Bell, J. and Machover, M. 1977.** *A Course in Mathematical Logic.* Amsterdam : North-Holland, 1977.

**Ben-Menahem, Y. 2000.** Idealization. [ed.] W. H. Newton-Smith. *A Companion to the Philosophy of Science.* Oxford : Blackwell, 2000.

**Benzer, S. 1955.** Fine structure of a genetic region in bacteriophage. *Proceedings of the National Academy of Science.* 1955, Vol. 41, pp. 344–354.

**Berk, A. and Sharp, P. 1978.** Structure of the Adenovirus 2 Early mRNAs . *Cell.* 1978, Vol. 14, pp. 695-711.

**Bloor, D. 1991.** *Knowledge and Social Imagery.* Chicago : Chicago Univ. Press, 1991.

**Boring, E. G. 1923.** Intelligence as the Tests Test it. *New Republic.* 1923, Vol. 35, pp. 35-37.

**Boyd, R. N. 1984.** The Current Status of Scientific Realism. [ed.] Jarrett Leplin. *Scientific Realism.* Berkeley : Univ. of California Press, 1984, pp. 41-43.

**Boyd, R. 1985.** Observations, Explanatory Power, and Simplicity. [ed.] P Achinstein and O. Hannaway. *Observation, Experiment, and Hypothesis In Modern Physical Science.* Cambridge : MIT Press, 1985.

**—. 1990.** Realism, Approximate Truth and Philosophical Method. [ed.] W. Savage. *Scientific Theories, Minnesota Studies in the Philosophy of Science.* Minneapolis : University of Minnesota Press, 1990, Vol. XIV.

**Brachet, J. 1933.** Recherches sur la synthese de l'acide thymonucleique pendant le developpement de l'oeuf d'Oursin. *Arch. de Biol.* 1933, Vol. 44, pp. 519-576.

**Braithwaite, R. B. 1953.** *Scientific Explanation.* Cambridge : Cambridge Univ. Press, 1953.

**Bridgman, P. 1927.** *The Logic of Modern Physics.* New York : Macmillan, 1927.

**Bromberger, S. 1966.** Why-Questions. [ed.] B. A. Brody. *Readings in the Philosophy of Science.* Englewood Cliffs : Prentice Hall, 1966.

**Bunge, M. 1973.** *Philosophy of Physics.* Dordrecht : Reidel, 1973.

**Campbell, N. 1920.** *Campbell, N.* Cambridge : Cambridge Univ. Press, 1920.

**Carlson, E. O. 1967.** *The Gene: A Critical History.* Philadelphia : W.B. Saunders, 1967.

**Carnap, R. 1939.** Foundations of Logic and Mathematics. *International Encyclopaedia of Unified Science.* Chicago : Univ. of Chicago Press, 1939, Vol. 1(3).

—. **1936.** Testability and Meaning. *Philosophy of Science.* 1936, 3, pp. 419-471.

—. **1928.** *The Logical Structure of the World.* Berkely : Univ. of California Press, 1928.

—. **1956.** The Methodological Character of Theoretical Concepts. [ed.] H. Feigl and M. Scriven. *The Foundations of Science and the Concepts of Psychology and Psychoanalysis.* Minneapolis : Univ. of Minnesota Press, 1956.

**Cartwright, N. 1983.** *How the Laws of Physics Lie.* Oxford : Oxford Univ. Press, 1983.

—. **1999.** *The Dappled World. A Study of the Boundaries of Science.* Cambridge : Cambridge University Press, 1999.

**Cartwright, N., Shomar, T. and Suárez, M. 1995.** The Tool-box of Science. [ed.] W. Herfel, et al. *Theories and Models in Scientific Process. (Poznan Studies in the Philosophy of Science and the Humanities 44.* Amsterdam : Rodopi, 1995.

**Castren, E. 2005.** Is Mood Chemistry? *Nat. Rev. Neurosci.* 2005, Vol. 6, 3, pp. 241-246.

**Chow, l., et al. 1977.** An Amazing Sequence Arrangement at the 5' Ends of Adenovirus 2 Messenger RNA. *Cell.* 1977, Vol. 12, pp. 1-18.

**Churchland, P. M. 1988.** *Matter and Consciousness.* Cambridge : MIT Press, 1988.

—. **1985.** The Ontological Status of Unobservables: In Praise of Superempirical Virtues. [ed.] P. M. Churchland and C. A. Hooker. *Images of Science.* Chicago : Chicago Univ. Press, 1985.

**Coffa, J. A. 1967.** Feyerabend on Explanation and Reduction. *Journal of Philosophy.* 1967, Vol. 64, 16, pp. 500-508.

**Crow, Ernest W. and Crow, James F. 2002.** 100 Years Ago: Walter Sutton and the Chromosome Theory of Heredity. *Genetics* . 2002, Vol. 160, pp. 1-4.

**Da Costa, N. and French, S. 2003.** *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning.* Oxford : Oxford University Press, 2003.

**Darden, L. and Craver, C. F. 2002.** Strategies in the Interfield Discovery of the Mechanism of Protein Synthesis . *Studies in History and Philosophy of Biological and Biomedical Sciences.* 2002, Vol. 33, pp. 1-28.

**Darden, L. 2006.** *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution.* Cambridge : Cambridge University Press, 2006.

—. **2002.** Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/Backward Chaining. *Philosophy of Science.* 2002, Vol. 69, pp. S354-365.

—. **1991.** *Theory Change in Science: Strategies from Mendelian Genetics.* New York : Oxford Univ. Press, 1991.

**Davidson, D. 1984.** *Inquiries into Truth and Interpretation.* Oxford : Oxford University Press, 1984.

**Devitt, M. and Sterelny, K. 1999.** *Language and Reality.* 2nd edition. Cambridge : MIT Press, 1999.

**Devitt, M. 1981.** *Designation.* New York : Columbia University Press, 1981.

—. **1990.** Meanings just ain't in the head. *Meaning and Method: Essays in Honor of Hilary Putnam.* Cambridge : Cambridge University Press, 1990, pp. 79-104.

—. **2004.** The Case for Referential Descriptions. [ed.] Bezuidenhout and Reimer. 2004.

**Dieter, E., et al. 2007.** Developmental reprogramming after chromosome transfer into mitotic mouse zygotes. *Nature.* 2007, Vol. 447, pp. 679-685.

**Dobson, K. and Franche, R. L. 1989.** A conceptual and empirical review of the depressive realism hypothesis. *Canadian Journal of Behavioural Science.* 1989, Vol. 21, pp. 419-433.

**Dretske, F. 1981.** *Knowledge and the Flow of Information.* Cambridge : MIT Press, 1981.
**Dreyer, J. L. E. 1953.** *History of Astronomy from Thales to Kepler, 2nd edition. (revised reprint of History of the Planetary Systems from Thales to Kepler, 1906).* New York : Dover, 1953.
**Duhem, P. 1906.** *The Aim and Structure of Physical Theory.* [trans.] P. Wiener. 1954. Princeton : Princeton Univ. Press, 1906.
**Dummett, M. 1978.** *Truth and Other Enigmas.* London : Duckworth, 1978.
**Edwards, A.W.F. 1977.** *Foundations of Mathematical Genetics.* Cambridge : Cambridge Univ. Press, 1977.
**Evans, G. 1973.** The Causal Theory of Names. *Proceedings of the Aristotelian Society.* 1973, Vol. 47, pp. 187-208.
—. **1982.** *The Varieties of Reference.* Oxford : Oxford University Press, 1982.
**Falk, R. 1986.** What is a gene? *Studies in the History and Philosophy of Science.* 1986, Vol. 17, pp. 133-170.
**Feigl, H. 1950.** Existential Hypotheses: Realistic versus Phenomenalistic Interpretations. *Philosophy of Science.* 1950, Vol. 17, pp. 35-62.
**Ferguson, K. 2002.** *Tycho & Kepler: the unlikely partnership that forever changed our understanding of the heavens.* New York : Walker, 2002.
**Feyerabend, P. 1962.** Explanation, Reduction, and Empiricism. *Minnesota Studies in the Philosophy of Science.* Minneapolis : Univ. of Minessota Press, 1962, Vol. III.
—. **1987.** *Farewell to Reason.* London : Verso, 1987.
**Fiers, W., et al. 1971.** Recent progress in the sequence determination of bacteriophage MS2 RNA. *Biochimie.* 1971, Vol. 53, pp. 495–506.
**Fine, A. 1984.** The Natural Ontological Attitude. [ed.] Jarrett Leplin. *Scientific Realism.* Berkeley : Univ. of California Press, 1984.
**Fogle, T. 2001.** The dissolution of the protein coding genese in molecular biology. [ed.] R. Falk, H. J. Rheinberger and P. Beurton. *The Concept of the Gene in Development and Evolution.* Cambridge : Cambridge University Press, 2001.
**Frege, G. 1893.** On Sense and Reference. [ed.] P. Geach and M. Black. *Translations from the Philosophical Writings of Gottlob Frege.* 1952. Oxford : Blackwell, 1893.
**Friedman, M. 1983.** *Foundations of Sapce-Time Theories.* Chicago : University of Chicago Press, 1983.
**Frigg, R. 2006.** Scientific Representation and the Semantic View of Theories. *Theoria.* 2006, Vol. 55, pp. 37-53.
**Galilei, Galileo. 1632.** *Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican.* [trans.] S. Drake. 1962. Berkeley : Univ. of California Press, 1632.
**Gayon, J. 1998.** *Darwinism's Struggle for Survival. Heredity and the Hypothesis of Natural Selection.* Cambridge : Cambridge University Press, 1998.
—. **2000.** From measurement to organization: A philosophical scheme for the history of the concept of heredity. [ed.] Raphael Falk, and Hans-Jörg Rheinberger Peter Beurton. *The Concept of the Gene in Development and Evolution. Historical and Epistemological Perspectives.* Cambridge : Cambridge University Press, 2000.
**Gerstein, M. B., et al. 2007.** What is a gene, post-ENCODE? History and updated definition. *Genome Research.* 2007, Vol. 17, 6, pp. 669-681.
**Giere, R. 1988.** *Explaining Science: A Cognitive Approach.* Chicago : Univ. of Chicago Press, 1988.

—. **2004.** How Models Are Used to Represent Reality. *Philosophy of Science.* 2004, Vols. 71, Supplement, pp. S742-752.
—. **1999.** *Science Without Laws.* Chicago : University of Chicago Press, 1999.
—. **2000.** Theories. [ed.] W. H. Newton-Smith. *A Companion to the Philosophy of Science.* Oxford : Blackwell, 2000.
**Hacking, I. 1984.** Experimentation and Scientific Realism. [ed.] J. Leplin. *Scientific Realism.* Berkeley : Univ. of California Press, 1984.
—. **1982.** Experimentation and Scientific Realism. *Philosophical Topics.* 1982, Vol. 13, pp. 71-87.
—. **1983.** *Representing and Intervening.* Cambridge : Cambridge Univ. Press, 1983.
**Hanson, N. R. 1972.** *Patterns of Discovery.* Cambridge : Cambridge Univ. Press, 1972.
**Hardcastle, G. L. 1995.** S.S. Stevens and the origins of operationism. *Philosophy of Science.* 1995, Vol. 62, pp. 404-424.
**Heidelberger, M. 2003.** Theory-Ladenness and Scientific Instruments in Experimentation. [ed.] Hans Radder. *The Philosophy of Scientific Experimentation.* Pittsburgh : University of Pittsburgh Press, 2003.
**Helmholtz, H. von. 1866.** Concerning the perceptions in general. [trans.] J. P. C. Southall. *Treatise on physiological optics.* 1962. New York : Dover, 1866, Vol. III.
**Hempel, C. 1950.** A Note on Semantic Realism. *Philosophy of Science.* 1950, Vol. 17, pp. 169-173.
**Hempel, C. and Oppenheim, P. 1965.** Studies in the Logic of Explanation. *Aspects of Scientific Explanation.* New York : Free Press, 1965.
**Hempel, C. 1963.** Implications of Carnap's Work for the Philosophy of Science. [ed.] P. Schlipp. *The Philosophy of Rudolf Carnap.* La Salle : Open Court, 1963.
—. **1945.** Studies in the Logic of Confirmation I. *Mind NS.* 1945, Vol. 54, 213, pp. 1–26.
—. **1965.** *The Philosophy of Natural Science.* Englewood Cliffs : Prentice-Hall, 1965.
**Hershey, A.D. and Chase, M. 1955.** An upper limit to the protein content of the germinal substance of bacteriophage T2. *Virology.* 1955, Vol. 1, pp. 108–127.
**Hesse, M. 1966.** *Models and Analogies in Science.* Notre Dame : Univ. of Notre Dame Press, 1966.
**Hitchcock, C. 1995.** Discussion: Salmon on Explanatory Relevance. *Philosophy of Science.* 1995, Vol. 62, pp. 304-320.
**Hull, D. 1974.** *Philosophy of Biological Science.* Englewood Cliffs : Prentice Hall, 1974.
—. **1979.** Reduction in Genetics. *Philosophy of Science.* 1979, Vol. 46, pp. 316-320.
—. **1972.** Reduction in Genetics: Biology or Philosophy? *Philosophy of Science.* 1972, Vol. 39, pp. 491-499.
**Hyman, A. 1993.** A Simple Cartesian Treatment of Planetary Motion. *European Journal of Physics.* 1993, Vol. 14, pp. 145-147.
**Jacob, F. and Monod, J. 1961.** Genetic Regulatory Mechanisms in the Synthesis of Proteins . *Journal of Molecular Biology.* 1961, Vol. 3, pp. 318-356.
**Jammer, M. 1960.** *Concepts of Space: The History of Theories of Space in Physics.* New York : Harper, 1960.
**Johannsen, W. 1909.** *Elemente der exakten Erblichkeitslehre.* Jena : s.n., 1909.
**Kemeny, J. and Oppenheim, P. 1956.** On Reduction. *Philosophical Studies.* 1956, Vol. 7, pp. 6-19.

**Kitcher, P. 1984.** 1953 and All That. A Tale of Two Sciences. *The Philosophical Review.* 1984, Vol. 93, pp. 335-373.
—. **1982.** *Abusing Science.* Cambridge : MIT press, 1982.
—. **1981.** Explanatory Unification . *Philosophy of Science.* 1981, Vol. 48, pp. 507-531.
—. **1989.** Explanatory Unification and the Causal Structure of the World. [ed.] P. Kitcher and W. Salmon. *Minnesota Studies in the Philosophy of Science.* Minneapolis : Minnesota Univ. Press, 1989, Vol. XIII.
—. **1982.** Genes. *British Journal for the Philosophy of Science.* 1982, Vol. 33, pp. 337-359.
—. **1999.** The hegemony of molecular biology. *Biology and Philosophy.* 1999, Vol. 14, 2, pp. 195-210.
**Klee, R. 1997.** *Introduction to Philosophy of Science: Cutting Nature at Its Seams.* Oxford : Oxford Univ. Press, 1997.
**Kornblith, H. 2003.** *Knowledge and Its Place in Nature.* Oxford : Clarendon Press, 2003.
**Koyré, A. 1957.** *From the Closed World to the Infinite Universe.* Baltimore : Johns Hopkins Press, 1957.
**Kripke, S. 1977.** Speaker's Reference and Semantic Reference. *Midwest Studies in Philosophy.* 1977, Vol. 2, pp. 255-276.
**Kripke, Saul. 1972.** Naming and Necessity. [ed.] D.Davidson and G. Harman. *Semantics of Natural Language.* Dordrecht : Reidel, 1972.
**Kuhn, T. and Heilbron, J. L. 1969.** The Genesis of the Bohr Atom. *Historical Studies in the Physical Sciences.* 1969, Vol. 1, pp. 211-290.
**Kuhn, T. 1982.** Commensurability, Comparability, Communicability. [ed.] P. Asquith and T. Nickels. *Philosophy of Science Association.* 1982, Vol. 2, pp. 669-688.
—. **1977.** *The Essential Tension: Selected Studies in Scientific Tradition and Change.* Chicago : University of Chicago Press, 1977.
—. **1970.** *The Structure of Scientific Revolutions.* Chicago : Chicago Univ. Press, 1970.
—. **1976.** Theory-Change as Structure-Change: Comments on the Sneed Formalism. *Erkenntnis.* 1976, Vol. 10, pp. 179-199.
**Lakatos, I. and Zahar, E.G. 1976.** Why Did Copernicus's Programme Supersede Ptolemy's? [ed.] R. Westman. *The Copernican Achievement.* Los Angeles : University of California Press, 1976.
**Lakatos, I. 1970.** Falsification and the Methodology of Scientific Research Programmes. [ed.] I. Lakatos and A. Musgrave. *Criticism and the Growth of Knowledge.* London : Cambridge Univ. Press, 1970.
**Laudan, L.** A Confutation of Convergent Realism. [ed.] Jarrett Leplin. *Scientific Realism.* Berkeley : Univ. of California Press.
—. **1984.** A Confutation of Convergent Realism. [ed.] Jarrett Leplin. *Scientific Realism.* Berkeley : Univ. of California Press, 1984.
—. **1996.** *Beyond Positivism and Relativism.* Boulder : Westview Press, 1996.
**Laymon, R. 1985.** Idealizations and the Testing of Theories by Experimentation. [ed.] P. Achinstein and O. Hannaway. *Observation Experiment and Hypothesis in Modern Physical Science.* Cambridge : MIT Press, 1985.
—. **1984.** The Path from Data to Theory. [ed.] Jarrett Leplin. *Scientific Realism.* Berkeley : Univ. of California Press, 1984.
**Lewontin, R. and Levins, R. 1985.** *The Dialectical Biologist.* Cambridge : Cambridge University Press, 1985.

**Lipton, P. 1993.** Is the Best Good Enough? *Proceedings of the Aristotelian Society.* 1993, Vol. 93, 2, pp. 89-104.

**Mach, E. 1893.** *The Science of Mechanics: A Critical and Historical Account of Its Development.* [trans.] T. J. McCormack. 1960. Lasalle : Open Court, 1893. sixth edition, with an introduction by K. Menger.

**Machamer, P., Darden, L. and Craver, C. F. 2000.** Thinking About Mechanisms . *Philosophy of Science.* 2000, Vol. 67, pp. 1-25.

**Maxwell, G. 1962.** The Ontological Status of Theoretical Entities. [ed.] H. Feigl and G. Maxwell. *Scientific Explanation, Space and Time.* Minneapolis : Univ. of Minnesota Press, 1962.

**McClintock, B. 1929.** A cytological and genetical study of triploid maize. *Genetics.* 1929, Vol. 14, pp. 180–222.

**McMullin, E. 1987.** Explanatory Successes and the Theory of Truth. [ed.] N. Rescher. *Scientific Inquiry in Philosophical Perspective.* Lanham : University Press of America, 1987.

**—. 1985.** Galilean Idealization. *Studies in the History and Philosophy of Science.* 1985, Vol. 16, pp. 247-273.

**Medin, D. L. and Schaffer, M. M. 1978.** Context theory of classification learning. *Psychological Review.* 1978, Vol. 85, pp. 207-238.

**Mendel, J.G. 1866.** Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn 4 Abhandlungen . 1866, pp. 3–47 (http://www. mendel-web.org/MWolby.html).

**Menuge, A. 1995.** The Scope of Observation. *Philosophical Quarterly.* 1995, Vol. 45, pp. 60-69.

**Meselson, M. and Stahl, F.W. 1958.** The Replication of DNA in Escherichia coli. *Proc Natl Acad Sci U S A.* 1958, Vol. 44, 7, pp. 671-682.

**Michels, C. W. 1973.** Rigorous Elementary Derivation of the Inverse Square Law from Kepler's Laws. *American Journal of Physics.* 1973, Vol. 41, p. 1007.

**Miller, R. 1987.** *Fact and Method.* Princeton : Princeton Univ. Press, 1987.

**Moran, S., et al. 1994.** *Biochemistry* . Englewood Cliffs : Prentice Hall, 1994.

**Morgan, T. H. 1935.** The relation of genetics to physiology and medicine. *Les prix Nobel en 1933. Imprimerie Royale.* 1935, pp. 1-16.

**Morgan, T. H., et al. 1915.** *The Mechanism of Mendelian Heredity.* New York : Henry Holt and Company, 1915.

**Morrison, M. 1999.** Models as Autonomous Agents. [ed.] M. Morgan and M. Morrison. *Models as Mediators. Perspectives on Natural and Social Science.* Cambridge : Cambridge University Press, 1999.

**—. 2000.** *Unifying Scientific Theories: Physical Concepts and Mathematical Structures.* Cambridge : Cambridge Univ. Press, 2000.

**Moss, L. 2003.** *What Genes Can't Do.* Cambridge : MIT Press, 2003.

**Muller, H. J. 1936.** Physics in the Attack on the Fundamental Problems of Genetics. *Scientific Monthly.* 1936, Vol. 44, pp. 210-214.

**—. 1951.** The development of the gene theory. [ed.] Leslie C. Dunn. *Genetics in the 20th Century. Essays on the Progress of Genetics During its First 50 Years.* New York : MacMillan, 1951, pp. 77-99.

**Murray, C. 2004.** *Reconstructing Reason and Representation.* Cambridge : MIT Press, 2004.

**Nagel, E. 1961.** Experimental Laws and Theories. *The Structure of Science: Problems in the Logic of Scientific Explanation.* New York and Burlingame : Harcourt, Brace & World, 1961.

—. **1974.** Issues in the Logic of Reductive Explanations. *Teleology Revisited.* New York : Columbia Univ. Press, 1974.
—. **1950.** Science and Semantic Realism. *Philosophy of Science.* 1950, Vol. 17, pp. 174-181.
—. **1961.** The Structure of Science. *Problems in the Logic of Scientific Explanation.* New York : Harcourt, Brace and World, 1961.
**Newton, I. 1687.** *Mathematical Principles of Natural Philosophy.* [trans.] A. Motte. 1946. Berkeley : Univ. of California Press, 1687.
**Newton-Smith, W. 2000.** Underdetermination of Theory by Data. [ed.] W. H. Newton-Smith. *A Companion to the Philosophy of Science.* Oxford : Blackwell, 2000.
**Nickles, T. 1973.** Two concepts of inter-theoretic reduction. *Journal of Philosophy.* 1973, Vol. 70, pp. 181–201.
**Niiniluoto, I. 2000.** *Critical Scientific Realism.* Cambridge : Cambridge Univ. Press, 2000.
**Nirenberg, M. W. and Matthaei, J. H. 1961.** The Dependence of Cell-Free Protein Synthesis in E. coli Upon Naturally Occurring or Synthetic Polyribonucleotides", Proceedings of the National Academy of Sciences. 1961, Vol. 47, pp. 1588-1602.
**Nirenberg, M., et al. 1965.** RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl.Acad. Sci.* 1965, Vol. 53, pp. 1161–1168.
**Nosofsky, R. M. and Johansen, M. J. K. 2000.** Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review.* 2000, Vol. 7, pp. 375-402.
**Oddie, G. 1986.** *Likeness to Truth.* Dordrecht : Reidel, 1986.
**Olby, R. 1985.** *Origins of Mendelism.* Chicago : University of Chicago Press, 1985.
**Painter, T. 1934.** A New Method for the Study of Chromosome Aberrations and the Plotting of Chromosome Maps in Drosophila Melanogaster. *Genetics.* 1934, Vol. 19, pp. 175-188.
**Palmer, S. E. 1999.** *Vision Science: Photons to Phenomenology.* Cambridge : MIT Press, 1999.
**Podolsky, S. H. and Tauber, A. I. 1997.** *The Generation of Diversity: Clonal Selection Theory and the Rise of Molecular Immunology* . Cambridge, MA : Harvard University Press, 1997.
**Popper, K. 1976.** Background Knowledge and Scientific Growth. [ed.] Sandra G. Harding. *Can Theories Be Refuted?: Essays on the Duhem-Quine Thesis.* Dordrecht : Springer, 1976.
—. **1963.** *Conjectures and Refutations.* London : Routledge and K. Paul, 1963.
—. **1959.** *The Logic of Scientific Discovery.* New York : Basic Books, 1959.
—. **1965.** Truth, Rationality and the Growth of Scientific Knowledge. *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York : Harper, 1965.
**Posner, M. I. and Keele, S. W. 1968.** On the genesis of abstract ideas. *Journal of Experimental Psychology.* 1968, Vol. 77, pp. 353-363.
**Prinz, J. J. 2002.** *Furnishing the Mind: Concepts and Their Perceptual Basis.* Cambridge : MIT Press, 2002.
**Prior, E. 1982.** The Dispositional/Categorical Distinction. *Analysis.* 1982, Vol. 42, pp. 93–96.
**Prior, E., Pargetter, R. and Jackson, F. 1982.** Three Theses About Dispositions. *American Philosophical Quarterly.* 1982, Vol. 19, pp. 251–257.
**Psillos, S. 1999.** *Scientific Realism: How Science Tracks Truth.* London : Routlege, 1999.
**Putnam, H. 1991.** Explanation and Reference. [ed.] R. Boyd, P. Gasper and J. D. Trout. *The Philosophy of Science.* Cambridge : MIT Press, 1991.
—. **1975.** Explanation and Reference. *Philosophical Papers.* Cambridge : Cambridge Univ. Press, 1975, Vol. 2.
—. **1990.** *Realism with a Human Face.* Cambridge : Harvard University Press, 1990.

—. **1991.** The 'Corroboration' of Theories. [ed.] R. Boyd, P. Gasper and J. D. Trout. *The Philosophy of Science.* Cambridge : MIT Press, 1991.
—. **1962.** What Theories Are Not. *Philosophical Papers.* 1975. Cambridge : Cambridge Univ. Press, 1962, Vol. 1.
**Quine, W. V. 1975.** On Empirically Equivalent Systems of the World. *Erkenntnis.* 1975, Vol. 9, pp. 313-328.
—. **1975.** On Empirically Equivalent Theories of the World. *Erkenntnis.* 1975, Vol. 9, pp. 313-328.
—. **1951.** Two Dogmas of Empiricism. *Philosophical Review.* 1951, Vol. 60, pp. 20-43.
**Ramsey, F. P. 1929.** Theories. [ed.] D. H. Mellor. *Foundations: Essays in Philosophy, Mathematics and Economics.* 1978. London : Routledge and K. Paul, 1929.
**Ravnik, S. E. and Wolgemuth, D. J. 1999.** Regulation of meiosis during mammalian spermatogenesis: the A-type cyclins and their associated cyclin-dependent kinases are differentially expressed in the germ-cell lineage. *Dev. Biol.* 1999, Vol. 207, 2, pp. 408-418.
**Redhead, M. 1980.** Models in Physics. *British Journal for the Philosophy of Science.* 1980, Vol. 31, pp. 145-163.
**Rheinberger, H.-J. 2000.** Gene concepts: Fragments from the perspective of molecular biology. [ed.] Raphael Falk, and Hans-Jörg Rheinberger Peter Beurton. *The Concept of the Gene in Development and Evolution. Historical and Epistemological Perspectives.* Cambridge : Cambridge University Press, 2000.
**Richard, G. 1997.** Knowledge and Perception in Illusion. 1997, Vol. 352, pp. 1121-1128.
**Rindler, W. 2001.** *Relativity: Special, General and Cosmological.* Oxford : Oxford Univ. Press, 2001.
**Roll-Hansen, Nils. 1989.** The crucial experiment of Wilhelm Johannsen. *Biol. Philos.* 1989, Vol. 4, pp. 303–329.
**Rosen, G. 1994.** What is Constructive Empiricism? *Philosophical Studies.* 1994, Vol. 74, pp. 143-178.
**Rosenberg, A. 2007.** Reductionism (and Antireductionsim) in Biology. [ed.] Hull. D. and M. Ruse. *The Cambridge Companion to the Philosophy of Biology.* Cambridge : Cambridge University Press, 2007.
—. **1985.** *The Structure of Biological Science.* Cambridge : Cambridge Univ. Press, 1985.
—. **1978.** The Supervenience of Biological Concepts. *Philosophy of Science.* 1978, Vol. 45, pp. 368-386.
**Rueger, A. 2005.** Perspectival Models and Theory Unification. *British Journal for the Philosophy of Science.* 2005, Vol. 56.
**Ruse, M. 1971.** Reduction, Replacement, and Molecular Biology. *Dialectica.* 1971, Vol. 25, pp. 39-72.
—. **1973.** *The Philosophy of Biology.* London : Hutchinson, 1973.
**Russell, B. 1919.** Knowledge by Acquaintance and Knowledge by Description. *Mysticism and Logic.* 1917. London : George Allen and Unwin, 1919.
—. **1905.** On Denoting. *Mind.* 1905, Vol. 14, pp. 479-493.
**Ryan, K. J. and Ray, C. G. 2004.** *Sherris Medical Microbiology.* s.l. : McGraw Hill, 2004.
**Saatsi, J. 2005.** Reconsidering the Fresnel–Maxwell theory shift: how the realist can have her cake and EAT it too. *Studies in History and Philosophy of Science.* 2005, Vol. 36, pp. 509-538.

**Salmon, W. 1985.** Empiricism: The Key Question. [ed.] N. Rescher. *The Heritage of Logical Positivism.* Lanham : University Press of America, 1985.

—. **1989.** Four Decades of Scientific Explanation. [ed.] P. Kitcher and W. Salmon. *Minnesota Studies in the Philosophy of Science.* Minneapolis : Minnesota Univ. Press, 1989, Vol. XIII.

—. **1984.** *Scientific Explanation and the Causal Structure of the World.* Princeton : Princeton University Press, 1984.

—. **1984.** *Scientific Explanation and the Causal Structure of the World.* Princeton : Princeton University Press, 1984.

—. **1971.** Statistical Explanation. *Statistical Explanation and Satistical Relevance.* Pittsburgh : University of Pittsburgh Press, 1971.

**Schaffner, K. F. 1967.** Approaches to reduction. *Philosophy of Science.* 1967, Vol. 34, pp. 137–147.

—. **1994.** Interactions among Theory, Experiment, and Technology in Molecular Biology. *PSA: Proceedings of the Biennial Meeting of the Philosophy of cience Association.* 1994, Vol. 2, pp. 192-205.

—. **1969.** The Watson-Crick Model and Reductionism. *British Journal of Philosophy of Science.* 1969, Vol. 20, pp. 325-348.

**Scriven, M. 1962.** Explanations, predictions, laws. [ed.] P. Kitcher and W. Salmon. *Minnesota Studies in the Philosophy of Science.* Minneapolis : Minnesota Univ. Press, 1962, Vol. III.

**Searle, J. 1983.** *Intentionality.* Cambridge : Cambridge Univ. Press, 1983.

**Sharrock, W. and Read, R. 2002.** *Kuhn: Philosopher of Scientific Revolution.* Cambridge : Polity, 2002.

**Silverstein, A. M. 1989.** *A history of immunology.* San Diego : Academic Press, 1989.

**Sklar, L. 2002.** *Theory and Truth: Philosophical Critique within Foundational Science.* Oxford : Oxford Univ. Press, 2002.

**Smart, J. J. C. 1968.** *Between Science and Philosophy.* New York : Random House, 1968.

—. **1963.** *Philosophy and Scientific Realism.* London : Routledge and K. Paul, 1963.

**Stein, H. 1989.** Yes, but... Some Skeptical Remarks on Realism and Antirealism. *Dialectica.* 1989, Vol. 43, pp. 47-65.

**Stevens, S. S. 1963.** Operationism and logical positivism. [ed.] M. H. Marx. *Theories in contemporary psychology.* New York : MacMillan, 1963.

**Sturtevant, H. 1913.** The linear arrangement of six sex-linked factors in Drosophila as shown by their mode of association. *J. Exp. Zool.* 1913, Vol. 14, pp. 43–59.

**Suppe, F. 1977.** *The Structure of Scientific Theories.* Chicago : Univ. of Illinois Press, 1977.

**Suppes, Patrick. 1960.** A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthèse.* 1960, Vol. 12, pp. 287-301.

—. **1962.** Models of Data. [ed.] Ernest Nagel, Patrick Suppes and Alfred Tarski. *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress.* Stanford : Stanford University Press, 1962, pp. 252-261.

**Sutton, W. S. 1903.** The chromosomes in heredity. *Biol. Bull.* 1903, Vol. 4, pp. 231-251.

**Teller, P. 2001.** Twilight of the Perfect Model. *Erkenntnis.* 2001, Vol. 55, pp. 393-415.

**Tinoco, I., Sauer, K. and Wang, J. C. 1995.** *Physical Chemistry: Principles and Applications in Biological Sciences.* Englewood Cliffs : Prentice Hall, 1995.

**van Fraassen, B. C. 1985.** Empiricism in Philosophy of Science. [ed.] P. M. Churchland and C. A. Hooker. *Images of Science.* Chicago : University of Chicago Press, 1985.

—. **1994.** Gideon Rosen on Constructive Empiricism. *Philosophical Studies.* 1994, Vol. 74, pp. 179-192.

**van Fraassen, B. C., LAdyman, J., Douven, I. and Horsten, L. 1997.** A Defence of van Fraassen's Critique of Abductive Reasoning: Reply to Psillos. *The Philosophical Quarterly.* 1997, Vol. 47, pp. 305-321.

**van Fraassen, B. 1989.** *Laws and Symmetry.* Oxford : Oxford Univ. Press, 1989.

—. **1980.** *The Scientific Image.* Oxford : Oxford Univ. Press, 1980.

**von Wright, G. 1971.** *Explanation and Understanding.* Ithaca : Cornell University Press, 1971.

**Wain, H.M., et al. 2002.** Guidelines for human gene nomenclature. *Genomics* . 2002, Vol. 79, pp. 464–470.

**Waskan, J. A. 2006.** *Models and Cognition.* Cambridge : MIT Press, 2006.

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

—. **1990.** Why the Anti-Reductionist Consensus Won't Survive: The Case of Classical Mendelian Genetics. *Proceedings to the Biennial Meeting of the Philosophy of Science Association.* 1990, pp. 125-139.

**Watson, J.D. and Crick, F.H.C. 1953.** A structure of deoxyribonucleic acid. *Nature* . 1953, Vol. 171, pp. 964–967.

**Weber, M. 2005.** *Philosophy of Experimental Biology.* Cambridge : Cambridge University Press, 2005.

**Wedeking, G. 1976.** Duhem, Quine and Grunbaum on Falsification. [ed.] Sandra G. Harding. *Can Theories Be Refuted?: Essays on the Duhem-Quine Thesis.* Dordrecht : Springer, 1976.

**Wimsatt, W. C. 2006.** Reductionism and its heuristics: Making ethodological reductionism honest. *Synthese.* 2006, Vol. 151, pp. 445–475.

—. **1976.** Reductive Explanation: A Functional Account. [ed.] A. C. Michalos, et al. *Boston Studies in the Philosophy of Science.* Dordrecht : Reidel, 1976, Vol. 30, pp. 671–710.

**Wimsatt, W. 1972.** Complexity and Organization. [book auth.] Proceedings of the Philosophy of Science Association. [ed.] K. F. Schaffner and R. S. Cohen. Dordrecht : Reidel, 1972, pp. 67-86.

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

**Worrall, J. 1982.** Scientific Realism and Scientific Change. *Philosophical Quarterly.* 1982, Vol. 32, pp. 201–231.

—. **1989.** Structural Realism: The best of both worlds? [ed.] D. Papineau. *The Philosophy of Science.* 1996. Oxford : Oxford Univ. Press, 1989.

**Wylie, A. 1986.** Arguments for Scientific Realism: The Ascending Spiral. *American Philosophical Quarterly.* 1986, Vol. 23, pp. 287-297.